Selectional Preference and Sense Disambiguation

Philip Resnik
Department of Linguistics and
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742 USA
resnik@umiacs.umd.edu

Abstract

The absence of training data is a real problem for corpus-based approaches to sense disambiguation, one that is unlikely to be solved soon. Selectional preference is traditionally connected with sense ambiguity; this paper explores how a statistical model of selectional preference, requiring neither manual annotation of selection restrictions nor supervised training, can be used in sense disambiguation.

1 Introduction

It has long been observed that selectional constraints and word sense disambiguation are closely linked. Indeed, the exemplar for sense disambiguation in most computational settings (e.g., see Allen's (1995) discussion) is Katz and Fodor's (1964) use of Boolean selection restrictions to constrain semantic interpretation. For example, although *burgundy* can be interpreted as either a color or a beverage, only the latter sense is available in the context of *Mary drank burgundy*, because the verb *drink* specifies the selection restriction +LIQUID for its direct objects.

Problems with this approach arise, however, as soon as the domain of interest becomes too large or too rich to specify semantic features and selection restrictions accurately by hand. This paper concerns the use of selectional constraints for automatic sense disambiguation in such broad-coverage settings. The approach combines statistical and knowledge-based methods, but unlike many recent corpus-based approaches to sense disambiguation (Yarowsky, 1993; Bruce and Wiebe, 1994; Miller et al., 1994), it takes as its starting point the assumption that sense-annotated training text is *not* available. Motivating this assumption is not only the limited availability of such text at present, but skepticism that the situation will change any time soon. In marked contrast to annotated training material for part-of-speech tagging, (a) there is no coarse-level set of sense distinctions widely agreed upon (whereas part-of-speech tag sets tend to differ in the details); (b) sense annotation has a comparatively high error rate (Miller, personal communication, reports an upper bound for human annotators of around 90% for ambiguous cases, using a non-blind evaluation method that may make even this estimate overly optimistic); and (c) no fully automatic method provides high enough quality output to support the "annotate automatically, correct manually" methodology used to provide high volume annotation by data providers like the Penn Treebank project (Marcus et al., 1993).

2 Selectional Preference as Statistical Association

The treatment of selectional preference used here is that proposed by Resnik (1993a; 1996), combining statistical and knowledge-based methods. The basis of the approach is a probabilistic model capturing the co-occurrence behavior of predicates and conceptual classes in the taxonomy. The intuition is illustrated in Figure 1. The *prior* distribution $\Pr_R(c)$ captures the probability of a class occurring as the argument in predicate-argument relation R, regardless of the identity of the predicate. For example, given the verb-subject relationship, the prior probability for ⟨person⟩ tends to be significantly higher than the prior probability for ⟨insect⟩. However, once the identity of the predicate is taken into account, the probabilities can change — if the verb is *buzz*, then the probability for ⟨insect⟩ can be expected to be higher than its prior, and ⟨person⟩ will likely be lower. In probabilistic terms, it is the difference between this conditional or *posterior* distribution and the prior distribution that determines selectional preference.

Information theory provides an appropriate way to quantify the difference between the prior and posterior distributions, in the form of relative entropy (Kullback and Leibler, 1951). The model defines the *selectional preference strength* of a predicate as:

$$\begin{aligned} S_R(p) &= \mathrm{D}(\Pr(c|p) \parallel \Pr(c)) \\ &= \sum_c \Pr(c|p) \log \frac{\Pr(c|p)}{\Pr(c)}. \end{aligned}$$

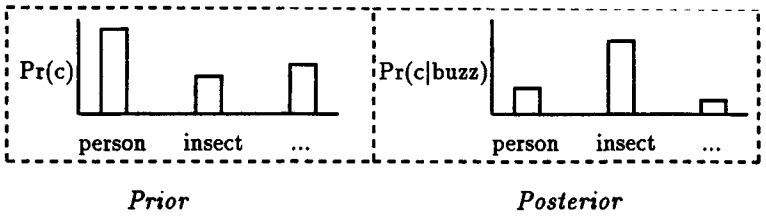

Figure 1: Prior and posterior distributions over argument classes.

Intuitively, $S_R(p)$ measures how much information, in bits, predicate p provides about the conceptual class of its argument. The better $\Pr(c)$ approximates $\Pr(c|p)$, the less influence p is having on its argument, and therefore the less strong its selectional preference.

Given this definition, a natural way to characterize the "semantic fit" of a particular class as the argument to a predicate is by its relative contribution to the overall selectional preference strength. In particular, classes that fit very well can be expected to have higher posterior probabilities, compared to their priors, as is the case for ⟨insect⟩ in Figure 1. Formally, *selectional association* is defined as:

$$A_R(p,c) = \frac{1}{S_R(p)} \Pr(c|p) \log \frac{\Pr(c|p)}{\Pr(c)}.$$

This model of selectional preference has turned out to make reasonable predictions about human judgments of argument plausibility obtained by psycholinguistic methods (Resnik, 1993a). Closely related proposals have been applied in syntactic disambiguation (Resnik, 1993b; Lauer, 1994) and to automatic acquisition of more KatzFodoresque selection restrictions in the form of weighted disjunctions (Ribas, 1994). The selectional association has also been used recently to explore apparent cases of syntactic optionality (Paola Merlo, personal communication).

3 Estimation Issues

If taxonomic classes were labeled explicitly in a training corpus, estimation of probabilities in the model would be fairly straightforward. But since text corpora contain words, not classes, it is necessary to treat each occurrence of a word in an argument position as if it might represent *any* of the conceptual classes to which it belongs, and assign frequency counts accordingly. At present, this is done by distributing the "credit" for an observation uniformly across all the conceptual classes containing an observed argument. Formally, given a predicate-argument relationship R (for example, the verb-object relationship), a predicate p, and a conceptual class c,

$$\text{freq}_R(p,c) \approx \sum_{w \in c} \frac{\text{count}_R(p,w)}{\text{classes}(w)},$$

where $\text{count}_R(p,w)$ is the number of times word w was observed as the argument of p with respect to R, and classes(w) is the number of taxonomic classes to which w belongs. Given the frequencies, probabilities are currently estimated using maximum likelihood; the use of word classes is itself a form of smoothing (cf. Pereira et al. (1993)).[1]

This estimation method is similar to that used by Yarowsky (1992) for Roget's thesaurus categories, and works for similar reasons. As an example, consider two instances of the verb-object relationship in a training corpus, *drink coffee* and *drink wine*. *Coffee* has 2 senses in the WordNet 1.4 noun taxonomy, and belongs to 13 classes in all, and *wine* has 2 senses and belongs to a total of 16 classes. This means that the observed $\text{count}_{\text{verb-obj}}(\textit{drink}, \textit{coffee}) = 1$ will be distributed by adding $\frac{1}{13}$ to the joint frequency with *drink* for each of the 13 classes containing *coffee*. Similarly, the joint frequency with *drink* will be incremented by $\frac{1}{16}$ for each of the 16 classes containing *wine*. Crucially, although each of the two words is ambiguous, only those taxonomic classes containing *both* words — e.g., ⟨beverage⟩ — receive credit for both observed instances. In general, because different words are ambiguous in different ways, credit tends to accumulate in the taxonomy only in those classes for which there is real evidence of co-occurrence; the rest tends to disperse unsystematically, resulting primarily in noise. Thus, despite the absence of class annotation in the training text, it is still possible to arrive at a usable estimate of class-based probabilities.

4 An Unsupervised Method for Sense Disambiguation

Table 1 presents a selected sample of Resnik's (1993a) comparison with argument plausibility judgments made by human subjects. What is most interesting here is the way in which strongly selecting

[1] Word w is typically the head of a noun phrase, which could lead the model astray — for example, *toy soldiers* behave differently from *soldiers* (McCawley, 1968). In principle, addressing this issue requires that noun phrases be mapped to taxonomic classes based on their compositional interpretation; however, such complications rarely arise in practice.

Verb	Object	A(*Verb*, *Object*)	Class
write	letter	7.26	⟨writing⟩
read	article	6.80	⟨writing⟩
warn	driver	4.73	⟨person⟩
hear	story	1.89	⟨communication⟩
remember	reply	1.31	⟨statement⟩
expect	visit	0.59	⟨act⟩

Table 1: Selectional ratings for plausible objects

verbs "choose" the sense of their arguments. For example, *letter* has 3 senses in WordNet,[2] and belongs to 19 classes in all. In order to approximate its plausibility as the object of *write*, the selectional association with *write* was computed for all 19 classes, and the highest value returned — in this case, ⟨writing⟩ ("anything expressed in letters; reading matter"). Since only one sense of *letter* has this class as an ancestor, this method of determining argument plausibility has, in essence, performed sense disambiguation as a side effect.

This observation suggests the following simple algorithm for disambiguation by selectional preference. Let n be a noun that stands in relationship R to predicate p, and let $\{s_1, ..., s_k\}$ be its possible senses. For i from 1 to k, compute:

$$C_i = \{c \mid c \text{ is an ancestor of } s_i\}$$
$$a_i = \max_{c \in C_i} A_R(p, c)$$

and assign a_i as the score for sense s_i. The simplest way to use the resulting scores, following Miller et al. (1994), is as follows: if n has only one sense, select it; otherwise select the sense s_i for which a_i is greatest, breaking ties by random choice.

5 Evaluation

Task and materials. Test and training materials were derived from the Brown corpus of American English, all of which has been parsed and manually verified by the Penn Treebank project (Marcus et al., 1993) and parts of which have been manually sense-tagged by the WordNet group (Miller et al., 1993). A parsed, sense-tagged corpus was obtained by merging the WordNet sense-tagged corpus (approximately 200,000 words of source text from the Brown corpus, distributed across genres) with the corresponding Penn Treebank parses.[3] The rest of the Brown corpus (approximately 800,000 words of source text) remained as a parsed, but not sense-tagged, training set.

The test set for the verb-object relationship was constructed by first training a selectional preference model on the training corpus, using the Treebank's `tgrep` utility to extract verb-object pairs from parse trees. The 100 verbs that select most strongly for their objects were identified, excluding verbs appearing only once in the training corpus; test instances of the form (*verb*, *object*, *correct sense*) were then extracted from the merged test corpus, including all triples where *verb* was one of the 100 test verbs.[4]

Evaluation materials were obtained in the same manner for several other surface syntactic relationships, including verb-subject (John ⇐ admires), adjective-noun (tall ⇒ building), modifier-head (river ⇒ bank), and head-modifier (river ⇐ bank).

Baseline. Following Miller et al. (1994), disambiguation by random choice was used as a baseline: if a noun has one sense, use it; otherwise select at random among its senses.

Results. Since both the algorithm and the baseline may involve random choices, evaluation involved multiple runs with different random seeds. Table 2 summarizes the results, taken over 10 runs, considering *only* ambiguous test cases. All differences between the means for algorithm and baseline were statistically significant.

Discussion. The results of the experiment show that disambiguation using automatically acquired selectional constraints leads to performance significantly better than random choice. Not surprisingly, though, the results are far from what one might expect to obtain with supervised training. In that respect, the most direct point of comparison is the performance of Miller et al.'s (1994) frequency heuristic — always choose the most frequent sense of a word — as evaluated using the full sense-tagged corpus, including nouns, verbs, adjectives, and adverbs. For ambiguous words, they report 58.2% correct, as compared to a random baseline of 26.8%.

Crucially, however, the frequency heuristic requires sense-tagged training data (Miller et al. evaluated via cross-validation), and this paper starts from the assumption that such data are unavailable. A fairer comparison, therefore, considers al-

[2](1) Written message, (2) varsity letter, (3) alphabetic character.

[3]The merge was mostly automatic, requiring manual intervention for only 3 of 103 files.

[4]Excluded were some inapplicable cases, e.g. where *object* was a proper noun tagged as ⟨person⟩.

Relationship		% Correct			
		mean	σ	min	max
verb-object	(baseline)	*28.5*	*5.91*	*18.0*	*36.0*
	(sel. pref.)	44.3	4.90	36.0	51.0
verb-subject	(baseline)	*29.1*	*5.23*	*20.0*	*38.0*
	(sel. pref.)	40.8	2.86	36.0	44.0
head-modifier	(baseline)	*32.8*	*7.00*	*23.0*	*44.0*
	(sel. pref.)	40.2	5.99	33.0	51.0
modifier-head	(baseline)	*30.8*	*6.25*	*24.0*	*40.0*
	(sel. pref.)	39.9	2.60	35.0	43.0
adjective-noun	(baseline)	*29.1*	*8.40*	*16.0*	*38.0*
	(sel. pref.)	35.3	3.95	31.0	40.0

Table 2: Experimental results

ternative unsupervised algorithms — though unfortunately the literature contains more proposed algorithms than quantitative evaluations of those algorithms. One experiment where results were reported was conducted by Cowie et al. (1992); their method involved using a stochastic search procedure to maximize the overlap in dictionary definitions (LDOCE) for alternative senses of words co-occurring in a sentence. They report an accuracy of 72% for disambiguation to the homograph level, and 47% for disambiguation to the sense level. Since the task here involved WordNet sense distinctions, which are rather fine grained, the latter value is more appropriate for comparison. Their experiment was more general in that they did not restrict themselves to nouns; on the other hand, their test set involved disambiguating words taken from full sentences, so the percentage correct may have been improved by the presence of unambiguous words.

Sussna (1993) has also looked at unsupervised disambiguation of nouns using WordNet. Like Cowie et al., his algorithm optimizes a measure of semantic coherence over an entire sentence, in this case pairwise semantic distance between nouns in the sentence as measured using the noun taxonomy. Comparison of results is somewhat difficult, however, for two reasons. First, Sussna used an earlier version of WordNet (version 1.2) having a significantly smaller noun taxonomy (35K nodes vs. 49K nodes). Second, and more significant, in creating the test data, Sussna's human sense-taggers (tagging articles from the *Time* IR test collection) were permitted to tag a noun with as many senses as they felt were "good," rather than making a forced choice; Sussna develops a scoring metric based on that fact rather than requiring exact matches to a single best sense. This is quite a reasonable move (see discussion below), but unfortunately not an option in the present experiment. Nonetheless, some comparison is possible, since he reports a "% correct," apparently treating a sense assignment as correct if any of the "good" senses is chosen — his experiments have a lower bound (chance) of about 40% correct, with his algorithm performing at 53–55%, considering only ambiguous cases.

The best results reported for an unsupervised sense disambiguation method are those of Yarowsky (1992), who uses evidence from a wider context (a window of 100 surrounding words) to build up a co-occurrence model using classes from Roget's thesaurus. He reports accuracy figures in the 72–99% range (mean 92%) in disambiguating test instances involving twelve "interesting" polysemous words. As in the experiments by Cowie et al., the choice of coarser distinctions presumably accounts in part for the high accuracy. By way of comparison, some words in Yarowsky's test set would require choosing among ten senses in WordNet, as compared to a maximum of six using the Roget's thesaurus categories; the mean level of polysemy for the tested words is a six-way distinction in WordNet as compared to a three-way distinction in Roget's thesaurus.

As an aside, a rich taxonomy like WordNet permits a more continuous view of the sense vs. homograph distinction. For example, *town* has three senses in WordNet, corresponding to an administrative district, a geographical area, and a group of people. Given *town* as the object of *leave*, selectional preference will produce a tie between the first two senses, since both inherit their score from a common ancestor, ⟨`location`⟩. In effect, the automatic selection of a class higher in the taxonomy as having the highest score provides the same coarse category that might be provided by a homograph/sense distinction in another setting. The choice of coarser category varies dynamically with the context: as the argument in *rural town*, the same two senses still tie, but with ⟨`region`⟩ (a subclass of ⟨`location`⟩) as the common ancestor that determines the score.

In other work, Yarowsky (1993) has shown that local collocational information, including selectional constraints, can be used to great effect in sense disambiguation, though his algorithm requires super-

vised training. The present work can be viewed as an attempt to take advantage of the same kind of information, but in an unsupervised setting.

6 Conclusions and Future Work

Although the definition of selectional preference strength is motivated by the use of relative entropy in information theory, selectional association is not; the approach would benefit from experimentation with alternative statistical association measures, particularly a comparison with simple mutual information and with the likelihood ratio. Combining information about selectional preference could also be helpful, e.g., where a noun is both the object of a verb and modified by an adjective, though such cases are rarer than one might expect.

More important is information beyond selectional preference, notably the wider context utilized by Yarowsky (1992). Performance of the method explored here is limited at present, though not surprisingly so when taken in the context of previous attempts at unsupervised disambiguation using fine-grained senses. One main message to take away from this experiment is the observation that, although selectional preferences are widely viewed as an important factor in disambiguation, their practical broad-coverage application appears limited — at least when disambiguating nouns — because many verbs and modifiers simply do not select strongly enough to make a significant difference. They may provide *some* evidence, but most likely only as a complement to other sources of information such as frequency-based priors, topical context, and the like.

Acknowledgements

Much of this work was conducted at Sun Microsystems Laboratories in Chelmsford, Massachusetts.

References

James Allen. 1995. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company.

Rebecca Bruce and Janyce Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics*, Las Cruces, New Mexico, June.

Jim Cowie, Joe Guthrie, and Louise Guthrie. 1992. Lexical disambiguation using simulated annealing. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 359–365, Nantes, France, August.

J. J. Katz and J. A. Fodor. 1964. The structure of a semantic theory. In J. A. Fodor and J. J. Katz, editors, *The Structure of Language*, chapter 19, pages 479–518. Prentice Hall.

S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Mark Lauer. 1994. Conceptual association for compound noun analysis. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, June. Student Session.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.

James McCawley. 1968. The role of semantics in a grammar. In Emmon Bach and Robert Harms, editors, *Universals in Linguistic Theory*, pages 124–169. Holt, Rinehart and Winston.

George Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *ARPA Workshop on Human Language Technology*. Morgan Kaufmann, March.

George Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert Thomas. 1994. Using a semantic concordance for sense identification. In *ARPA Workshop on human Language Technology*, Plainsboro, NJ, March.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93)*, Morristown, New Jersey, June. Association for Computational Linguistics.

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

Philip Resnik. 1993b. Semantic classes and syntactic ambiguity. In *Proceedings of the 1993 ARPA Human Language Technology Workshop*. Morgan Kaufmann, March.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159.

Francesc Ribas. 1994. An experiment on learning appropriate selectional restrictions from a parsed corpus. In *Proceedings of COLING 1994*.

Michael Sussna. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management (CIKM-93)*, Arlington, Virginia.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 454–460, Nantes, France, July.

David Yarowsky. 1993. One sense per collocation. ARPA Workshop on Human Language Technology, March. Princeton.