Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences

Diana McCarthy*
University of Sussex

John Carroll*
University of Sussex

Selectional preferences have been used by word sense disambiguation (WSD) systems as one source of disambiguating information. We evaluate WSD using selectional preferences acquired for English adjective–noun, subject, and direct object grammatical relationships with respect to a standard test corpus. The selectional preferences are specific to verb or adjective classes, rather than individual word forms, so they can be used to disambiguate the co-occurring adjectives and verbs, rather than just the nominal argument heads. We also investigate use of the one-sense-per-discourse heuristic to propagate a sense tag for a word to other occurrences of the same word within the current document in order to increase coverage. Although the preferences perform well in comparison with other unsupervised WSD systems on the same corpus, the results show that for many applications, further knowledge sources would be required to achieve an adequate level of accuracy and coverage. In addition to quantifying performance, we analyze the results to investigate the situations in which the selectional preferences achieve the best precision and in which the one-sense-per-discourse heuristic increases performance.

1. Introduction

Although selectional preferences are a possible knowledge source in an automatic word sense disambiguation (WDS) system, they are not a panacea. One problem is coverage: Most previous work has focused on acquiring selectional preferences for verbs and applying them to disambiguate nouns occurring at subject and direct object slots (Ribas 1995; McCarthy 1997; Abney and Light 1999; Ciaramita and Johnson 2000; Stevenson and Wilks 2001). In normal running text, however, a large proportion of word tokens do not fall at these slots. There has been some work looking at other slots (Resnik 1997), and on using nominal arguments as disambiguators for verbs (Federici, Montemagni, and Pirrelli 1999; Agirre and Martinez 2001), but the problem of coverage remains. Selectional preferences can be used for WSD in combination with other knowledge sources (Stevenson and Wilks 2001), but there is a need to ascertain when they work well so that they can be utilized to their full advantage. This article is aimed at quantifying the disambiguation performance of automatically acquired selectional preferences in regard to nouns, verbs, and adjectives with respect to a standard test corpus and evaluation setup (SENSEVAL-2) and to identify strengths and weaknesses. Although there is clearly a limit to coverage using preferences alone, because preferences are acquired only with respect to specific grammatical roles, we show that when dealing with running text, rather than isolated examples, coverage can be increased at little cost in accuracy by using the one-sense-per-discourse heuristic.

* Department of Informatics, University of Sussex, Brighton BN1 9QH, UK. E-mail: {dianam, johnca}@sussex.ac.uk

We acquire selectional preferences as probability distributions over the WordNet (Fellbaum 1998) noun hyponym hierarchy. The probability distributions are conditioned on a verb or adjective class and a grammatical relationship. A noun is disambiguated by using the preferences to give probability estimates for each of its senses in WordNet, that is, for WordNet synsets. Verbs and adjectives are disambiguated by using the probability distributions and Bayes' rule to obtain an estimate of the probability of the adjective or verb class, given the noun and the grammatical relationship. Previously, we evaluated noun and verb disambiguation on the English all-words task in the SENSEVAL-2 exercise (Cotton et al. 2001). We now present results also using preferences for adjectives, again evaluated on the SENSEVAL-2 test corpus (but carried out after the formal evaluation deadline). The results are encouraging, given that this method does not rely for training on any hand-tagged data or frequency distributions derived from such data. Although a modest amount of English sense-tagged data is available, we nevertheless believe it is important to investigate methods that do not require such data, because there will be languages or texts for which sense-tagged data for a given word is not available or relevant.

2. Motivation

The goal of this article is to assess the WSD performance of selectional preference models for adjectives, verbs, and nouns on the SENSEVAL-2 test corpus. There are two applications for WSD that we have in mind and are directing our research. The first application is text simplification, as outlined by Carroll, Minnen, Pearce et al. (1999). One subtask in this application involves substituting words with thier more frequent synonyms, for example, substituting *letter* for *missive*. Our motivation for using WSD is to filter out inappropriate senses of a word token, so that the substituting synonym is appropriate given the context. For example, in the following sentence we would like to use *strategy*, rather than *dodge*, as a substitute for *scheme*:

A recent government study singled out the scheme as an example to others.

We are also investigating the disambiguation of verb senses in running text before subcategorization information for the verbs is acquired, in order to produce a subcategorization lexicon specific to sense (Preiss and Korhonen 2002). For example, if subcategorization were acquired specific to sense, rather than verb form, then distinct senses of *fire* could have different subcategorization entries:

fire(1) - **sack**:	`NP V NP`
fire(2) - **shoot**:	`NP V NP, NP V`

Selectional preferences could also then be acquired automatically from sense-tagged data in an iterative approach (McCarthy 2001).

3. Methodology

We acquire selectional preferences from automatically preprocessed and parsed text during a training phase. The parser is applied to the test data as well in the runtime phase to identify grammatical relations among nouns, verbs, and adjectives. The acquired selectional preferences are then applied to the noun-verb and noun-adjective pairs in these grammatical constructions for disambiguation.

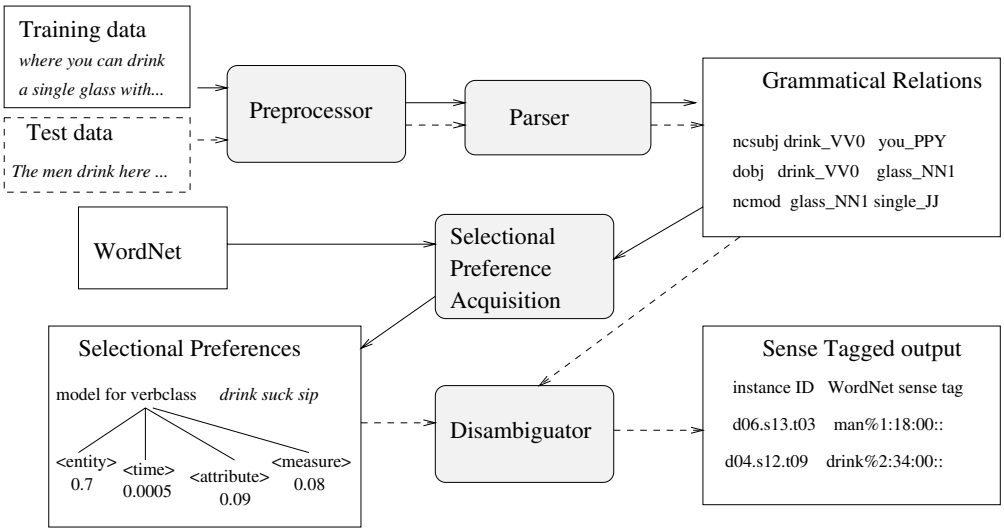

Figure 1
System overview. Solid lines indicate flow of data during training, and broken lines show that at run time.

The overall structure of the system is illustrated in Figure 1. We describe the individual components in sections 3.1–3.3 and 4.

3.1 Preprocessing

The preprocessor consists of three modules applied in sequence: a tokenizer, a part-of-speech (POS) tagger, and a lemmatizer.

The tokenizer comprises a small set of manually developed finite-state rules for identifying word and sentence boundaries. The tagger (Elworthy 1994) uses a bigram hidden Markov model augmented with a statistical unknown word guesser. When applied to the training data for selectional preference acquisition, it produces the single highest-ranked POS tag for each word. In the run-time phase, it returns multiple tag hypotheses, each with an associated forward-backward probability to reduce the impact of tagging errors. The lemmatizer (Minnen, Carroll, and Pearce 2001) reduces inflected verbs and nouns to their base forms. It uses a set of finite-state rules expressing morphological regularities and subregularities, together with a list of exceptions for specific (irregular) word forms.

3.2 Parsing

The parser uses a wide-coverage unification-based shallow grammar of English POS tags and punctuation (Briscoe and Carroll 1995) and performs disambiguation using a context-sensitive probabilistic model (Briscoe and Carroll 1993), recovering from extra-grammaticality by returning partial parses. The output of the parser is a set of **grammatical relations** (Carroll, Briscoe, and Sanfilippo 1998) specifying the syntactic dependency between each head and its dependent(s), taken from the phrase structure tree that is returned from the disambiguation phase.

For selectional preference acquisition we applied the analysis system to the 90 million words of the written portion of the British National Corpus (BNC); the parser produced complete analyses for around 60% of the sentences and partial analyses for over 95% of the remainder. Both in the acquisition phase and at run time, we extract from the analyser output subject–verb, verb–direct object, and noun–adjective

modifier dependencies.[1] We did not use the SENSEVAL-2 Penn Treebank–style bracketings supplied for the test data.

3.3 Selectional Preference Acquisition

The preferences are acquired for grammatical relations (subject, direct object, and adjective–noun) involving nouns and grammatically related adjectives or verbs. We use WordNet synsets to define our sense inventory. Our method exploits the hyponym links given for nouns (e.g., *cheese* is a hyponym of *food*), the troponym links for verbs [2] (e.g., *limp* is a troponym of *walk*), and the "similar-to" relationship given for adjectives (e.g., one sense of *cheap* is similar to *flimsy*).

The preference models are modifications of the tree cut models (TCMs) originally proposed by Li and Abe (1995, 1998). The main differences between that work and ours are that we acquire adjective as well as verb models, and also that our models are with respect to verb and adjective classes, rather than forms. We acquire models for classes because we are using the models for WSD, whereas Li and Abe used them for structural disambiguation.

We define a TCM as follows. Let NC be the set of noun synsets (noun classes) in WordNet: $NC = \{nc \in WordNet\}$, and NS be the set of noun senses [3] in Wordnet: $NS = \{ns \in WordNet\}$. A TCM is a set of noun classes that partition NS disjointly. We use Γ to refer to such a set of classes in a TCM. A TCM is defined by Γ and a probability distribution:

$$\sum_{nc \in \Gamma} p(nc) = 1 \tag{8}$$

The probability distribution is conditioned by the grammatical context. In this work, the probability distribution associated with a TCM is conditioned on a verb class (vc) and either the subject or direct-object relation, or an adjective class (ac) and the adjective–noun relation. Let VC be the set of verb synsets (verb classes) in WordNet: $VC = \{vc \in WordNet\}$. Let AC be the set of adjective classes (which subsume WordNet synsets; we elaborate further on this subsequently). Thus, the TCMs define a probability distribution over NS that is conditioned on a verb class (vc) or adjective class (ac) and a particular grammatical relation (gr):

$$\sum_{nc \in \Gamma} p(nc|vc, gr) = 1 \tag{9}$$

Acquisition of a *TCM* for a given vc and gr proceeds as follows. The data for acquiring the preference are obtained from a subset of the tuples involving verbs in the synset or troponym (subordinate) synsets. Not all verbs that are troponyms or direct members of the synset are used in training. We take the noun argument heads occurring with verbs that have no more than 10 senses in WordNet and a frequency of 20 or more occurrences in the BNC data in the specified grammatical relationship. The threshold of 10 senses removes some highly polysemous verbs having many sense distinctions that are rather subtle. Verbs that have more than 10 senses include very frequent verbs such as *be* and *do* that do not select strongly for their

1 In a previous evaluation of grammatical-relation accuracy with in-coverage text, the analyzer returned subject–verb and verb–direct object dependencies with 84–88% recall and precision (Carroll, Minnen, and Briscoe 1999).

2 In WordNet, verbs are organized by the troponymy relation, but this is represented with the same hyponym pointer as is used in the noun hierarchy.

3 We refer to nouns attached to synsets as noun senses.

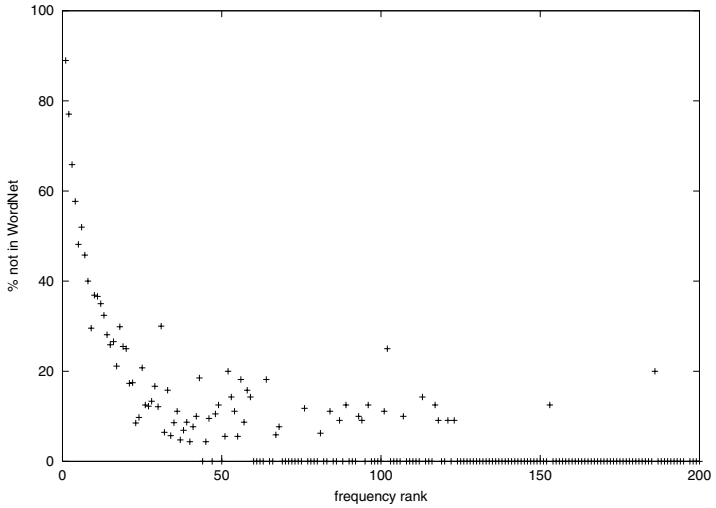

Figure 2
Verbs not in WordNet by BNC frequency.

arguments. The frequency threshold of 20 is intended to remove noisy data. We set the threshold by examining a plot of BNC frequency and the percentage of verbs at particular frequencies that are not listed in WordNet (Figure 2). Using 20 as a threshold for the subject slot results in only 5% verbs that are not found in WordNet, whereas 73% of verbs with fewer than 20 BNC occurrences are not present in WordNet.[4]

The selectional-preference models for adjective–noun relations are conditioned on an *ac*. Each *ac* comprises a group of adjective WordNet synsets linked by the "similar-to" relation. These groups are formed such that they partition all adjective synsets. Thus $AC = \{ac \in \textit{WordNet adjective synsets linked by similar-to}\}$. For example, Figure 3 shows the adjective classes that include the adjective *fundamental* and that are formed in this way.[5] For selectional-preference models conditioned on adjective classes, we use only those adjectives that have 10 synsets or less in WordNet and have 20 or more occurrences in the BNC.

The set of *nc*s in Γ are selected from all the possibilities in the hyponym hierarchy according to the minimum description length (MDL) principle (Rissanen 1978) as used by Li and Abe (1995, 1998). MDL finds the best TCM by considering the cost (in bits) of describing both the model and the argument head data encoded in the model. The cost (or description length) for a TCM is calculated according to equation (10). The number of parameters of the model is given by k, which is the number of *nc*s in Γ minus one. N is the sample of the argument head data. The cost of describing each noun argument head (n) is calculated by the log of the probability estimate for that noun:

$$\begin{aligned} \textit{description length} = \quad & \textit{model description length} \quad + \textit{data description length} \\ = \quad & \tfrac{k}{2} \times \log |N| \quad + - \sum_{n \in N} \log p(n) \end{aligned} \tag{10}$$

4 These threshold values are somewhat arbitrary, but it turns out that our results are not sensitive to the exact values.
5 For the sake of brevity, not all adjectives are included in this diagram.

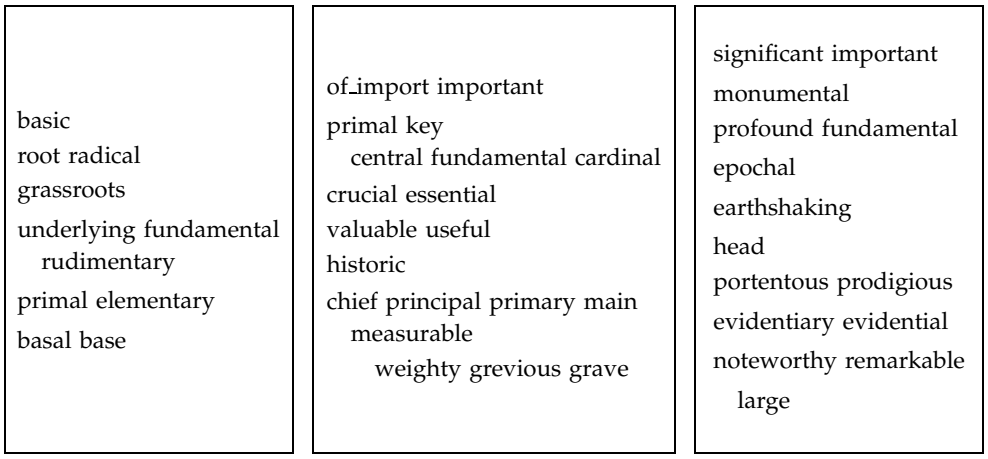

Figure 3
Adjective classes that include *fundamental*.

The probability estimate for each n is obtained using the estimates for all the nss that n has. Let C_n be the set of ncs that include n as a direct member: $C_n = \{nc \in NC | n \in nc\}$. Let nc' be a hypernym of nc on Γ (i.e. $nc' \in \{\Gamma | nc \in nc'\}$) and let $ns_{nc'} = \{ns \in nc'\}$ (i.e., the set of nouns senses at and beneath nc' in the hyponym hierarchy). Then the estimate $p(n)$ is obtained using the estimates for the hypernym classes on Γ for all the C_n that n belongs to:

$$p(n) = \sum_{nc \in C_n} \frac{p(nc')}{|ns_{nc'}|} \tag{11}$$

The probability at any particular nc' is divided by $ns_{nc'}$ to give the estimate for each $p(ns)$ under that nc'.

The probability estimates for the $\{nc \in \Gamma\}$ ($p(nc|vc, gr)$ or $p(nc|ac, gr)$) are obtained from the tuples from the data of nouns co-occurring with verbs (or adjectives) belonging to the conditioning vc (or ac) in the specified grammatical relationship ($< n, v, gr >$). The frequency credit for a tuple is divided by $|C_n|$ for any n, and by the number of synsets of v, C_v (or C_a if the gr is adjective-noun):

$$\text{freq}(nc|vc, gr) = \sum_{v \in vc} \sum_{n \in nc} \frac{\text{freq}(n|v, gr)}{|C_n||C_v|} \tag{12}$$

A hypernym nc' includes the frequency credit attributed to all its hyponyms ($\{nc \in nc'\}$).

$$\text{freq}(nc'|vc, gr) = \sum_{nc \in nc'} \text{freq}(nc|vc, gr) \tag{13}$$

This ensures that the total frequency credit at any Γ across the hyponym hierarchy equals the credit for the conditioning vc. This will be the sum of the frequency credit for all verbs that are direct members or troponyms of the vc, divided by the number of other senses of each of these verbs:

$$\text{freq}(vc|gr) = \sum_{verb \in vc} \frac{\text{freq}(verb|gr)}{|C_v|} \tag{14}$$

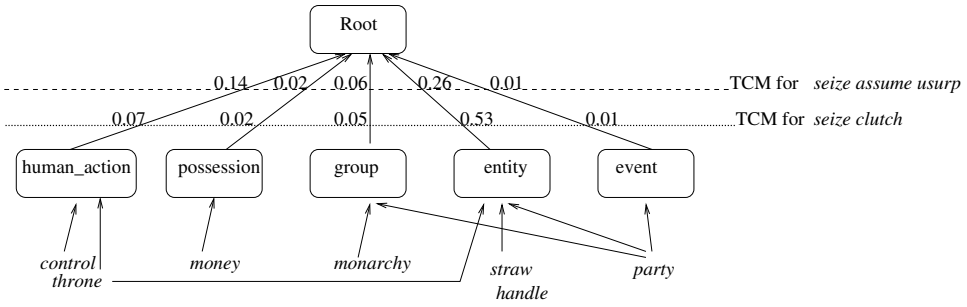

Figure 4
TCMs for the direct-object slot of two verb classes that include the verb seize.

To ensure that the TCM covers all *NS* in WordNet, we modify Li and Abe's original scheme by creating hyponym leaf classes below all WordNet's internal classes in the hyponym hierarchy. Each leaf holds the *ns* previously held at the internal class.

Figure 4 shows portions of two TCMs. The TCMs are similar, as they both contain the verb seize, but the TCM for the class that includes clutch has a higher probability for the **entity** noun class compared to the class that also includes *assume* and usurp. This example includes only top-level WordNet classes, although the TCM may use more specific noun classes.

4. Disambiguation

Nouns, adjectives and verbs are disambiguated by finding the sense (*nc*, *vc*, or *ac*) with the maximum probability estimate in the given context. The method disambiguates nouns and verbs to the WordNet synset level and adjectives to a coarse-grained level of WordNet synsets linked by the similar-to relation, as described previously.

4.1 Disambiguating Nouns

Nouns are disambiguated when they occur as subjects or direct objects and when modified by adjectives. We obtain a probability estimate for each *nc* to which the target noun belongs, using the distribution of the TCM associated with the co-occurring verb or adjective and the grammatical relationship.

Li and Abe used TCMs for the task of structural disambiguation. To obtain probability estimates for noun senses occurring at classes beneath hypernyms on the cut, Li and Abe used the probability estimate at the nc' on the cut divided by the number of *ns* descendants, as we do when finding Γ during training, so the probability estimate is shared equally among all nouns in the nc', as in equation (15).

$$p(ns \in ns_{nc'}) = \frac{p(nc')}{|ns_{nc'}|} \tag{15}$$

One problem with doing this is that in cases in which the TCM is quite high in the hierarchy, for example, at the **entity** class, the probability of any *ns*'s occurring under this nc' on the TCM will be the same and does not allow us to discriminate among senses beneath this level.

For the WSD task, we compare the probability estimates at each $nc \in C_n$, so if a noun belongs to several synsets, we compare the probability estimates, given the context, of these synsets. We obtain estimates for each *nc* by using the probability of the hypernym nc' on Γ. Rather than assume that all synsets under a given nc' on Γ

have the same likelihood of occurrence, we multiply the probability estimate for the hypernym nc' by the ratio of the prior frequency of the nc, that is, $p(nc|gr)$, for which we seek the estimate divided by the prior frequency of the hypernym nc' ($p(nc'|gr)$):

$$p(nc \in \textit{hyponyms of } nc'|vc, gr) = p(nc'|vc, gr) \times \frac{p(nc|gr)}{p(nc'|gr)} \tag{16}$$

These prior estimates are taken from populating the noun hyponym hierarchy with the prior frequency data for the gr irrespective of the co-occurring verbs. The probability at the hypernym nc' will necessarily total the probability at all hyponyms, since the frequency credit of hyponyms is propagated to hypernyms.

Thus, to disambiguate a noun occurring in a given relationship with a given verb, the $nc \in C_n$ that gives the largest estimate for $p(nc|vc, gr)$ is taken, where the verb class (vc) is that which maximizes this estimate from C_v. The TCM acquired for each vc of the verb in the given gr provides an estimate for $p(nc'|vc, gr)$, and the estimate for nc is obtained as in equation (16).

For example, one target noun was *letter*, which occurred as the direct object of *sign* in our parses of the SENSEVAL-2 data. The TCM that maximized the probability estimate for $p(nc|vc, \textit{direct object})$ is shown in Figure 5. The noun *letter* is disambiguated by comparing the probability estimates on the TCM above the five senses of *letter* multiplied by the proportion of that probability mass attributed to that synset. Although **entity** has a higher probability on the TCM, compared to **matter**, which is above the correct sense of *letter*,[6] the ratio of prior probabilities for the synset containing *letter*[7] under **entity** is 0.001, whereas that for the synset under **matter** is 0.226. This gives a probability of $0.009 \times 0.226 = 0.002$ for the noun class probability given the verb class

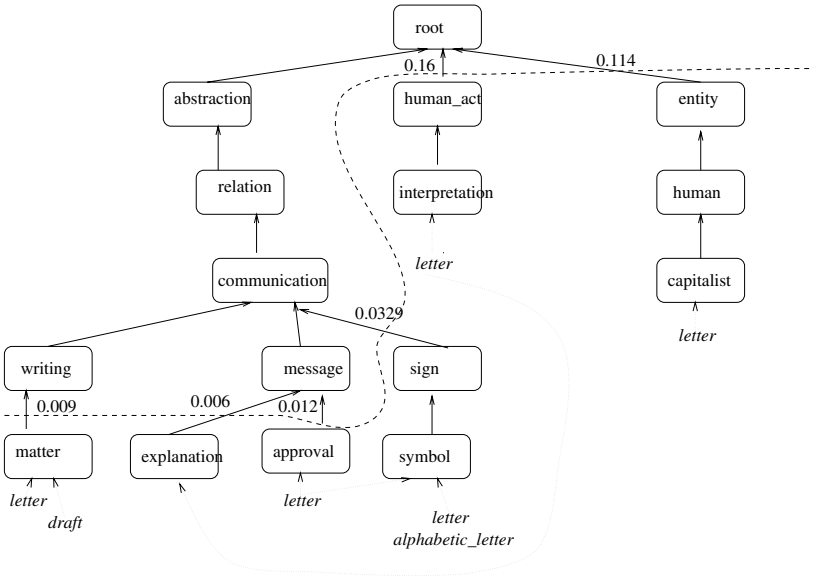

Figure 5
TCM for the direct-object slot of the verb class including *sign* and *ratify*.

6 The gloss is "a written message addressed to a person or organization; e.g. wrote an indignant letter to the editor."
7 The gloss is "owner who lets another person use something (housing usually) for hire."

(with maximum probability) and grammatical context. This is the highest probability for any of the synsets of *letter*, and so in this case the correct sense is selected.

4.2 Disambiguating Verbs and Adjectives

Verbs and adjectives are disambiguated using TCMs to give estimates for $p(nc|vc, gr)$ and $p(nc|ac, gr)$, respectively. These are combined with prior estimates for $p(nc|gr)$ and $p(vc|gr)$ (or $p(ac|gr)$) using Bayes' rule to give:

$$p(vc|nc, gr) = p(nc|vc, gr) \times \frac{p(vc|gr)}{p(nc|gr)} \tag{17}$$

and for adjective–noun relations:

$$p(ac|nc, adjnoun) = p(nc|ac, adjnoun) \times \frac{p(ac|adjnoun)}{p(nc|adjnoun)} \tag{18}$$

The prior distributions for $p(nc|gr)$, $p(vc|gr)$ and $p(ac|adjnoun)$ are obtained during the training phase. For the prior distribution over *NC*, the frequency credit of each noun in the specified *gr* in the training data is divided by $|C_n|$. The frequency credit attached to a hyponym is propagated to the superordinate hypernyms, and the frequency of a hypernym (nc') totals the frequency at its hyponyms:

$$\text{freq}(nc'|gr) = \sum_{nc \in nc'} \text{freq}(nc|gr) \tag{19}$$

The distribution over *VC* is obtained similarly using the troponym relation. For the distribution over *AC*, the frequency credit for each adjective is divided by the number of synsets to which the adjective belongs, and the credit for an *ac* is the sum over all the synsets that are members by virtue of the similar-to WordNet link.

To disambiguate a verb occurring with a given noun, the *vc* from C_v that gives the largest estimate for $p(vc|nc, gr)$ is taken. The *nc* for the co-occurring noun is the *nc* from C_n that maximizes this estimate. The estimate for $p(nc|vc, gr)$ is taken as in equation (16) but selecting the *vc* to maximize the estimate for $p(vc|nc, gr)$ rather than $p(nc|vc, gr)$. An adjective is likewise disambiguated to the *ac* from all those to which the adjective belongs, using the estimate for $p(nc|ac, gr)$ and selecting the *nc* that maximizes the $p(ac|nc, gr)$ estimate.

4.3 Increasing Coverage: One Sense per Discourse

There is a significant limitation to the word tokens that can be disambiguated using selectional preferences, in that they are restricted to those that occur in the specified grammatical relations and in argument head position. Moreover, we have TCMs only for adjective and verb classes in which there was at least one adjective or verb member that met our criteria for training (having no more than a threshold of 10 senses in WordNet and a frequency of 20 or more occurrences in the BNC data in the specified grammatical relationship). We chose not to apply TCMs for disambiguation where we did not have TCMs for one or more classes for the verb or adjective. To increase coverage, we experimented with applying the one-sense-per-discourse (OSPD) heuristic (Gale, Church, and Yarowsky 1992). With this heuristic, a sense tag for a given word is propagated to other occurrences of the same word within the current document in order to increase coverage. When applying the OSPD heuristic, we simply applied a tag for a noun, verb, or adjective to all the other instances of the same word type with the same part of speech in the discourse, provided that only one possible tag for that word was supplied by the selectional preferences for that discourse.

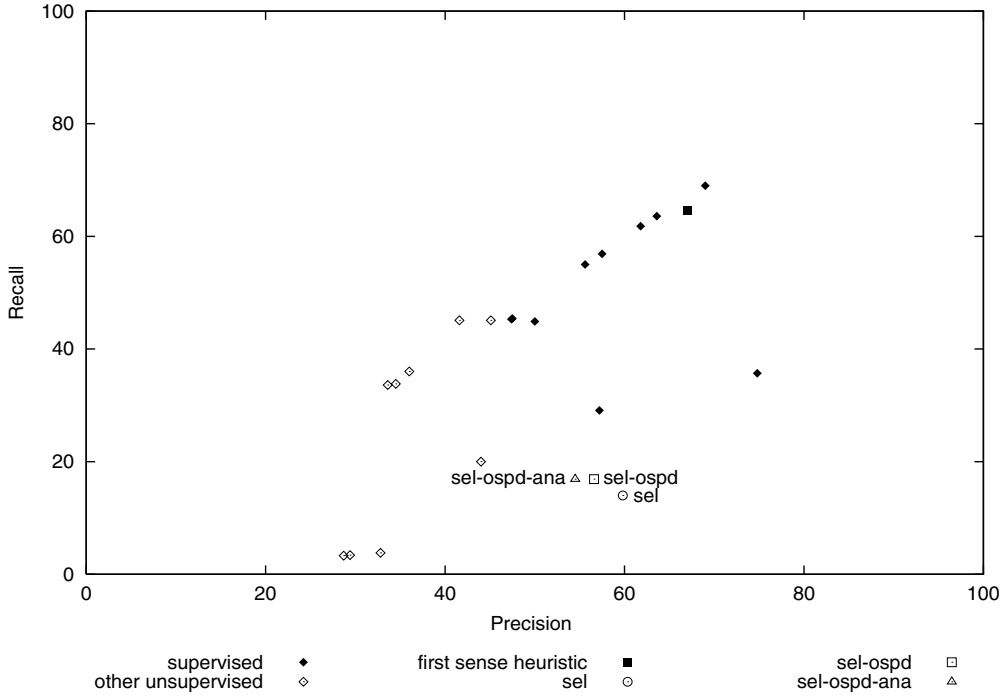

Figure 6
SENSEVAL-2 English all-words task results.

5. Evaluation

We evaluated our system using the SENSEVAL-2 test corpus on the English all-words task (Cotton et al., 2001). We entered a previous version of this system for the SENSEVAL-2 exercise, in three variants, under the names "sussex-sel" (selectional preferences), "sussex-sel-ospd" (with the OSPD heuristic), and "sussex-sel-ospd-ana" (with anaphora resolution). [8] For SENSEVAL-2 we used only the direct object and subject slots, since we had not yet dealt with adjectives. In Figure 6 we show how our system fared at the time of SENSEVAL-2 compared to other unsupervised systems.[9] We have also plotted the results of the supervised systems and the precision and recall achieved by using the most frequent sense (as listed in WordNet).[10]

In the work reported here, we attempted disambiguation for head nouns and verbs in subject and direct object relationships, and for adjectives and nouns in adjective-noun relationships. For each test instance, we applied subject preferences before direct object preferences, and direct object preferences before adjective–noun preferences. We also propagated sense tags to test instances not in these relationships by applying the one-sense-per-discourse heuristic.

We did not use the SENSEVAL-2 coarse-grained classification, as this was not available at the time when we were acquiring the selectional preferences. We therefore

8 The motivation for using anaphora resolution was increased coverage, but anaphora resolution turned out not actually to improve performance.
9 We use *unsupervised* to refer to systems that do not use manually sense-tagged training data, such as SemCor. Our systems, marked in the figure as *sel*, *sel-ospd*, and *sel-ospd-ana* are unsupervised.
10 We are indebted to Judita Preiss for the most-frequent-sense result. This was obtained using the frequency data supplied with the WordNet 1.7 version prereleased for SENSEVAL-2.

Table 1
Overall results.

With OSPD		Without OSPD	
Precision	51.1%	Precision	52.3%
Recall	23.2%	Recall	20.0%
Attempted	45.5%	Attempted	38.3%

Table 2
Precision results by part of speech.

	Precision (%)	Baseline precision (%)
Nouns	58.5	51.7
Polysemous nouns	36.8	25.8
Verbs	40.9	29.7
Polysemous verbs	38.1	25.3
Adjectives	49.8	48.6
Polysemous adjectives	35.5	24.0
Nouns, verbs, and adjectives	51.1	44.9
Polysemous nouns, verbs, and adjectives	36.8	27.3

do not include in the following the coarse-grained results; they are just slightly better than the fine-grained results, which seems to be typical of other systems.

Our latest overall results are shown in Table 1. In this table we show the results both with and without the OSPD heuristic. The results for the English SENSEVAL-2 tasks were generally much lower than those for the original SENSEVAL competition. At the time of the SENSEVAL-2 workshop, this was assumed to be due largely to the use of WordNet as the inventory, as opposed to HECTOR (Atkins 1993), but Palmer, Trang Dang, and Fellbaum (forthcoming) have subsequently shown that, at least for the lexical sample tasks, this was due to a harder selection of words, with a higher average level of polysemy. For three of the most polysemous verbs that overlapped between the English lexical sample for SENSEVAL and SENSEVAL-2, the performance was comparable. Table 2 shows our precision results including use of the OSPD heuristic, broken down by part of speech. Although the precision for nouns is greater than that for verbs, the difference is much less when we remove the trivial monosemous cases. Nouns, verbs, and adjectives all outperform their random baseline for precision, and the difference is more marked when monosemous instances are dropped.

Table 3 shows the precision results for polysemous words given the slot and the disambiguation source. Overall, once at least one word token has been disambiguated by the preferences, the OSPD heuristic seems to perform better than the selectional preferences. We can see, however, that although this is certainly true for the nouns, the difference for the adjectives (1.3%) is less marked, and the preferences outperform OSPD for the verbs. It seems that verbs obey the OSPD principle much less than nouns. Also, verbs are best disambiguated by their direct objects, whereas nouns appear to be better disambiguated as subjects and when modified by adjectives.

6. Discussion

6.1 Selectional Preferences

The precision of our system compares well with that of other unsupervised systems on the SENSEVAL-2 English all-words task, despite the fact that these other systems use a

Table 3
Precision results for polysemous words by part of speech and slot or disambiguation source.

	Subject (%)	Dobj (%)	Adjm (%)	OSPD (%)
Polysemous nouns	33.7	26.8	31.0	49.0
Polysemous verbs	33.8	47.3	—	29.8
Polysemous adjectives	—	—	35.1	36.4
Polysemous nouns, verbs, and adjectives	33.4	36.0	31.6	44.8

number of different sources of information for disambiguation, rather than selectional preferences in isolation. Light and Greiff (2002) summarize some earlier WSD results for automatically acquired selectional preferences. These results were obtained for three systems (Resnik 1997; Abney and Light 1999; Ciaramita and Johnson 2000) on a training and test data set constructed by Resnik containing nouns occurring as direct objects of 100 verbs that select strongly for their objects.

Both the test and training sets were extracted from the section of the Brown corpus within the Penn Treebank and used the treebank parses. The test set comprised the portion of this data within SemCor containing these 100 verbs, and the training set comprised 800,000 words from the Penn Treebank parses of the Brown corpus not within SemCor. All three systems obtained higher precision than the results we report here, with Ciaramita and Johnson's Bayesian belief networks achieving the best accuracy at 51.4%. These results are not comparable with ours, however, for three reasons. First, our results for the direct-object slot are for *all* verbs in the English all-words task, as opposed to just those selecting strongly for their direct objects. We would expect that WSD results using selectional preferences would be better for the latter class of verbs. Second, we do not use manually produced parses, but the output from our fully automatic shallow parser. Third and finally, the baselines reported for Resnik's test set were higher than those for the all-words task. For Resnik's test data, the random baseline was 28.5%, whereas for the polysemous nouns in the direct-object relation on the all-words task, it was 23.9%. The distribution of senses was also perhaps more skewed for Resnik's test set, since the first sense heuristic was 82.8% (Abney and Light 1999), whereas it was 53.6% for the polysemous direct objects in the all-words task. Although our results do show that the precision for the TCMs compares favorably with that of other unsupervised systems on the English all-words task, it would be worthwhile to compare other selectional preference models on the same data.

Although the accuracy of our system is encouraging given that it does not use hand-tagged data, the results are below the level of state-of-the-art supervised systems. Indeed, a system just assigning to each word its most frequent sense as listed in WordNet (the "first-sense heuristic") would do better than our preference models (and in fact better than the majority of the SENSEVAL-2 English all-words supervised systems). The first-sense heuristic, however, assumes the existence of sense-tagged data that are able to give a definitive first sense. We do not use any first-sense information. Although a modest amount of sense-tagged data is available for English (Miller et al. 1993, Ng and Lee 1996), for other languages with minimal sense-tagged resources, the heuristic is not applicable. Moreover, for some words the predominant sense varies depending on the domain and text type.

To quantify this, we carried out an analysis of the polysemous nouns, verbs, and adjectives in SemCor occurring in more than one SemCor file and found that a large proportion of words have a different first sense in different files and also in different genres (see Table 4). For adjectives there seems to be a lot less ambi-

Table 4
Percentages of words with a different predominant sense in SemCor, across files and genres.

	File	Genre
Nouns	70	66
Verbs	79	74
Adjectives	25	21

guity (this has also been noted by Krovetz [1998]; the data in SENSEVAL-2 bear this out, with many adjectives occurring only in their first sense. For nouns and verbs, for which the predominant sense is more likely to vary among texts, it would be worthwhile to try to detect words for which using the predominant sense is not a reliable strategy, for example, because the word shows "bursty" topic-related behavior.

We therefore examined our disambiguation results to see if there was any pattern in the predicates or arguments that were easily disambiguated themselves or were good disambiguators of the co-occurring word. No particular patterns were evident in this respect, perhaps because of the small size of the test data. There were nouns such as *team* (precision= $\frac{2}{2}$) and *cancer* ($\frac{8}{10}$) that did better than average, but whether or not they did better than the first-sense heuristic depends of course on the sense in which they are used. For example, all 10 occurrences of *cancer* are in the first sense, so the first sense heuristic is impossible to beat in this case. For the test items that are not in their first sense, we beat the first-sense heuristic, but on the other hand, we failed to beat the random baseline. (The random baseline is 21.8% and we obtained 21.4% for these items overall.) Our performance on these items is low probably because they are lower-frequency senses for which there is less evidence in the untagged training corpus (the BNC). We believe that selectional preferences would perform best if they were acquired from similar training data to that for which disambiguation is required. In the future, we plan to investigate our models for WSD in specific domains, such as sport and finance. The senses and frequency distribution of senses for a given domain will in general be quite different from those in a balanced corpus.

There are individual words that are not used in the first sense on which our TCM preferences do well, for example *sound* (precision = $\frac{2}{2}$), but there are not enough data to isolate predicates or arguments that are good disambiguators from those that are not. We intend to investigate this issue further with the SENSEVAL-2 lexical sample data, which contains more instances of a smaller number of words.

Performance of selectional preferences depends not just on the actual word being disambiguated, but the cohesiveness of the tuple <*pred*, *arg*, *gr*>. We have therefore investigated applying a threshold on the probability of the class (*nc*, *vc*, or *ac*) before disambiguation. Figure 7 presents a graph of precision against threshold applied to the probability estimate for the highest-scoring class. We show alongside this the random baseline and the first-sense heuristic for these items. Selectional preferences appear to do better on items for which the probability predicted by our model is higher, but the first-sense heuristic does even better on these. The first sense heuristic, with respect to SemCor, outperforms the selectional preferences when it is averaged over a given text. That seems to be the case overall, but there will be some words and texts for which the first sense from SemCor is not relevant, and use of a threshold on probability, and perhaps a differential between probability of the top-ranked senses suggested by the model, should increase precision.

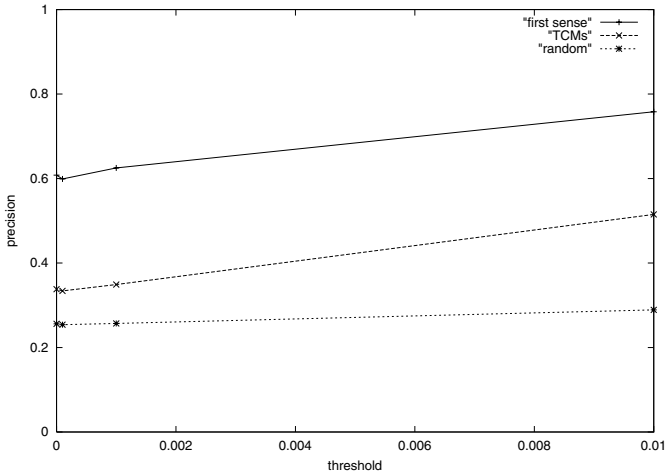

Figure 7
Thresholding the probability estimate for the highest-scoring class.

Table 5
Lemma/file combinations in SemCor with more than one sense evident.

Nouns	23%
Verbs	19%
Adjectives	1.6%

6.2 The OSPD Heuristic

In these experiments we applied the OSPD heuristic to increase coverage. One problem in doing this when using a fine-grained classification like WordNet is that although the OSPD heuristic works well for homonyms, it is less accurate for related senses (Krovetz 1998), and this distinction is not made in WordNet. We did, however, find that in SemCor, for the majority of polysemous[11] lemma and file combinations, there was only one sense exhibited (see Table 5). We refrained from using the OSPD in situations in which there was conflicting evidence regarding the appropriate sense for a word type occurring more than once in an individual file. In our experiments the OSPD heuristic increased coverage by 7% and recall by 3%, at a cost of only a 1% decrease in precision.

7. Conclusion

We quantified coverage and accuracy of sense disambiguation of verbs, adjectives, and nouns in the SENSEVAL-2 English all-words test corpus, using automatically acquired selectional preferences. We improved coverage and recall by applying the one-sense-per-discourse heuristic. The results show that disambiguation models using only selectional preferences can perform with accuracy well above the random baseline, although accuracy would not be high enough for applications in the absence of

11 Krovetz just looked at "actual ambiguity," that is, words with more than one sense in SemCor. We define *polysemy* as those words having more than one sense in WordNet, since we are using SENSEVAL-2 data, and not SemCor.

other knowledge sources (Stevenson and Wilks 2001). The results compare well with those for other systems that do not use sense-tagged training data.

Selectional preferences work well for some word combinations and grammatical relationships, but not well for others. We hope in future work to identify the situations in which selectional preferences have high precision and to focus on these at the expense of coverage, on the assumption that other knowledge sources can be used where there is not strong evidence from the preferences. The first-sense heuristic, based on sense-tagged data such as that available in SemCor, seems to beat unsupervised models such as ours. For many words, however, the predominant sense varies across domains, and so we contend that it is worth concentrating on detecting when the first sense is not relevant, and where the selectional-preference models provide a high probability for a secondary sense. In these cases evidence for a sense can be taken from multiple occurrences of the word in the document, using the one-sense-per-discourse heuristic.

Acknowledgments

This work was supported by UK EPSRC project GR/N36493 "Robust Accurate Statistical Parsing (RASP)" and EU FW5 project IST-2001-34460 "MEANING." We are grateful to Rob Koeling and three anonymous reviewers for their helpful comments on earlier drafts. We would also like to thank David Weir and Mark Mclauchlan for useful discussions.

References

Abney, Steven and Marc Light. 1999. Hiding a semantic class hierarchy in a Markov model. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, pages 1–8.

Agirre, Eneko and David Martinez. 2001. Learning class-to-class selectional preferences. In *Proceedings of the Fifth Workshop on Computational Language Learning (CoNLL-2001)*, pages 15–22.

Atkins, Sue. 1993. Tools for computer-aided lexicography: The Hector project. In *Papers in Computational Lexicography: COMPLEX 93*, Budapest.

Briscoe, Ted and John Carroll. 1993. Generalised probabilistic LR parsing of natural language (corpora) with unification-based grammars. *Computational Linguistics*, 19(1):25–59.

Briscoe, Ted and John Carroll. 1995. Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels. In *fourth ACL/SIGPARSE International Workshop on Parsing Technologies*, pages 48–58, Prague, Czech Republic.

Carroll, John, Ted Briscoe, and Antonio Sanfilippo. 1998. Parser evaluation: A survey and a new proposal. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 447–454.

Carroll, John, Guido Minnen, and Ted Briscoe. 1999. Corpus annotation for parser evaluation. In *EACL-99 Post-conference Workshop on Linguistically Interpreted Corpora*, pages 35–41, Bergen, Norway.

Carroll, John, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying English text for language impaired readers. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270, Bergen, Norway.

Ciaramita, Massimiliano and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In *Proceedings of the 18th International Conference of Computational Linguistics (COLING-00)*, pages 187–193.

Cotton, Scott, Phil Edmonds, Adam Kilgarriff, and Martha Palmer. 2001. SENSEVAL-2. Available at http://www.sle.sharp.co.uk/senseval2/.

Elworthy, David. 1994. Does Baum-Welch re-estimation help taggers? In *Proceedings of the fourth ACL Conference on Applied Natural Language Processing*, pages 53–58, Stuttgart, Germany.

Federici, Stefano, Simonetta Montemagni, and Vito Pirrelli. 1999. SENSE: An analogy-based word sense disambiguation system. *Natural Language Engineering*, 5(2):207–218.

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

Gale, William, Kenneth Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*,

26:415–439.
Krovetz, Robert. 1998. More than one sense per discourse. In *Proceedings of the ACL-SIGLEX SENSEVAL Workshop*. Available at http://www.itri.bton.ac.uk/events/senseval/ARCHIVE/PROCEEDINGS/.
Li, Hang and Naoki Abe. 1995. Generalizing case frames using a thesaurus and the MDL principle. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 239–248, Bulgaria.
Li, Hang and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.
Light, Marc and Warren Greiff. 2002. Statistical models for the induction and use of selectional preferences. *Cognitive Science*, 26(3):269–281.
McCarthy, Diana. 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*, pages 52–61.
McCarthy, Diana. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. Ph.D. thesis, University of Sussex.
Miller, George, A., Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308. Morgan Kaufmann.
Minnen, Guido, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.
Ng, Hwee Tou and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 40–47.
Palmer, Martha, Hoa Trang Dang, and Christiane Fellbaum. 2003. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. Forthcoming, *Natural Language Engineering*.
Preiss, Judita and Anna Korhonen. 2002. Improving subcategorization acquisition with WSD. In *Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, PA.
Resnik, Philip. 1997. Selectional preference and sense disambiguation. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* pages 52–57, Washington, DC.
Ribas, Francesc. 1995. On learning more appropriate selectional restrictions. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 112–118.
Rissanen, Jorma. 1978. Modelling by shortest data description. *Automatica*, 14:465–471.
Stevenson, Mark and Yorick Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 17(3):321–349.