Word Sense Disambiguation for Acquisition of Selectional Preferences

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

7th May 1997

Abstract

The selectional preferences of verbal predicates are an important component of lexical information useful for a number of NLP tasks including disambiguation of word senses. Approaches to selectional preference acquisition without word sense disambiguation are reported to be prone to errors arising from erroneous word senses. Large scale automatic semantic tagging of texts in sufficient quantity for preference acquisition has received little attention as most research in word sense disambiguation has concentrated on quality word sense disambiguation of a handful of target words. The work described here concentrates on adapting semantic tagging methods that do not require a massive overhead of manual semantic tagging and that strike a reasonable compromise between accuracy and cost so that large amounts of text can be tagged relatively quickly. The results of some of these adaptations are described here along with a comparison of the selectional preferences acquired with and without one of these methods. Results of a bootstrapping approach are also outlined in which the preferences obtained are used for coarse grained sense disambiguation and then the partially disambiguated data is fed back into the preference acquisition system.

1 Introduction

The automatic acquisition of lexical information is widely seen as a way of overcoming the bottleneck of producing NLP applications (Zernik, 1991). The selectional preferences that predicates have for their arguments provide useful information that can help with resolution of both lexical and structural ambiguity and anaphora as well as being important for identifying the underlying semantic roles in argument slots.

The work reported here concentrates on acquisition for verbal predicates since verbs are of such obvious importance for the lexicon. However it could also be applied to any other type of predication.

The main contribution of this work is that it uses shallow parses produced by a fully automatic parser and that some word sense disambiguation (WSD) is performed on the heads collected from these parses. Most current research on selectional preference acquisition has used the Penn Treebank parses (Resnik, 1993a, 1993b; Ribas, 1995; Li & Abe, 1995; Abe & Li, 1996) These are obtained semi- automatically with a deterministic parser and manual correction. Additionally the other approaches do not perform any WSD on the input data and most report a major source of error arising from the contribution of erroneous senses sometimes giving incorrect preferences and at other times a noticeable effect of over-generalisation (Ribas, 1995; Resnik, 1993a).

The relationship between selectional preference acquisition and WSD is a circular one. One potential use of selectional preferences is WSD yet their acquisition appears to require disambiguated data. There are ways of breaking this circle. This paper describes work comparing the preferences acquired with and without semantic tagging of the input data. The cost of word sense disambiguation is kept low enough to permit tagging of a sufficient quantity of data. An iterative approach is also outlined whereby the preferences so acquired are used to disambiguate the argument heads which are then fed back into the preference acquisition system.

It is hoped that with enough data erroneous senses can be filtered out as noise (Li & Abe, 1995). However tagged data should produce more appropriate selectional restrictions provided the tagging is sufficiently accurate. Tagging should be particularly useful in cases where the data is scarce.

2 Previous Work

2.1 Selectional preference acquisition

Other approaches to selectional preference acquisition closely related to this are those of Resnik (Resnik, 1993b, 1993a) Ribas (1994, 1995), and Li and Abe (Li & Abe, 1995; Abe & Li, 1996)[1]. All use a class based approach using the WordNet hypernym hierarchy (Beckwith, Felbaum, Gross, & Miller, 1991) as the noun classification and finding selectional preferences as sets of disjoint noun classes (not related as descendants of one another) within the hierarchy. The key to obtaining good selectional preferences is obtaining classes at an appropriate level of generalisation. These researchers also use variations on the association score given in equation 1, the log of which gives the measure from information theory known as mutual information. This measures the "relatedness" between two words, or in the class-based work on selectional preferences between a class (c) and the predicate (v).

$$A(c, v) = \frac{p(c|v)}{p(c)} \tag{1}$$

In comparison to the conditional distribution $p(c|v)$ of the predicate (v) and noun class (c) the association score takes into account the marginal distribution of the noun class so that higher values are not obtained because the noun happens to occur more frequently irrespective of context. The conditional distribution might, for example, bias a class containing "people" as the direct object of "fly" in comparison to the class of "BIRDS" simply because the "PEOPLE" class occurs more in the corpus to start with.

Resnik and Ribas both search for disjoint classes with the highest score. Since smaller classes will fit the predicate better and will hence have a higher association value they scale up the mutual information value by the conditional distribution giving the association score in equation 2. The conditional distribution will be larger for larger classes and so in this way they hope to obtain an appropriate level of generalisation.

$$A(v, c) = P(c|v) log \frac{P(c|v)}{P(c)} \tag{2}$$

The work described here uses the approach of Li and Abe who rather than modifying the association score use a principle of data compression from information theory to find the appropriate level of generalisation. This principle is known as the Minimum Description Length (MDL) principle. In their approach selectional preferences are represented as a set of classes or a "tree cut" across the hierarchy which dominates all the leaf nodes exhaustively and disjointly. The tree cut features in a model, termed an "association tree cut model" (ATCM) which identifies an association score for each of the classes in the cut. In the MDL principle the best model is taken to be the one that minimises the sum of:

1. The Model Description Length - the number of bits to encode the model

2. The Data Description Length - the number of bits to encode the data in the model.

In this way, rather than searching for the classes with the highest association score, MDL searches for the classes which make the best compromise between explaining the data well by having a high association score and providing as simple (general) a model as possible and so minimising the model description length.

In all the systems described above the input is not disambiguated with respect to word senses. Resnik and Ribas both report erroneous word senses being a major source of error. Ribas explains that this occurs because some individual nouns occur particularly frequently as complements to a given verb and so all senses of these nouns also get unusually high frequencies.

Li and Abe place a threshold on class frequencies before consideration of a class. In this way they hope to avoid the noise from erroneous senses. In this paper some modifications to Li and Abe's system

[1] I shall refer to the work in papers (Li & Abe, 1995) and (Abe & Li, 1996) as "Li and Abe" throughout, since the two pieces of work relate to each other and both involve the same two authors

are described and a comparison is made of the use of some word sense disambiguation.

2.2 Word Sense Disambiguation

Since the literature on WSD is vast there will be no attempt to describe the variety of current work here.

Two approaches were investigated as possible ways for pretagging the head nouns that are used as input to the preference acquisition system. These were selected for having a low enough cost to enable tagging of a sufficient amount of text.

One strategy has been suggested by Wilks and Stevenson in which the most frequent sense is picked regardless of context. In this work they distinguish senses to the homograph level given the correct part of speech and report a 95% accuracy using the most frequent sense specified by LDOCE ranking. This approach has the advantage of simplicity and training data is only needed for the estimation of one parameter, the sense frequencies.

The other approach selected was Yarowsky's unsupervised algorithm (1995). This has the advantage that it does not require any manually tagged data. His approach relies on initial seed collocations to discriminate senses that can be observed in a portion of the data. This portion is then labelled accordingly. New collocations are extracted from the labelled sample and ordered by log-likelihood as in equation 3. The new ordered collocations are then used to relable the data and the system iterates between observing and ordering new collocations and re-labelling the data until the stopping condition is met. The final decision list of collocations can then be used at run-time.

$$\text{log-likelihood} = \log \frac{p(senseA|coll_i)}{p(other\_senses|coll_i)} \qquad (3)$$

3 Experiments

3.1 Word Sense Disambiguation

Preliminary experiments have been carried out using adaptations of the two approaches mentioned above.

3.1.1 Experiment 1

This experiment followed the approach of using the first sense regardless of context. Wilks and Stevenson did this in order to disambiguate LDOCE homographs. Distinguishing between WordNet senses is a much harder problem and so performance was not expected to be as good.

Table 1: Threshold 5 ratio 2

DATA	RECALL	PRECISION
Brown	61	86
LOB(179)	44	69
WSJ (191)	41	68

The only frequency information available for WordNet senses, assuming large scale manual tagging is out of the question, is the portion of the Brown corpus that has been semantically tagged with WordNet senses for the SemCor project (Miller, Leacock, Tengi, & Bunker, 1993). Criteria were used alongside this frequency data specifying when to use the first sense and when to leave the ambiguity untouched. Two criteria were used initially:

1. FREQ - a threshold on the frequency of the first sense
2. RATIO - a threshold ratio between the frequencies of the first sense and next most frequent sense.

At first FREQ was set at 5 and RATIO at 2. The method was then evaluated against the manually tagged sample of the Brown corpus (200,000 words of text) from which the frequency data was obtained and two small manually tagged samples from the LOB corpus (sample size of nouns 179) and the Wall Street Journal corpus (size 191 nouns). The results are shown in table 1. As expected the performance was superior when scored against the same data from which the frequency estimates had been taken.

Further experimentation was performed using the LOB sample and varying FREQ and RATIO. Additionally a third constraint was added (D). In this nouns identified on the SemCor project as being difficult for humans to tag were ignored.

The results are shown in table 2. Although the results indicate this is rather a limited method of disambiguating it was hoped that it would improve the selectional preference acquisition process whilst also avoiding a heavy cost in terms of human time (for manual tagging) or computer time (for unsupervised

Table 2: Variation of thresholds on the LOB data

CRITERIA	RECALL	PRECISION
freq 5 ratio 2	44	69
freq 3 ratio 2	47	69
freq 1 ratio 2	49	67
freq 3 ratio 1.5	50	67
freq 3 ratio 3	39	76
freq 3 ratio 2 (D)	45	71

training). For the selectional preference acquisition experiments 4 and 6 described below it was decided to use the criteria FREQ 3, RATIO 2 and D (ignore difficult nouns).

3.1.2 Experiment 2

Yarowksy's unsupervised algorithm (1995) was also investigated using WordNet to generate the initial seed collocations. This has the advantage that it does not rely on a quantity of handtagged data however the time taken for training remains an issue. Without optimisation the algorithm took 15 minutes of elapsed time for 710 citations of the word "plant". Accuracy was reasonable considering a) the quantity of data used (a corpus of 90 million words compared with Yarowsky's 460 million) and b) the simplifications made, imparticular the use of only one type of collocation. [2].

On initial experimentation it was evident that predominant senses quickly became favoured. For this reason the measure to order the decision list was changed from log-likelihood to a log of the ratio of association scores as show in equation 4

$$\log \frac{A(senseA_{noun}, collocation_i)}{A(other\_senses_{noun}, collocation_i)} \qquad (4)$$

Where

$$A(sense, collocation_i) = \frac{prob(sense|collocation_i)}{prob(sense)} \qquad (5)$$

This helped overcome the bias of conditional probabilities towards the most frequent sense. Recall is 71% and precision is 72% when using the log-likelihood to order the decision list with a stopping condition that the tagged portion exceeds 95% of the target data. The ratio of association scores compensates for the relative frequencies of the senses and on stopping the recall is 76% and precision is 78%

Unfortunately evaluation on the target word "plant" was rather optimistic when contrasted with an evaluation on randomly selected targets involving finer word sense distinctions. In a experiment 390 mid-frequency nouns were trained and the algorithm used to disambiguate the same nouns appearing in the SemCor files of the Brown corpus. This produced only 29% for both recall and precision which was only just better than chance. An important source of error seems to have been the poor quality of the automatically derived seeds.

On account of the training time that would be required Yarowsky's unsupervised algorithm was abandoned for the purpose of tagging the argument heads. The Wilks and Stevenson style strategy was chosen instead because it requires storage of one parameter only and is exceptionally easy to apply. A major disadvantage for this approach is that lower rank senses do not feature in the data at all. It is hoped that this will not matter where we are collecting information from many heads in a particular slot because any mistagging will be outweighed by correct taggings overall. However this approach would be unhelpful where we want to distinguish behaviour for different word senses. A potential use of Yarowky's algorithm might be verb sense distinction. The experiments outlined in the next section have been conducted using verb form rather than sense. If verbs sense distinction were to be performed it would be important to obtain the preferences for the different senses and would not be appropriate to lump the preferences together under the predominant sense. It is hoped that with some alteration to the automatic seed derivation and allowance for a coarser grained distinction this would be viable.

3.2 Acquisition of Selectional Preferences

Representation and acquisition of selectional preferences is based on Li and Abe's concept of an ATCM. The details of how such a model is acquired from corpus data using WordNet and the MDL prin-

[2]The only collocation used was within a window of 10 words either side of the target. Other simplifications include the use of a constant for smoothing, a rudimentary stopping condition, no use of the one sense per discourse strategy and no alteration of the parameters at run time.

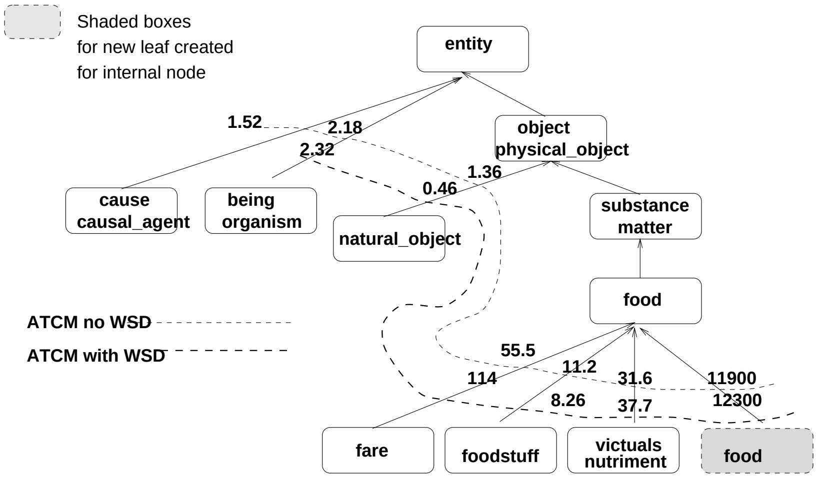

Figure 1: ATCM for 'eat' Object slot

ciple is detailed in the papers (Li & Abe, 1995; Abe & Li, 1996).

The WordNet hypernym noun hierarchy is used here as it is available and ready made. Using a resource produced by humans has its drawbacks, particularly that the classification is not tailored to the task and data at hand and is prone to the inconsistencies and errors that beset any man-made lexical resource. Still the alternative of using an automatically clustered hierarchy has other disadvantages, a particular problem being that techniques so far developed often give rise to semantically incongruous classes (Pereira, Tishby, & Lee, 1993).

Calculation of the class frequencies is key to the process of acquisition of selectional preferences. Li and Abe estimate class frequencies by dividing the frequencies of nouns occurring in the set of synonyms of a class between all the classes in which they appear. Class frequencies are then inherited up the hierarchy. In order to keep to their definition of a "tree cut" all nouns in the hierarchy need to be positioned at leaves. WordNet does not adhere to this stipulation and so they prune the hierarchy at classes where a noun featured in the set of synonyms has occurred in the data. This strategy was abandoned in the work described here because some words in the data belonged at root classes. For example in the direct object of "build" one instance of the word "entity" occurred which appears at one of the roots in WordNet. If the tree were pruned at the "ENTITY" class there would be no possibility for the preference of "build" to distinguish between the subclasses "LIFE FORM" and "PHYSICAL OBJECT".

As an alternative strategy in this work, new leaf classes were created for every internal class in the WordNet hierarchy so that terminals only occurred at leaves but the detail of WordNet was left intact. Li and Abe's strategy of pruning at classes less than a threshold of 0.1 was adhered to as this not only avoided noise but also reduced the search space.

3.2.1 Experiments 3 and 4

The input data was produced by the system described in (Briscoe & Carroll, 1997) and comprised 2 million words of parsed text with argument heads and subcategorisation frames identified. Only argument heads consisting of common nouns, days of the week and months and personal pronouns with the exception of "it" were used. The personal pronouns were all tagged with the "SOMEONE" class which is unambiguous in WordNet. Selectional preferences were acquired for a handful of verbs using either subject or object position. In experiment 3 class frequencies were calculated in much the same way as in Li and Abe's original experiments, dividing frequencies for each noun between the set of classes in which they featured as synonyms. In experiment 4 the nouns in the target slots were disambiguated using the approach outlined in experiment 1. Where frequency data was not available for the target word the word was simply treated as ambiguous and class frequencies were calculated as in experiment 3.

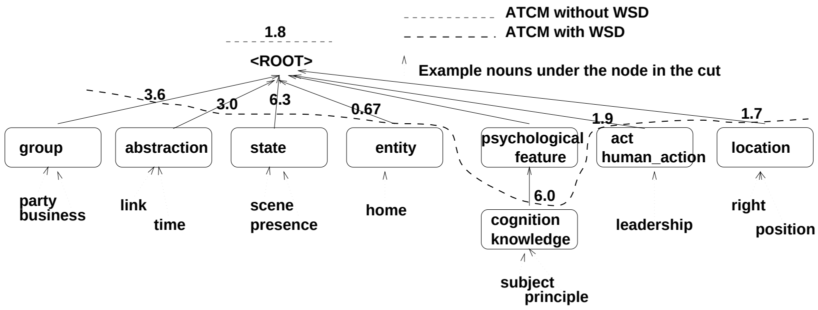

Figure 2: ATCM for 'establish' Object slot

Since ATCMs have only been obtained for the subject and object slot and for 10 target verbs no formal evaluation has been performed as yet. Instead the ATCMs were examined and some observations are given below along with diagrams showing some of the models obtained. For clarity only some of the nodes have been shown and classes are labelled with some of the synonyms belonging to that class in WordNet.

In order to obtain the ATCMs "tree cut models" (TCMs) for the target slot, irrespective of verb are obtained. A TCM is similar to an ATCM except that the scores associated with each class on the cut are probabilities and should sum to 1. The TCMs obtained for a given slot with and without WSD were similar.

In contrast ATCMs are produced with a small data set specific to the target verb. The verbs in our target set having between 32 ('clean') and 2176 ('make') instances. Because of this the noise from erroneous senses is not as easily filtered and WSD does seem to make a difference although this depends on the verb and the degree of polysemy of the most common arguments.

"Eat" is a verb which selects strongly for its object slot. The ATCMs with and without WSD are pictured in figure 1. The ATCMs are similar but WSD gives slightly stronger scores to the appropriate nodes. Additionally the NATURAL OBJECT class changes from a slight preference in the ATCM without WSD to a score below 1 (indicating no evidence for a preference) with WSD. WSD appears to slightly improve the preferences acquired but the difference is small. The reasons are that there is a predominant sense of "eat" which selects strongly for its direct object and many of the heads in the data were monosemous (e.g. food, sandwich and pretzel).

In contrast "establish" only has 79 instances and without any WSD the ATCM consists of the root node with a score of 1.8. This shows that without WSD the variety of erroneous senses causes gross over-generalisation when compared to the cut with WSD as pictured in figure 2. There are cases where the WSD is faulty and many heads are not covered by the criteria outlined in experiment 1. The head "right" for example contributes to a higher association score at the LOCATION node though its correct sense really falls under the ABSTRACTION node. However even with these inadequacies the cut with WSD appears to provide a reasonable set of preferences as opposed to the cut at the root node which is uninformative.

There was no distinction of verb senses for the preferences acquired and the data and ATCM for "serve" highlights this. "Serve" has a number of senses including the sense of "meet the needs of" or "set food on the table" or "undergo a due period". The heads in direct object position could on the whole be identified as belonging to one or other of these senses. The ATCM with WSD is illustrated in figure 3 The GROUP, RELATION and MENTAL OBJECT nodes relate to the first sense, the SUBSTANCE to the second and the third sense to the STATE and RELATION nodes. The ATCM without WSD was again an uninformative cut at the root. Ideally preferences should be acquired respective to verb sense otherwise the preferences for the different predicates will be confused.

Although formal evaluation has as yet to be performed the models examined so far with the crude WSD seem to improve on those without. This is especially so in cases of sparse data.

Some errors were due to the parser. For example

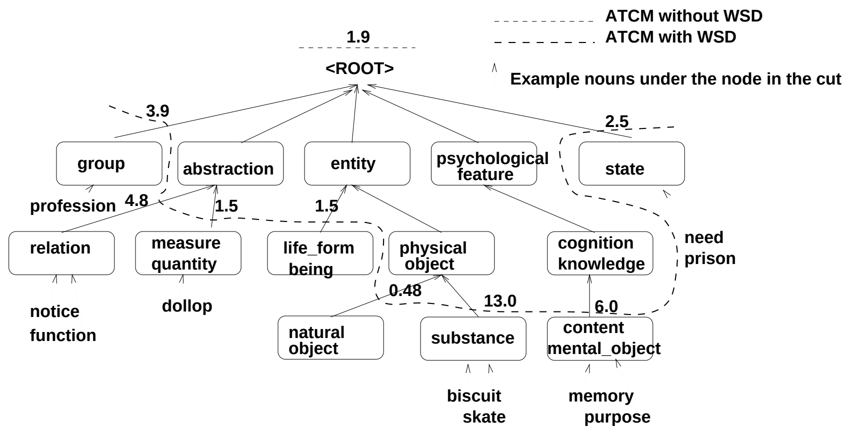

Figure 3: ATCM for 'serve' Object slot

time adverbials such as "the night before" were mistaken as direct objects when the parser failed to identify the passive as in :-

"... presented a lamb, killed the night before".

Errors also arose because collocations such as "post office" were not recognised as such. Despite these errors the advantages of using automatic parsing are significant in terms of the quantity of data thereby made available and portability to new domains.

3.3 Word Sense Disambiguation using Selection Preferences

The tree cuts obtained in experiments 3 and 4 have been used for WSD in a bootstrapping approach where heads, disambiguated by selectional preferences, are then used as input data to the preference acquisition system. WSD using the ATCMs simply selects all senses for a noun that fall under the node in the cut with the highest association score with senses for this word. For example the sense of "chicken" under VICTUALS would be preferred over the senses under LIFE FORM when occurring as the direct object of "eat". The granularity of the WSD depends on how specific the cut is. The approach has not been evaluated formally although we have plans to so with SemCor. A small evaluation has been performed comparing the manually tagged direct objects of "eat" with those selected using the cuts from experiment 3. The coarse tagging is considered correct when the identified senses contain the manually selected one. This provides a recall of 62% and precision of 93% which can be compared to a baseline precision of 55% which is calculated as in equation 6

$$\frac{\sum_{n \epsilon Heads} \frac{Number\_Senses_n Under\_Cut}{Number\_Senses_n}}{Total\_Heads\_Covered} \tag{6}$$

Naturally this approach will work better for verbs which select more strongly for their arguments.

Further experiments have been conducted which feed the disambiguated heads back into the selectional preference acquisition system.

3.3.1 Experiments 5 and 6

In experiment 5 cuts obtained in experiment 3, without any initial WSD, are used to disambiguate the heads before these are then fed back in. In contrast experiment 6 uses the cuts obtained with Wilks and Stevenson style WSD from experiment 4 to disambiguate the heads. In both cases the cuts are only used to disambiguate the heads appearing with the target verb and the full data sample required for the prior distribution TCM is left as in experiments 3 and 4.

Where the verb selects strongly for its arguments, for example "eat", the cuts obtained in experiments 5 and 6 were similar to those achieved with initial Wilks and Stevenson WSD, for example both have the effect of taking the class NATURAL OBJECT below 1 (i.e. removing the weak indication of a preference).

In contrast where the quantity of data is sparse and the verb selects less strongly the cut obtained from fully ambiguous data (experiment 5) is unhelpful for WSD. However if the Wilks and Stevenson style disambiguation is performed on the initial data the cuts in experiment 6 show great improvement on those from experiment 4. For example the ATCM in experiment 6 for "establish" showed no preferences for the LOCATION and POSSESSION nodes where preferences in experiment 4 had arisen because of erroneous word senses.

4 Conclusions

From inspection of the ATCMs obtained so far it appears that even crude WSD does help the selectional preference acquisition especially in cases of sparse data, however this still needs formal evaluation to verify whether the difference is significant.

WSD is particularly useful when the quantity of data is small as is the case when collecting data for a specific predicate. WSD selecting the most frequent sense regardless of context certainly seems to help overall despite mistakes. The preferences are improved still further if an iterative approach is taken and the preferences produced with initial WSD are used to disambiguate the heads which can then be fed back into the preference acquisition system. This has the effect of removing preferences caused by erroneous senses.

So far experiments using Yarowsky's unsupervised algorithm take too long for training each word to produce semantic tagging of sufficient quantity of text for preference acquisition but may be useful for disambiguation of target verbs, particularly with adaptations to allow a coarser granularity than the exact WordNet sense.

5 Future

The importance of word sense disambiguation on the input data needs to be subjected to formal evaluation.

Undoubtedly different underlying semantic roles will occur in a specified argument slot and this will confuse the issue. It would be interesting to examine the preferences acquired where the data used is specific to the subcategorisation frame as well as to the argument slot. Naturally this can only be done if we have sufficient data to start with.

Where there is insufficient data for a target verb it may be worth merging the data for similar verbs. The WordNet verb hierarchies might provide a useful classification for this purpose.

Although WordNet is used here, the classification could easily be changed. It would be interesting to compare results from a re-implementation using an alternative hierarchy more in tune with the corpus data. The hierarchy could be produced by humans or automatically.

The encoding of the model and data for the MDL principle needs attention as this will affect the level of generalisation. As yet the description length has assumed a tree rather than a DAG and it is apparent that cuts at nodes with shared daughters will be penalised in the current scheme.

Reference

Abe, N., & Li, H. (1996). Learning word association norms using tree cut pair models.. Unpublished. cmp-lg/9605029.

Beckwith, R., Felbaum, C., Gross, D., & Miller, G. A. (1991). WordNet: A lexical database organised on psycholinguistic principles. In Zernik, U. (Ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pp. 211–232. Lawrence Erlbaum Associates., Hillsdale NJ.

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

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

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

Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguists.*, pp. 183–190.

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

Resnik, P. (1993b). Semantic classes and syntactic ambiguity. In *Proceedings of the ARPA Workshop on Human Language Technology.*, pp. 278–283. Morgan Kaufman.

Ribas, F. (1994). Learning more appropriate selectional restrictions. Tech. rep. 41, ACQUILEX-II.

Ribas, F. (1995). 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.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods.. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguists.*, pp. 189–196.

Zernik, U. (1991). Introduction.. In Zernik, U. (Ed.), *Lexical Acquisition : Exploiting On-Line Resources to Build a Lexicon.*, pp. 1–26. Lawrence Erlbaum Associates., Hillsdale NJ.