Estimation of a Probability Distribution over a Hierarchical Classification.

Diana McCarthy
Cognitive & Computing Sciences
University of Sussex
Brighton BN1 9QH, UK
diana.mccarthy@cogs.susx.ac.uk

Abstract

Many natural language processing applications make use of hierarchical classifications whilst also having a statistical framework which requires an estimation of the probability distribution over the taxonomy. Data for estimating the probability distributions typically comes from corpora but estimation is complicated by the ambiguity of the data. One application involving such a task is the automatic acquisition of selectional preferences. The method of estimating these class probabilities is crucial to the success of the technique and this paper compares the previous schemes and concludes that correct modelling of class inclusion is essential. Modifications to previous approaches are described which help to curb the effect of ambiguity and avoid overly-general results.

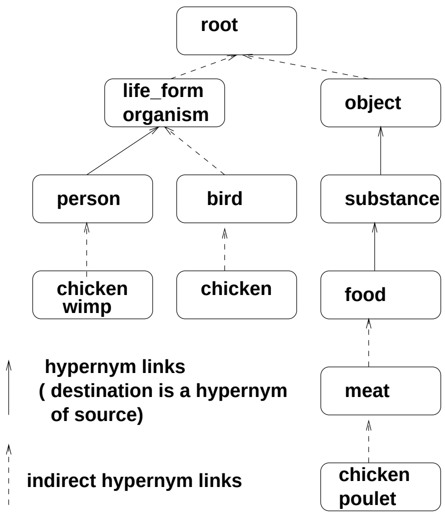

Figure 1: Hypernym hierarchy showing the senses of "chicken".

1 Introduction

The automatic acquisition of lexical information is an active area for current NLP research. The selectional preferences that predicates have for their arguments are one such valuable piece of information. Such work usually relies on a hierarchical classification of word senses and this has typically involved the lexical thesaurus WordNet (Miller, Beckwith, Felbaum, Gross, & Miller, 1993b).

WordNet is a lexical resource for open class words of English which is organised by part of speech (verb, adjective, adverb and noun). Within these categories all entries are defined in terms of word senses rather than word form, i.e. it is a thesaurus rather than a dictionary. WordNet entries (classes or synsets) are sets of synonymous word senses which are related to other classes of the same POS category by relations such as hyponymy. Any noun may belong to more than one class and any class may have more than one member. To illustrate the noun *"chicken"* belongs to three classes. The first sense is that of **MEAT** the second sense **BIRD** and the third **WIMP** [1]. These are illustrated along with some synonyms and hypernyms in figure 1.

This paper concerns the issues raised when counting occurrences of nouns in data and attempting to estimate the frequency distribution over word senses of these nouns by designating the appropriate Word-

[1] In diagrams I represent WordNet classes by some of the words stored at the class in WordNet. Meanwhile in text I use one of these close synonyms (if available) or a salient word at a hypernym class

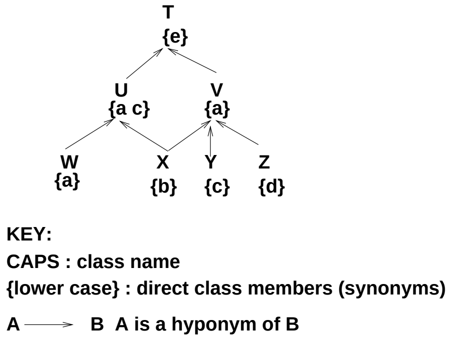

Figure 2: Taxonomy A.

Net class or classes for each noun token. For example which classes should we increment when we see the word *"chicken"*. How should the frequency (and therefore probability) distribution be divided between the various classes with direct membership and those with indirect membership? A probability distribution should sum to 1 but how should the probabilities at hypernyms relate to the probabilities of their hyponyms? This paper describes how these questions are tackled in two schemes for estimating probability distributions over the WordNet noun hypernym hierarchy and describes two modifications to one of these schemes which reduce the effect of ambiguity and thereby reduce over-generalisation.

2 Previous Work

Previous work estimating the probability distribution over the WordNet has typically involved the noun hypernym hierarchy for the acquisition of selectional preferences. There are two main approaches used, that by Resnik (Resnik, 1993) and that by Ribas (Ribas, 1995b, 1995a), and Li and Abe (Li & Abe, 1995; Abe & Li, 1996). The description of how the class probabilities are calculated differs between Li and Abe and Ribas but the methods are essentially the same.

To illustrate the difference in the approaches we will consider taxonomy A in figure 2. The lower case letters represent direct membership of the classes (in UPPER CASE) under which they appear. The arrows indicate the hypernymy relationship and any lower case member of a hyponym is an implicit indirect member of the hypernym classes. In this way the class **T** has *"e"* as a direct member and *"a b c d"* as indirect members. A class may have more than one member e.g. **U**, and an item may belong to more than one class e.g. *"a"*.

Resnik's approach doesn't distinguish between hypernymy and polysemy when estimating the frequency distribution. The frequency of a noun, in a given sample, contributes to all classes the noun belongs to, regardless of direct or indirect membership. Furthermore each frequency count is divided by the number of these classes to ensure that the sum of probabilities over the entire hierarchy equals one, equation 1 describes his estimation of the frequency of a class (c).

$$freq(c) = \sum_{n \in nouns\_at\_or\_under(c)} \frac{1}{|classes(n)|} \times freq(n) \tag{1}$$

In the work of Resnik and all other works described in this paper the estimation of class probabilities from the class frequencies is the straightforward maximum likelihood estimate:

$$\hat{p}(c) = \frac{freq(c)}{N}, where\ N = \sum_{c' \in all\_classes} freq(c') \tag{2}$$

The crucial flaw in Resnik's scheme arises from the lack of distinction between direct and indirect membership. This means that the probabilities of hyponym classes are not subsumed by their hypernyms. This also gives rise to a number of anomalies because the contribution of a noun depends on the depth of the classes (Ribas, 1995a).

For example, given the taxonomy A in figure 2, if the string *"a b d a e"* is observed the frequency and probability distributions would be as shown in table 1. The frequency contribution from each "word" (lower case letter) is shown separately at each of the classes to which it belongs, directly or indirectly. The class probabilities in Resnik's scheme sum to 1 but there are anomalies that Resnik himself acknowledges. For example **X** and **Z** both have members (*"b"* and *"d"* respectively) which occur the same number of times (once) but the classes end up having different frequency counts because of the difference in the number of hypernyms above them. Resnik suggested in his thesis that the assignment of class probabilities warranted further attention.

Ribas, in contrast, wants to adhere to the maxim that the probabilities of all the possible senses of a

Table 1: Resnik's Frequency and Probability Distributions.

Class	Freq	Prob $= \frac{Freq}{5}$
T	$\frac{2}{4}+\frac{1}{4}+0+\frac{1}{3}+\frac{1}{1}=2.0833$	$0.41\dot{6}$
U	$\frac{2}{4}+\frac{1}{4}+0=0.75$	0.15
V	$\frac{2}{4}+\frac{1}{4}+0+\frac{1}{3}=1.08\dot{3}$	$0.21\dot{6}$
W	$\frac{2}{4}=0.5$	0.1
X	$\frac{1}{4}=0.25$	0.05
Y	0	0
Z	$\frac{1}{3}=0.\dot{3}$	$0.0\dot{6}$

Table 2: Ribas' Frequency and Probability Distributions.

Class	Freq	Prob $= \frac{Freq}{5}$
T	$(2\times\frac{3}{3})+\frac{1}{1}+0+\frac{1}{1}+\frac{1}{1}=5$	1
U	$(2\times\frac{2}{3})+\frac{1}{1}+0=2.\dot{3}$	$0.4\dot{6}$
V	$(2\times\frac{1}{3})+\frac{1}{1}+0+\frac{1}{1}=2.\dot{6}$	$0.5\dot{3}$
W	$2\times\frac{1}{3}=0.\dot{6}$	$0.1\dot{3}$
X	$\frac{1}{1}=1$	0.2
Y	0	0
Z	$\frac{1}{1}=1$	0.2

word should sum to one rather than the probabilities of all senses and hypernyms. In this case the frequency of a class must be divided by the number of classes in which the word has direct membership. Additionally the probability of a hypernym should be the sum of the probabilities of all hyponyms plus any probability it has by virtue of direct membership. In this way the probability at the root of the hierarchy should equal one. He formalises this as a weight that is applied to the frequency count for each noun. The weight is specific to the noun (n) and class (c) and is the ratio between the number of classes directly containing n beneath and including c and the total number of classes containing n, this is shown in equation 3 and is equivalent to Ribas' estimation of the conditional probability of a class given the noun.

$$weight(n,c) = \frac{|direct\_classes_n\_at\_or\_under\_c|}{|direct\_classes_n|} \tag{3}$$

The unweighted frequency count at any class from any noun belonging at or beneath the class is simply the number of occurrences of that noun in the corpus. This is then multiplied by the weight as in equation 3. The frequency of a class is then the sum of all the weighted frequencies of nouns in the corpus belonging by direct or indirect membership:

$$freq(c) = \sum_{n \in nouns\_at\_or\_under(c)} freq(n) \times weight(n,c) \tag{4}$$

The frequency and probability counts for the example string *"a b d a e"* are shown in table 2. The frequency estimate from each letter is again shown separately. The contribution of *"a"* is shown as the first component of the addition for the frequency estimation of each class it belongs to. It belongs to classes **W**,**U**,**V** and **T**. For each of these classes the unweighted frequency 2 is multiplied by the appropriate weight. The denominator of the weight is 3 in all these cases, the number or polysemes of *"a"*. The numerator however is 1 for **V** and **T** since only one direct sense falls under these classes. Meanwhile for **U** and **T** the numerators are 2 and 3 respectively reflecting the membership at and under these classes.

The basic scheme for frequency assignment used by Li and Abe is effectively the same as that of Ribas but just expressed a little differently. The frequency estimate for a class is initially calculated only with direct membership in mind using the sum of the frequencies of each noun belonging at that class divided by the total number of classes this noun directly belongs to. These frequencies are then cumulated up the hierarchy ensuring that a hyponym only contributes its frequency once to each hypernym (this being necessary because WordNet is a DAG with multiple inheritance). The frequency estimation, and therefore probability estimation, is in effect the same as that of Ribas but the contributions of direct members are calculated by a separate process than those of indirect members, this is shown by equation 5.

$$freq(c) = \sum_{n \in nouns\_at\_class(c)} \frac{freq(n)}{|direct\_classes_n|} + \sum_{n \in nouns\_under\_class(c)} \frac{freq(n)}{|direct\_classes_n|} \quad (5)$$

With either formulation the results are the same. As can be seen in table 2 the probability of the root T equals 1. Also, this method ensures that for any noun the sum of probabilities $p(c|n)$ at the direct classes equals one whereas in Resnik's scheme it is the sum of probabilities from all classes for a noun (direct or hypernyms) that equals one.

3 Modifications to Improve Estimation.

The methods for estimating the class probabilities proposed by Ribas and Li and Abe make sense because they take the hypernym relation into account in a different way to the handling of polysemy. Hyponym probabilities are correctly assumed to be subsumed by hypernyms which accords with the class inclusion assumptions associated with a hypernym or "is a" relationship. Furthermore the frequency distribution given a particular noun is sensibly divided between the possible senses of that noun, and not the potentially large number of hypernyms of the noun. Li and Abe's approach for calculating the probability distribution and acquiring selectional preferences is taken as a starting point. Two modifications are made to the basic approach.

In this work selectional preferences are obtained largely as in Li and Abe's work (Li & Abe, 1995; Abe & Li, 1996). The preferences are represented by a cut across the WordNet noun hypernym hierarchy where the class members of the cut are not ancestors of one another and each is assigned a score indicating the strength of the preference. These models are

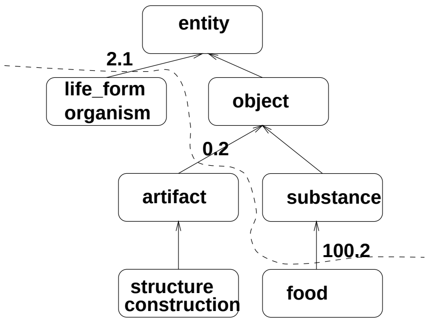

Figure 3: ATCM for 'eat' object slot.

known as association tree cut models (ATCMs) as the scores indicating preference are given by the association score in equation 6, the log of which gives the measure known as mutual information. This measures the "relatedness" between two words, or in this class-based work on selectional preferences between a class (c) and the verbal predicate (v).

$$A(c, v) = \frac{p(c|v)}{p(c)} \quad (6)$$

For example an ATCM for the direct-object slot of *"eat"* might look like that in figure 3.

The method of calculating class probabilities follows that of Li and Abe except for two important departures which are described in the two following sub-sections.

3.1 Modifications to WordNet

Li and Abe's method of obtaining selectional preferences requires a hierarchy where all words (terminals) appear on leaf classes. The method relies on finding cuts across the tree where no members of the cut are related by hypernymy and where they partition the set of words exhaustively and disjointly, i.e. all words appear at or under these classes on the cut. The sum of class probabilities on any such tree cut will equal 1. All the classes in WordNet's hierarchy contain noun word forms as members so in order to meet this requirement Li and Abe prune the tree at all classes where they have observed an occurrence in the data. This means that a far shallower version of WordNet is used. This however can give rise to overly-general preferences when the data contains an argument which occurs higher in the hierarchy than the prototypical arguments. For example, if a

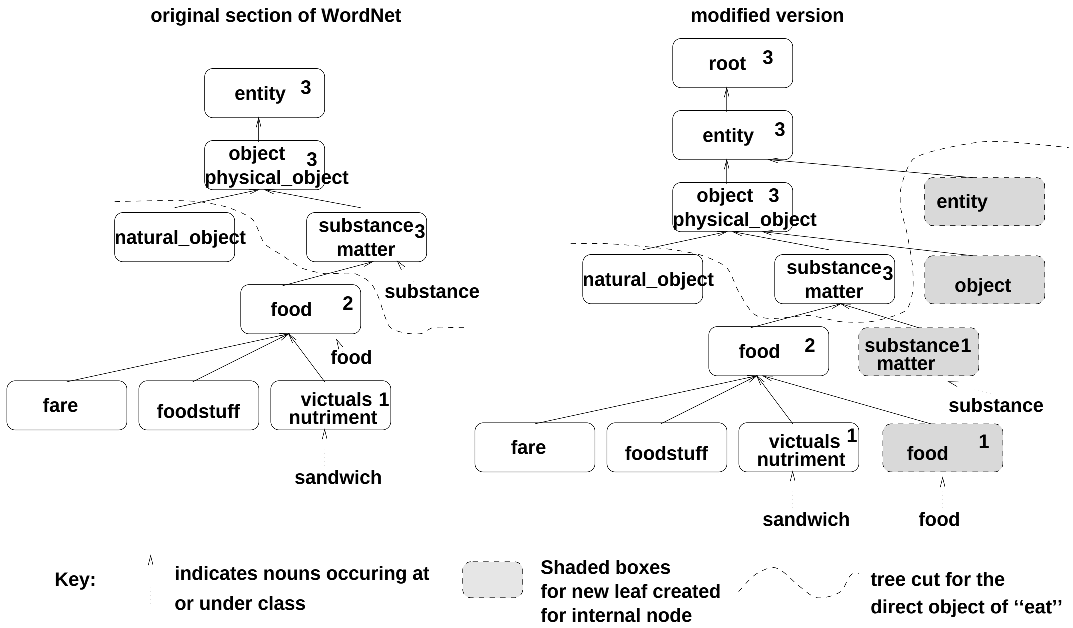

Figure 4: Creating New Leaves for Internal Nodes.

direct object such as *"entity"* in the predicate argument pair "build:entity" is observed then pruning the hierarchy at this level would result in loosing detail on the types of entities that are built (e.g. **OBJECT** vs **LIFE_FORM**).

In contrast, here the tree is not pruned but new leaves are created beneath each internal node. The frequency count attributed to nouns listed at the internal node is instead placed at the new leaf and then cumulated up the hierarchy in the usual way.

To illustrate look at the transformation of a small portion of the hierarchy in figure 4. The class frequency distribution for three heads, *"substance"*, *"sandwich"* and *"food"*, from the direct object slot of *"eat"* is shown by the numbers in the right-hand corner of the class boxes [2]. All classes without numbers have 0 frequency with respect to this small sample. In the unmodified version of WordNet it is quite possible to have a cut which does not cover all the noun senses and so the probability distribution along it would not sum to one. This is because terminal nouns such as *"substance"* occur at internal classes which can occur above a candidate cut. In our scheme the frequency, and therefore probability, contributions from such terminals are moved down to newly created leaves and so the probability axioms are maintained. In contrast the strategy adopted by Li and Abe would be to prune the tree at the **SUBSTANCE** class which would result in a shallow tree with no possibility of distinguishing between a preference for **FOOD** or **DRINK** at the direct object slot of *"eat"*.

One outstanding problem with the structure of WordNet arises because WordNet is actually a DAG rather than a tree. This is acknowledged by Li and Abe but nothing is done to circumvent this problem. This is also true of our implementation. We have not as yet attempted to compensate for this fact because of the small extent of the problem. On examination of the cases of multiple inheritance we find that less than 1% of the classes in the noun hypernym network have more than one parent class. Moreover, the majority of such classes are deep down in the network, as can be seen from table 3, and many concern compound nouns e.g. "school boy" "head nurse" which currently are not handled by our system. In cases of multiple-inheritance the probability distribution is still valid in that any hypernyms above more than one parent of a class will have the frequency count incremented only once for the multiply inherited class. Thus, in the single case at depth 3 pictured in figure 5 a frequency count of 1 detected at the **PERSON** node will only be incre-

[2] For the purposes of this example assume that these words are monosemous for the sake of simplicity

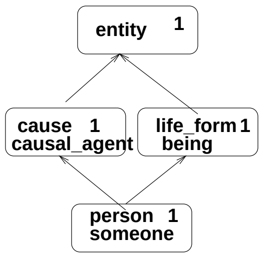

Figure 5: An example of multiple-inheritance.

mented once at the superordinate **ENTITY** class. A cut at this class will therefore meet the probability axioms. A cut at the layer of **CAUSAL_AGENT** and **LIFE_FORM** would however give rise to the sum of the classes on a cut including these to be greater than 1. This however is understandable when we consider that multiple-inheritance gives rise to overlap between the two parent classes both in terms of the subsumed nouns senses and their probability. The result is a coherent probability distribution across the cut. [3]

3.2 Word Sense Disambiguation

The second modification concerns the pervasive problem of word sense ambiguity. All the researchers whose work is discussed here acknowledge the problem and hope that the effect will be reduced with enough data. However Li and Abe and Ribas all report interference from erroneous senses. Li and Abe try and get rid of some of this interference by placing a threshold on class probabilities before including the classes in the cut representing selectional preferences. Even though a certain amount of noise can be removed we expect that resolution of the ambiguity would improve the results.

A method was sought for sense tagging the argument heads that did not require a large amount of training time (unsupervised approaches) nor manual effort to create large amounts of training data (supervised approaches). The approach selected was that where the most frequent sense is chosen regardless of context (Wilks & Stevenson, 1996). Criteria on when to apply the most frequent sense is specified and if these were not met then the ambiguity was left unresolved as in the work of Li and Abe and Ribas. The criteria included two thresholds, one on the frequency of the predominant sense and the other on the ratio between the first sense and the second. The details are described in (McCarthy, 1997). Although supervised training material is required for estimation of the sense frequencies this is smaller than would be the case with context dependent methods since fewer parameters are involved. There is already a portion of the Brown corpus that has been manually tagged with WordNet senses for the SEMCOR project (Miller, Leacock, Tengi, & Bunker, 1993a) and this was used for the sense frequency data.

To indicate the differences in estimation of class probabilities suppose we have an occurrence of the verb direct-object pair "eat:chicken" in our corpus, what would the class probability distribution look like? There are three senses of the noun *"chicken"*. These are indicated in figure 1. In Resnik's scheme a frequency count of 1/20 would be added to all direct classes containing *"chicken"* as well as all hypernyms. This is because there is a total of 20 such classes. In contrast Li and Abe and Ribas would place 1/3 at each direct sense and then these would be incremented up the hierarchy. Thus the probability distribution for the noun *"chicken"* over its direct senses would sum to one. In the scheme we are advocating, the predominant sense of *"chicken"* meets our criteria in the SEMCOR data (its first sense occurs 16 times whilst the second sense occurs 7 times). Thus the frequency 1 would all be placed at a new leaf (since this is an internal node) under the **MEAT** sense of *"chicken"*.

This is helpful because in the case of *eat:chicken* the predominant sense is the correct one. There are however occasions when the disambiguation might work against us for example given an occurrence of the verb direct-object pair *eat:fish* the predominant sense of fish is the **ANIMAL** sense rather than the **MEAT** sense. The rationale for using this rather rudimentary form of sense tagging is that it will help more than it hinders and so the overall results will be improved.

[3] Although the probability distribution is valid even in cases of multiple-inheritance there are unresolved problems arising because the method used to obtain the most appropriate cut across WordNet relies on a tree structure and not a DAG. These issues have been left for further research

Table 3: Frequency by depth of classes with multiple-inheritance

Depth	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Freq	0	0	1	3	22	111	161	147	68	21	12	11	0	1

3.3 Experimental Results

ATCMs were obtained using the subcategorization acquisition system described by Briscoe and Carroll (Briscoe & Carroll, 1997). This system was used to construct lexicons which list argument heads at the appropriate slots. Two sets of data were used both from shallow parses of the British National Corpus. The first was from 1.8 million words of parsed text and which formed a subset of the 10.8 million words used to construct the second lexicon. The point of trying two sample sizes was to see if the benefits of the crude disambiguation diminished with the larger sample. Evaluation was performed both by informal observation and quantitative evaluation comparing the acquired preferences to data from a "gold standard".

The ATCMs were obtained for a test set of 30 verbs:
add, agree, allow, ask, begin, believe, bring, build, call, cause, change, charge, choose, consider, cut, decide, end, establish, expect, feel, find, fix, give, help, like, move, produce, provide, seem, swing.

These verbs were selected because of the range of complement patterns they exhibit and were not picked with respect to their selection properties.

3.4 Informal Evaluation

From informal observation it could be seen that calculation of class probabilities using the crude disambiguation resulted in more appropriate (and usually less general) cuts than those where no disambiguation was used. Over-generalisation is cited as a result of noise from erroneous senses (Ribas, 1995a).

This can be seen in figure 6 which shows a small part of two ATCMs for the direct object slots of *"begin"* and *"produce"* respectively. These were produced from the larger lexicon built using the 10.8 million word sample. For clarity only a few classes on the cuts are pictured, these are classes involving strong preferences. The ATCMs obtained using the first sense heuristic are indicated with a bolder line and fall below the ATCMs produced without this heuristic. This specificity often coincides with intuition for example the direct object slot of *"begin"* features a strong preference for the class **TIME_PERIOD** when the first sense heuristic is used whilst without this heuristic a weaker preference is indicated for the hypernym class **MEASURE**. In the case of *"produce"* a preference is shown for the class **OBJECT**, without any disambiguation of the input data, whilst with the first sense heuristic the preference is more distinctive indicating preferences for *"inventions"* whilst not for nouns such as *"bridge"* belonging to the **STRUCTURE** class. In addition to the specificity the first sense heuristic tends to produce fewer but stronger preferences than without WSD, though this needs to be quantified.

An even more striking difference was shown with the smaller lexicon from 1.8 million words. With the smaller sample of argument heads the cut for some verbs was at the root indicating that no preferences could be observed for these verbs. This occurred more frequently without word sense disambiguation affecting 12 of the target verbs as opposed to 8 when the first sense heuristic was used. [4]. All 30 verbs had ATCMs below the root node with the larger data set.

The improvement observed informally is supported by some preliminary evaluation.

3.5 Gold Standard Evaluation

The "gold standard" used here is the Cambridge International Dictionary of English (CIDE) (Procter, 1995). The dictionary, as with most others, provides alongside each entry a list of example uses. From these the head nouns at direct object slot position have been obtained manually and each assigned a WordNet class that best represents the sense of the noun. As much as possible these senses have contained the noun in their set of synonyms but in cases where an appropriate class cannot be found as close

[4] Preference acquisition was performed on only 28 verbs from the smaller sample as 2 had frequencies below a threshold set at 20.

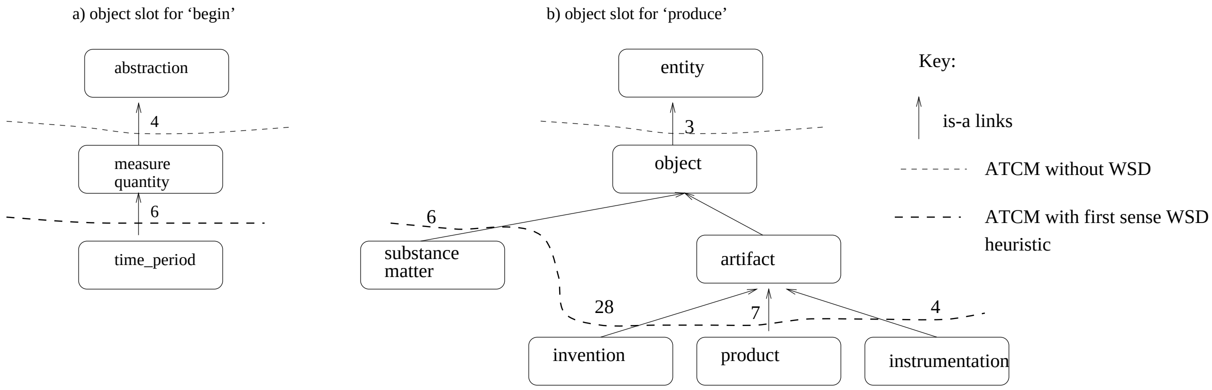

Figure 6: ATCM for a) ‘begin’ object slot and b) ‘produce’ object slot.

Table 4: CIDE Evaluation

Data (Million)	1st S	% correct	BL	diff
1.8	No	53	37	16
1.8	Yes	70	46	24
10.8	No	92	58	34
10.8	Yes	89	48	41

a sense as possible to the meaning of the head noun is chosen.

The preferences for the target verb have been evaluated against this set of test direct object senses. Each test item is scored correctly if it falls at or under one of the classes on the cut with an association score greater than 1. A baseline has been formulated that attempts to test the chance that a given sense will fall at or under a preference (class on the ATCM with association score above 1). This baseline takes an average of the ratios between the number of classes with a score above 1 and the number of classes on the cut for all instances in the test set. This is expressed in equation 7 where i is an instance from the test sample S, and v is the verb specified in the instance i.

$$\sum_{i \epsilon S} \frac{|prefs_v|}{|classes\_in\_ATCM_v|} \qquad (7)$$

This baseline does not account for the specificity of the cut explicitly but a more specific cut will typically have more classes and less preferences (but stronger ones). It simply assumes that all target senses will fall under the cut and measures the chance that an item will fall under one of the classes on the cut that expresses a preference.

For example, the verb : direct object pair *“believe:robber”* was observed in the dictionary. The sense assigned for *“robber”* was that under **PERSON**. This falls under both ATCMs with and without the first sense heuristic as can be seen in figure 7. Without disambiguation the baseline ratio for this instance is $\frac{12}{25}$ whereas with the first sense heuristic the ratio is $\frac{9}{26}$.

Results are shown in table 4. The last column indicates the difference between the % correct and the baseline (BL). For both sample sizes the difference is increased by the first sense heuristic (indicated by 1st S).

From both informal inspection and preliminary evaluation these modifications do appear to improve the selectional preferences obtained.

4 Summary

All the methods for acquiring selectional preferences from hierarchies such as WordNet are dependent on a good estimation of class probabilities. Methods should respect the class inclusion relationship of hypernyms for their hyponyms. Additionally care needs to be taken when distributing the frequency of polysemous nouns. The frequency distribution for any particular noun should be spread over its direct classes in the absence of any means of disambiguation. The problem of word sense ambiguity is so extensive that although with large samples some noise can be removed there is evidence that even crude sense disambiguation helps matters. It is predicted that a more detailed disambiguation, taking

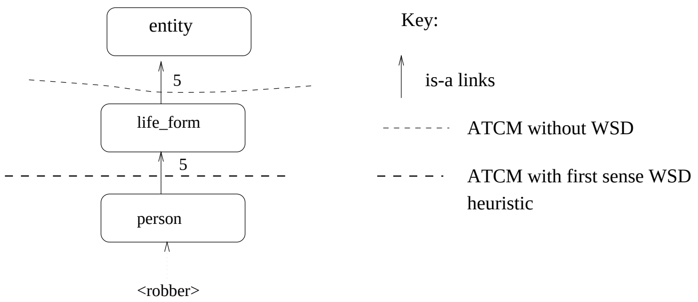

Figure 7: 'robber' under ATCMs for 'believe' object slot.

context into account, would improve accuracy still further but the pay-off would have to be worth the additional costs.

Reference

Abe, N., & Li, H. (1996). Learning word association norms using tree cut pair models.. In *Proceedings of the 13th International Conference on Machine Learning ICML.*

Briscoe, T., & Carroll, J. (1997). Automatic extraction of subcategorization from corpora.. In *Fifth Applied Natural Language Processing Conference.*, pp. 356–363.

Li, H., & Abe, N. (1995). Generalizing case frames using a thesaurus and the MDL principle. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pp. 239–248 Bulgaria.

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

Miller, George, A., Leacock, C., Tengi, R., & Bunker, R. T. (1993a). A semantic concordance.. In *Proceedings of the ARPA Workshop on Human Language Technology.*, pp. 303–308. Morgan Kaufman.

Miller, G., Beckwith, R., Felbaum, C., Gross, D., & Miller, K. (1993b). *Introduction to WordNet: An On-Line Lexical Database.* ftp//clarity.princeton.edu/pub/WordNet/5papers.ps.

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

Ribas, F. (1995a). *On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy.* Ph.D. thesis, University of Catalonia.

Ribas, F. (1995b). On learning more appropriate selectional restrictions. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics.*, pp. 112–118.

Wilks, Y., & Stevenson, M. (1996). The grammar of sense - is word-sense tagging much more than part-of-speech tagging?. cmp-lg/9607028. Submitted to Natural Language Engineering.