Verb Class Disambiguation Using Informative Priors

Mirella Lapata*
University of Edinburgh

Chris Brew†
Ohio State University

Levin's (1993) study of verb classes is a widely used resource for lexical semantics. In her framework, some verbs, such as give *exhibit no class ambiguity. But other verbs, such as* write*, have several alternative classes. We extend Levin's inventory to a simple statistical model of verb class ambiguity. Using this model we are able to generate preferences for ambiguous verbs without the use of a disambiguated corpus. We additionally show that these preferences are useful as priors for a verb sense disambiguator.*

1 Introduction

Much research in lexical semantics has concentrated on the relation between verbs and their arguments. Many scholars hypothesize that the behavior of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extent determined by its meaning (Talmy, 1985; Jackendoff, 1983; Goldberg, 1995; Levin, 1993; Pinker, 1989; Green, 1974; Gropen et al., 1989; Fillmore, 1965). The correspondence between verbal meaning and syntax has been extensively studied in Levin (1993) who argues that verbs which display the same *diathesis alternations*—alternations in the realization of their argument structure— can be assumed to share certain meaning components and to form a semantically coherent class.

The converse of this assumption is that verb behavior (i.e., participation in diathesis alternations) can be used to provide clues about aspects of meaning, which in turn can be exploited to characterize verb senses (referred to as classes in Levin's 1993 terminology). A major advantage of this approach is that criteria for assigning senses can be more concrete than is traditionally assumed in lexicographic work concerned with sense distinctions (e.g., WordNet or machine-readable dictionaries) (Palmer, 2000). As an example consider sentences (1)–(4) taken from Levin. Examples (1) and (2) illustrate the dative and benefactive alternations, respectively. Dative verbs alternate between the prepositional frame 'NP1 V NP2 *to* NP3' (see (1a)) and the double object frame 'NP1 V NP2 NP3' (see (1b)), whereas benefactive verbs alternate between the double object frame (see (2a)) and the prepositional frame 'NP1 V NP2 *for* NP3' (see (2b)). To decide whether a verb is benefactive or dative it suffices to test the acceptability of the *for* and *to* frames. Verbs undergoing the conative alternation can be attested either as transitive or as intransitive with a prepositional phrase headed by the word *at*[1]. The role filled by the object of the transitive variant is shared by the noun-phrase complement of *at* in the intransitive variant (see (3)). This example makes explicit that class assignment depends not only on syntactic facts but also on judgments about semantic roles. Similarly,

* Division of Informatics, 2 Buccleuch Place, Edinburgh EH8 9LW, UK. E-mail: mlap@cogsci.ed.ac.uk
† Department of Linguistics, Oxley Hall,1712 Neil Avenue, Columbus OH. E-mail: cbrew@ling.ohio-state.edu

1 *At* is the most likely choice, but for some conative verbs the preposition is instead *on* or *onto*

the possessor object alternation involves a possessor and a possessed attribute which can be manifested either as the verbal object or as the object of a prepositional phrase headed by *for* (see (4)).

(1) a. Bill sold a car to Tom.
b. Bill sold Tom a car.
(2) a. Martha carved the baby a toy.
b. Martha carved a toy for the baby.
(3) a. Paula hit the fence.
b. Paula hit at the fence.
(4) a. I admired his honesty.
b. I admired him for his honesty.

Observation of the semantic and syntactic behavior of *pay* and *give* reveals that they pattern with *sell* in licensing the dative alternation. These verbs are all members of the GIVE class. Verbs like *make* and *build* behave similarly to *carve* in licensing the benefactive alternation and are members of the class of BUILD verbs. The verbs *beat* and *kick*, and *hit* undergo the conative alternation; they are all members of the HIT verb class. By grouping together verbs which pattern together with respect to diathesis alternations, Levin (1993) defines approximately 200 verb classes, which she argues reflect important semantic regularities. These analyses (and many similar ones by Levin and her successors) rely primarily on straightforward syntactic and syntactico-semantic criteria. To adopt this approach is to accept some limitations on the reach of our analyses, since not all semantically interesting differences will have the appropriate reflexes in syntax. Nevertheless, the emphasis on concretely available observables makes Levin's methodology a good candidate for automation(Palmer, 2000).

Therefore, Levin's (1993) classification has formed the basis for many efforts that aim to acquire lexical semantic information from corpora. These exploit syntactic cues, or at least cues that are plausibly related to syntax (Merlo and Stevenson, 2001; Schulte im Walde, 2000; Lapata, 1999; McCarthy, 2000). Other work has used Levin's classification (in conjunction with other lexical resources) to create dictionaries that express the systematic correspondence between syntax and meaning (Dorr, 1997; Dang, Rosenzweig, and Palmer, 1997; Dorr and Jones, 1996). Levin's inventory of verbs and classes has been also useful for applications such as machine translation (Dorr, 1997; Palmer and Wu, 1995), generation (Stede, 1998), information retrieval (Levow, Dorr, and Lin, 2000), and document classification (Klavans and Kan, 1998).

Although the classification provides a general framework for describing verbal meaning, it says only which verb meanings are *possible*, staying silent on the relative likelihoods of the different meanings. The inventory captures systematic regularities in the meaning of words and phrases, but falls short of providing a probabilistic model of these regularities. Such a model would be useful in applications that need to resolve ambiguity in the presence of multiple and conflicting probabilistic constraints.

More precisely, Levin (1993) provides an index of $3,024$ verbs for which she lists the semantic classes and diathesis alternations. The mapping between verbs and classes is not one-to-one. Of the $3,024$ verbs which she covers, 784 are listed as having more than one class. Even though Levin's monosemous verbs outnumber her polysemous verbs by a factor of nearly four to one, the total frequency of the former $(4,252,715)$ is comparable to the total frequency of the latter $(3,986,014)$. This means that close to half of the cases processed by a semantic tagger would manifest some degree of ambiguity. The frequencies are detailed in Table 1 and were compiled from a lemmatized version of British National Corpus (BNC, Burnage, 1996). Furthermore, as shown in Figure 1, the level of ambiguity increases in tandem with the number of alternations licensed

Table 1
Polysemous verbs according to Levin

Classes	Verbs	BNC frequency
1	2,239	4,252,715
2	536	2,325,982
3	173	738,854
4	43	395,212
5	23	222,747
6	7	272,669
7	2	26,123
10	1	4,427

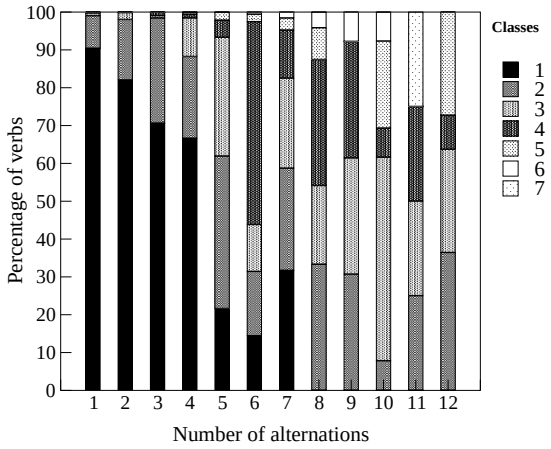

Figure 1 Relation between number of classes and alternations

by a given verb. Consider for example verbs participating in one alternation only: of these, 90.4% have one semantic class, 8.6% have two classes, .7% have three classes, and .3% have four classes. In contrast, of the verbs licensing six different alternations, 14% have one class, 17% have two classes, 12.4% have three classes, 53.6% have four classes, 2% have six classes, and 1% has seven classes. As ambiguity increases, so does the availability and potential utility of information about diathesis alternations.

Palmer (2000) and Dang et al. (1998) argue that syntactic frames and verb classes are useful for developing principled classifications of verbs. We go beyond this, showing that they can also be of assistance in disambiguation. Consider for instance the verb *serve*, which is a member of four Levin classes: GIVE, FIT, MASQUERADE, and FULFILLING. Each of these classes can in turn license four distinct syntactic frames. As shown in the examples[2] below, in (5a) *serve* appears ditransitively and belongs to the semantic class of GIVE verbs, in (5b) it occurs transitively and is a member of the class of FIT verbs, in (5c) it takes the predicative complement *as minister of the interior* and is a member of MASQUERADE verbs. Finally, in sentence (5d) *serve* is a FULFILLING verb and takes two complements, a noun phrase (*an apprenticeship*) and a prepositional phrase headed by *to* (*to a still-life photographer*). In the case of verbs like *serve* we can guess their semantic class solely on the basis of the frame with which they appear.

(5) a. I'm desperately trying to find a venue for the reception which can serve our guests an authentic Italian meal. NP1VNP2NP3
b. The airline serves 164 destinations in over 75 countries. NP1VNP2
c. Jean-Antoine Chaptal was a brilliant chemist and technocrat who served Napoleon as minister of the interior from 1800 to 1805. NP1VNP2*as*NP3
d. Before her brief exposure to pop stardom, she served an apprenticeship to a still-life photographer. NP1VNP2*to*NP3

But sometimes we do not have the syntactic information that would provide cues

2 Unless otherwise stated, our example sentences were taken (possibly in simplified form) from the BNC.

for semantic disambiguation. Consider example (6). The verb *write* is a member of three Levin classes, two of which (MESSAGE TRANSFER, PERFORMANCE) take the double object frame. In this case, we have the choice between the MESSAGE TRANSFER reading (see (6a)) and the PERFORMANCE reading (see (6b)). The same situation arises with the verb *toast* which is listed as a PREPARE verb and a JUDGMENT verb; both these classes license the prepositional frame 'NP1 V NP2 *for* NP3'. In sentence (7a) the preferred reading is that of PREPARE instead of JUDGMENT (see sentence (7b)). The verb *study* is ambiguous among three classes when attested in the transitive frame: LEARN (see example (8a)), SIGHT (see example (8b)), and ASSESSMENT (see example (8c)). The verb *convey* when attested in the prepositional frame 'NP1 V NP2 *to* NP3' can be ambiguous between the SAY class (see example (9a)) and the SEND class (see example (9b)). In order to correctly decide the semantic class for a given ambiguous verb, we would not only need detailed semantic information about the verb's arguments but also a considerable amount of world knowledge. Admittedly, selectional restrictions are sufficient for distinguishing (7a) from (7b) (one normally heats up inanimate entities and salutes animate ones), but selectional restrictions alone are probably not enough to disambiguate (6a) from (6b) since both *letter* and *screenplay* are likely to be described as written material. Rather, we need fine-grained world knowledge: both scripts and letters can be written for someone, only letters can be written to someone.

(6) a. A solicitor wrote him a letter at the airport.
b. I want you to write me a screenplay called "The Trip".

(7) a. He sat by the fire and toasted a piece of bread for himself.
b. We all toasted Nigel for his recovery.

(8) a. Chapman studied medicine at Cambridge.
b. Romanov studied the old man carefully, looking for some sign that he knew exactly what had been awaiting him at the bank.
c. The alliance will also study the possibility of providing service to other high-volume products, such as IBM and multi-vendor workstations.

(9) a. By conveying the news to her sister, she would convey by implication something of her own anxiety.
b. The judge signed the committal warrant and the police conveyed Mr. Butler to prison, giving the warrant to the governor.

This need for world knowledge (or at least a convenient way of approximating this knowledge) is not an isolated phenomenon, but manifests itself across a variety of classes and frames (e.g., double object, transitive, prepositional frame, see examples (6)–(9)). We have argued that the concreteness of Levin-style verb classes is an advantage, but this advantage would be compromised if we tried to fold too much world knowledge into the classification. We do not do this. Instead section 5 of the current paper describes disambiguation experiments in which our probabilistic Levin classes are used in tandem with proxies for appropriate world knowledge.

Levin's (1993) classification has other limitations:

- It is not intended as an exhaustive description of English verbs, their meanings and their likelihood.
- Many other classifications could have been built using the same principles. A different grouping might for example have occurred, if finer or coarser semantic distinctions were taken into account (see (Merlo and Stevenson, 2001; Dang, Rosenzweig, and Palmer, 1997) for alternative classifications) or if the containment of ambiguity was one of the classification objectives.

- As pointed out by Kipper, Dang, and Palmer (2000), Levin classes exhibit inconsistencies and verbs are listed in multiple classes, some of which have conflicting sets of syntactic frames. This means that some ambiguities may also arise due to accidental errors or inconsistencies.
- The classification was not created with computational uses in mind, but for human readers, so it has not been necessary to remedy all the errors and omissions that might cause trouble for machines.

Similar issues arise in almost all efforts to make use of pre-existing lexical resources for computational purposes (Briscoe and Carroll, 1997), so none of the above comments should be taken as criticisms of Levin's achievement

The objective of this paper is to show how to train and use a probabilistic version of Levin's classification in verb sense disambiguation. We treat errors and inconsistencies in the classification as noise. While all our tests have used Levin's classes and the British National Corpus, the method itself depends neither on on the details of Levin's classification nor on parochial facts about the English language. Our future work will include tests on other languages, other classifications and other corpora.

The model developed in this paper takes as input a partially parsed corpus and generates, for each combination of a verb and its syntactic frame, a probability distribution over the available verb classes. The corpus itself does not have to be labeled with classes. This makes it feasible to use large corpora. Our model is not immediately useful for disambiguation, because it cannot discriminate between different occurrences of the same verb and frame, but it can (as we show in section 5) be used as a prior in a full disambiguation system that does take appropriate account of context. The model relies on several gross simplifications does not take selectional restrictions, discourse, or pragmatic information into account, but is demonstrably superior to simpler priors that make no use of subcategorization.

In Section 2 we describe the probabilistic model and the estimation of the various model parameters. In Sections 3 and 4 we report on the results of two experiments which use the model to derive the dominant class for polysemous verbs. Section 5 discusses our verb class disambiguation experiments. We base our results on the BNC, a 100 million word collection of samples of written and spoken language from a wide range of sources designed to represent a wide cross-section of current British English, both spoken and written (Burnard, 1995). We discuss our results in Section 7 and review related work in Section 8.

2 The Prior Model

Consider again the sentences in (6). Assuming that we more often write something to someone rather than for someone, we would like to derive MESSAGE TRANSFER as the prevalent class for *write* rather than PERFORMANCE. We view the choice of a class for a polysemous verb in a given frame as maximizing the joint probability $P(c, f, v)$ where v is a verb subcategorizing for the frame f and inhabiting more than one Levin class c:

$$P(c, f, v) = P(v) \cdot P(f|v) \cdot P(c|v, f) \tag{10}$$

Although the terms $P(v)$ and $P(f|v)$ can be estimated from the BNC ($P(v)$ reduces to the number of times a verb is attested in the corpus and $P(f|v)$ can be obtained through parsing), the estimation of $P(c|v, f)$ is somewhat problematic since it relies on the frequency $F(c, v, f)$. The latter could be obtained straightforwardly if we had access to a parsed corpus annotated with subcategorization and semantic class information. Lack-

Table 2
Estimation of Model Parameters

(14)	$\hat{P}(v) = \frac{F(v)}{\sum_i F(v_i)}$	(15)	$\hat{P}(f\|v) = \frac{F(f,v)}{F(v)}$	(16)	$\hat{P}(f\|c) = \frac{F(f,c)}{F(c)}$
(17)	$\hat{P}(c) = \frac{F(c)}{\sum_i F(c_i)}$	(18)	$\hat{P}(f) = \frac{F(f)}{\sum_i F(f_i)}$	(19)	$F(f,c) = \sum_i F(c,f,v_i)$
(20)	$F(c) = \sum_i F(v_i,c)$	(21)	$F(v,c) = F(v) \cdot P(c\|v)$		

ing such a corpus we will assume that the semantic class determines the subcategorization patterns of its members independently of their identity (see (11)).

(11) $P(c|v,f) \approx P(c|f)$

The independence assumption is a simplification of Levin's (1993) hypothesis that the argument structure of a given verb is a direct reflection of its meaning. The rationale behind the approximation in (11) is that since class formation is determined on the basis of diathesis alternations, it is the differences in subcategorization structure that determine class likelihood rather than the identity of the individual verbs. For example, if we know that some verb subcategorizes for the double object and the prepositional 'NP1 V NP2 *to* NP3' frames, we can guess that it is a member of the GIVE class or the MESSAGE TRANSFER class without knowing whether this verb is *give*, or *write*, or *tell*.

Note that the approximation in (11) assumes that verbs of the same class uniformly subcategorize (or not) for a given frame. This is evidently not true for all classes of verbs. For example, all GIVE verbs undergo the dative diathesis alternation and therefore we would expect them to be attested in both the double object and prepositional frame, but only a subset of CREATE verbs undergo the benefactive alternation. For example, the verb *invent* is a CREATE verb and can be attested only in the benefactive prepositional frame (*I will invent a tool for you* versus ?*I will invent you a tool*, see Levin (1993) for details). By applying Bayes Law we write $P(c|f)$ as:

(12) $P(c|f) = \frac{P(f|c) \cdot P(c)}{P(f)}$

By substituting (12) into (10), $P(c, f, v)$ can be written as:

(13) $P(c,f,v) = \frac{P(v) \cdot P(f|v) \cdot P(f|c) \cdot P(c)}{P(f)}$

It is easy to obtain $P(v)$ from the lemmatized BNC (see (14) in Table 2). In order to estimate the probability $P(f|v)$ we need to know how many times a verb is attested with a given frame. We acquired Levin compatible subcategorization frames from the BNC after performing a coarse-grained mapping between Levin's frame descriptions and surface syntactic patterns without preserving detailed semantic information about argument structure and thematic roles. This resulted in 80 frame types that were grossly compatible with Levin. We used Gsearch (Corley et al., 2001), a tool which facilitates the search of arbitrary part-of-speech tagged corpora for shallow syntactic patterns based

on a user-specified context-free grammar and a syntactic query. We specified a chunk grammar for recognizing the verbal complex, NPs and PPs and used Gsearch to extract tokens matching the frames specified in Levin. We discarded all frames with a frequency smaller than 5, as they were likely to be unreliable given our heuristic approach. The frame probability $P(f)$ (see the denominator in (13) and equation (18) in Table 2) was also estimated on the basis of the Levin compatible subcategorization frames that were acquired from the BNC.

We cannot read off $P(f|c)$ in (13) directly from the corpus, because it is not annotated with verb classes. Nevertheless Levin's (1993) classification records the syntactic frames that are licensed by a given verb class (for example GIVE verbs license the double object and the 'NP1 V NP2 *to* NP3' frame) and also the number and type of classes a given verb exhibits (e.g. *write* inhabits two classes, PERFORMANCE and MESSAGE TRANSFER). Furthermore, we know how many times a given verb is attested with a certain frame in the corpus as we have acquired Levin compatible frames from the BNC (see (15)). We first explain how we obtain $F(f, c)$ which we re-write as the sum of all occurrences of verbs v that are members of class c and are attested in the corpus with frame f (see (16) and (19) in Table 2).

For monosemous verbs the count $F(c, f, v)$ reduces to the number of times these verbs have been attested in the corpus with a certain frame. For polysemous verbs, we additionally need to know the class in which they were attested in the corpus. Note that we don't necessarily need an annotated corpus for class ambiguous verbs whose classes license distinct frames (see example (5) in Section 1), provided that we have extracted verb frames relatively accurately. For genuinely ambiguous verbs (i.e., verbs licensed by classes that take the same frame), given that we don't have access to a corpus annotated with verb class information, we distribute the frequency of the verb and its frame evenly across its semantic classes.

$$F(c, f, v) = \frac{F(f, v)}{|classes(v, f)|} \tag{22}$$

Here $F(f, v)$ is the co-occurrence frequency of a verb and its frame and $|classes(v, f)|$ is the number of classes verb v is a member of when found with frame f. The joint frequency of a class and its frame $F(f, c)$ is then the sum of all verbs that are member of the class c and are attested with frame f in the corpus (see (19)). Table 3 shows the estimation of the frequency $F(c, f, v)$ for six verbs that are members of the GIVE class. Consider for example *feed* which is a member of four classes: GIVE, GORGE, FEEDING, and FIT. Of these classes, only FEEDING and GIVE license the double object and prepositional frame. This is why the co-occurrence frequency of *feed* with these frames is divided by two. The verb *serve* inhabits four classes. The double object frame is licensed by the GIVE class, whereas the prepositional frame is additionally licensed by the FULFILLING class and therefore the co-occurrence frequency $F(\text{NPVNP}to\text{NP}, serve)$ is equally distributed between these two classes. This is clearly a simplification, since one would expect $F(c, f, v)$ to vary for different verb classes. However, note that according to this estimation $F(f, c)$ will vary across frames reflecting differences in the likelihood of a class being attested with a certain frame.

Both terms $P(f|c)$ and $P(c)$ in (13) rely on the class frequency $F(c)$ (see (16) and (17)). We rewrite $F(c)$ as the sum of all verbs attested in the corpus with class c (see (20) in Table 2). For monosemous verbs the estimate of $F(v, c)$ reduces to the count of the verb in the corpus. Once again we cannot estimate $F(v, c)$ for polysemous verbs directly. The task would be straightforward if we had a corpus of verbs, each labeled explicitly with class information. All we have is the overall frequency of a given verb in the BNC and the number of classes it is a member of according to Levin (1993). Since pol-

Table 3
Estimation of $F(c, f, v)$ and $F(v, c)$

GIVE	$F(\text{GIVE}, \text{NPVNPNP}, v)$	$F(\text{GIVE}, \text{NPVNP}to\text{NP}, v)$	$F(v, \text{GIVE})$
feed	$\frac{98}{2}$	$\frac{40}{2}$	$\frac{3,263}{4}$
give	$25,705$	$7,502$	$126,894$
lend	343	648	$2,650$
rent	$\frac{6}{2}$	10	$\frac{1,060}{2}$
pass	$\frac{181}{3}$	$\frac{256}{3}$	$\frac{19,459}{4}$
serve	85	$\frac{58}{2}$	$\frac{15,457}{4}$

ysemous verbs can generally be the realization of more than one semantic class, counts of semantic classes can be constructed by dividing the contribution from the verb by the number of classes it belongs to (Resnik, 1993; Lauer, 1995). We rewrite the frequency $F(v, c)$ as shown in (21) in Table 2, and approximate $P(c|v)$, the true distribution of the verb and its classes, as follows:

$$F(v, c) \approx F(v) \cdot \frac{1}{|classes(v)|} \tag{23}$$

Here, $F(v)$ is the number of times the verb v was observed in the corpus and $|classes(v)|$ is the number of classes c it belongs to. For example, in order to estimate the frequency of the class GIVE we consider all verbs that are listed as members of this class in Levin (1993). The class contains 13 verbs, among which six are polysemous. We will obtain $F(\text{GIVE})$ by taking into account the verb frequency of the monosemous verbs ($|classes(v)|$ is one in this case) as well as distributing the frequency of the polysemous verbs among their classes. For example, *feed* inhabits the classes GIVE, GORGE, FEEDING, and FIT and occurs in the corpus 3,263 times. We will increment the count of $F(\text{GIVE})$ by $\frac{3,263}{4}$. Table 3 illustrates the estimation of $F(v, c)$ for six members of the GIVE class. The total frequency of the class is obtained by summing over individual the $F(v, c)$'s (see equation (20)).

The approach in (23) relies on the simplifying assumption that the frequency of a verb is distributed evenly across its semantic classes. This is clearly not true for all verbs. Consider for example the verb *rent* which inhabits classes GIVE (*Frank rented Peter his room*) and GET (*I rented my flat for my sister*). Intuitively speaking, the GIVE sense of *rent* is more frequent than GET, however this is not taken into account in (23), primarily because we do not know the true distribution of the classes for *rent*. An alternative to (23) is to distribute the verb frequency unequally among verb classes. Even though we don't know how likely classes are in relation to a particular verb, we can approximate how likely classes are in general on the basis of their size (i.e., number of verbs that are members of each class). So then we can distribute a verb's frequency unequally, according to class size. This time we approximate $P(c|v)$ (see (21) in Table 2) by $P(c|amb\_class)$, the probability of class c given the ambiguity class[3] amb_class. The latter represents the set of classes a verb might inhabit:

$$F(v, c) \approx F(v) \cdot P(c|amb\_class) \tag{24}$$

3 Our use of ambiguity classes is inspired by a similar use in HMM based part-of-speech tagging (Kupiec, 1992).

Table 4
Estimation of $F(v,c)$ for the verb *feed*

c	$\|c\|$	$P(c\|amb\_class)$	$F(v,c)$
GIVE	15	.39	1,272.57
GORGE	8	.21	685.23
FEED	3	.08	261.04
FIT	12	.32	1,044.16

Table 5
The ten most frequent classes using equal distribution of verb frequencies

c	$F(c)$
CHARACTERIZE	601,647.4
GET	514,308.0
SAY	450,444.6
CONJECTURE	390,618.4
FUTURE HAVING	369,229.3
DECLARE	264,923.6
AMUSE	258,857.9
DIRECTED MOTION	252,775.6
MESSAGE TRANSFER	248,238.7
GIVE	208,884.1

Table 6
The ten most frequent classes using unequal distribution of verb frequencies

c	$F(c)$
GET	453,843.6
SAY	447,044.2
CHARACTERIZE	404,734.2
CONJECTURE	382,193.8
FUTURE HAVING	370,717.7
DECLARE	285,431.7
DIRECTED MOTION	255,821.6
POCKET	247,392.7
AMUSE	205,729.4
GIVE	197,828.8

We collapse verbs into ambiguity classes in order to reduce the number of parameters which must be estimated: we certainly lose information, but the approximation makes it easier to get reliable estimates from limited data. We simply approximate $P(c|amb\_class)$ using a heuristic based on class size:

$$P(c|amb\_class) \approx \frac{(|c|)}{\sum_{c \in amb\_class} |c|)} \tag{25}$$

For each class we recorded the number of its members after discarding verbs whose frequency was less than one per million in the BNC. This gave us a first approximation of the size of each class. We then computed, for each polysemous verb, the total size of the classes of which it was a member. We calculated $P(c|amb\_class)$ by dividing the former by the latter (see equation (25)). We obtained the class frequency $F(c)$ by multiplying $P(c|amb\_class)$ by the observed frequency of the verb in the BNC (see equation (24)). As an example consider again $F(\text{GIVE})$ which is calculated by summing over all verbs that are members of this class (see 20). In order to add the contribution of the verb *feed* we need to distribute its corpus frequency among the classes GIVE, GORGE, FEED, FIT. The respective $P(c|amb\_class)$ for these classes are $\frac{15}{38}$, $\frac{8}{38}$, $\frac{3}{38}$, and $\frac{12}{38}$. By multiplying these by the frequency of *feed* in the BNC (3,263) we obtain the the $F(v,c)$s given in Table 4. Only the frequency $F(feed, \text{GIVE})$ is relevant for $F(\text{GIVE})$.

The estimation process described above involves at least one gross simplification, since $P(c|amb\_class)$ is calculated without reference to the identity of the verb in question. For any two verbs which fall into the same set of classes, $P(c|amb\_class)$ will be the same, even though one or both may be atypical in its distribution across the classes. Furthermore, the estimation tends to favor large classes, again irrespectively of the iden-

Table 7
Smoothed Estimates

$$(26)\quad P(f|v) \approx \frac{F(f,v) + \frac{F(f,V)}{F(V)}}{F(v)+1} \qquad (27)\quad F(f,V) = \sum_i F(f,v_i)$$

$$(28)\quad P(f|c) \approx \frac{F(f,c) + \frac{Ff,C)}{F(C)}}{F(c)+1} \qquad (29)\quad F(C) = \sum_i F(c_i)$$

tity of the verb in question. For example, the verb *carry* has three classes, CARRY, FIT, and COST. Intuitively speaking, the CARRY class is the most frequent (e.g., *Smoking can impair the blood which carries oxygen to the brain*, *I carry sugar lumps around with me*). However, since the FIT class (e.g., *Thameslink presently carries 20,000 passengers daily*) is larger than the CARRY class, it will be given a higher probability (.45 versus .4). This is clearly wrong, but it is an empirical question how much it matters. Tables 5 and 6 show the ten most frequent classes as estimated using (23) and (24). We explore the contribution of the two estimation schemes for $P(c)$ in Experiments 1 and 2.

The probabilities $P(f|c)$ and $P(f|v)$ will be unreliable when the frequencies $F(f,v)$ and $F(f,c)$ are small, and undefined when the frequencies are zero. Following Hindle and Rooth (1993), we smooth the observed frequencies as shown in Table 7. When $F(f,v)$ is zero, the estimate used is proportional to the average $\frac{F(f,V)}{F(V)}$ across all verbs. Similarly, when $F(f,c)$ is zero, our estimate is proportional to the average $\frac{F(f,C)}{F(C)}$ across all classes. We do not claim that this scheme is perfect, but any deficiencies it may have are almost certainly masked by the effects of approximations and simplifications elsewhere in the system.

We evaluated the performance of the model on all verbs listed in Levin (1993) which are polysemous (i.e., members of more than one class) and take frames characteristic of the widely studied dative and benefactive alternations (Pinker, 1989; Boguraev and Briscoe, 1989; Levin, 1993; Goldberg, 1995; Briscoe and Copestake, 1999) and the less well-known conative and possessor object alternations (see the examples in (1)–(4)). All four alternations seem fairly productive, i.e., a large number of verbs undergo these alternations, according to Levin. A large number of classes licenses the frames that are relevant for these alternations and the verbs that inhabit these classes are likely to exhibit class ambiguity: 20 classes license the double object frame, 22 license the prepositional frame 'NP1 V NP2 *to* NP3', 17 classes license the benefactive 'NP1 V NP2 *for* NP3' frame, 118 (out of 200) classes license the transitive frame, and 15 classes license the conative 'NP1 V *at* NP2' frame.

In Experiment 1 we use the model to test the hypothesis that subcategorization information can be used to disambiguate polysemous verbs. In particular, we concentrate on verbs like *serve* (see example (5)) which can be disambiguated solely on the basis of their frame. In Experiment 2 we focus on verbs which are genuinely ambiguous, i.e., they inhabit a single frame and yet can be members of more than one semantic class (e.g., *write*, *study*, see examples (6)–(9) in Section 1). In this case, we use the probabilistic model to assign a probability to each class the verb inhabits. The class with the highest probability represents the dominant meaning for a given verb.

3 Experiment 1: Using Subcategorization to Resolve Verb Class Ambiguity

3.1 Method

In this experiment we focused solely on verbs whose meaning can be potentially disambiguated by taking into account their subcategorization frame. A model which performs badly on this task cannot be expected to produce any meaningful results for genuinely ambiguous verbs.

We considered 128 verbs with the double object frame (2.72 average class ambiguity), 101 verbs with the prepositional frame 'NP1 V NP2 *to* NP3' (2.59 average class ambiguity), 113 verbs with the frame 'NP1 V NP2 *for* NP3' (2.63 average class ambiguity), 42 verbs with the frame 'NP1 V *at* NP3' (3.05 average class ambiguity), and 39 verbs with the transitive frame (2.28 average class ambiguity). The task was the following: given that we know the frame of a given verb can we predict its semantic class? In other words by varying the class c in the term $P(c, f, v)$ we are trying to see whether the class which maximizes it is the one predicted by the lexical semantics and the argument structure of the verb in question. The model's responses were evaluated against Levin's (1993) classification. The model's performance was considered correct if it agreed with Levin in assigning a verb to an appropriate class given a particular frame. Recall from Section 2 that we proposed two approaches for the estimation of the class probability $P(c)$. We explore the influence of $P(c)$ by obtaining two sets of results corresponding to the two estimation schemes.

3.2 Results

The model's accuracy is shown in Tables 8 and 9. The results in Table 8 were obtained using the estimation scheme for $P(c)$ which relies on the even distribution of the frequency of a verb across its semantic classes (see equation (23)). The results in Table 9 were obtained using an alternative scheme which distributes verb frequency unequally among verb classes by taking class size into account (see equation (24)). As mentioned in Section 3.1, the results were obtained by comparing the model's performance against Levin's (1993) classification. We also compared the results to the baseline of choosing the most likely class $P(c)$ (without taking subcategorization information into account). The latter was determined on the basis of the approximations described in Section 2 (see equations (20), (23), (24), and (25)).

The model achieved an accuracy of 93.9% using either type of estimation for $P(c)$. It also outperformed the baseline by 38.1% (see Table 8) and 37.2% (see Table 9). One might expect an accuracy of 100% since these verbs can be disambiguated solely on the basis of their frame. However, the performance of our model is less, mainly because of the way we estimated the terms $P(c)$ and $P(f|c)$: we over-emphasize the importance of class information without taking into account how individual verbs distribute across classes. Furthermore, we rely on frame frequencies acquired from the BNC, using shallow syntactic analysis which means that the correspondence between Levin's (1993) frames and our acquired frames is not one-to-one. Except from the fact that our frames do not preserve much of the linguistic information detailed Levin, the number of frames acquired for a given verb can be a subset or superset of the frames available in Levin. Note that the two estimation schemes yield comparable performances. This is a positive result given the importance of $P(c)$ in the estimation of $P(c, f, v)$.

A more demanding task for our probabilistic model will be with genuinely ambiguous verbs (i.e., verbs for which the mapping between meaning and subcategorization is not one-to-one). Although native speakers may have intuitions about the dominant interpretation for a given verb, this information is entirely absent from Levin (1993) and from the corpus on which our model is trained on. In Experiment 2 we show how our

Table 8
Model accuracy using equal distribution of verb frequencies for the estimation of $P(c)$

Frame	Baseline	Model
NP1 V NP2 NP3	60.9%	93.8%
NP1 V NP *to* NP3	63.3%	95.0%
NP1 V NP *for* NP3	63.6%	98.2%
NP1 V *at* NP2	2.4%	83.3%
NP1 V NP2	43.6%	87.2%
Combined	55.8%	93.9%

Table 9
Model accuracy using unequal distribution of verb frequencies for the estimation of $P(c)$

Frame	Baseline	Model
NP1 V NP2 NP3	62.5%	93.8%
NP1 V NP *to* NP3	67.3%	95.0%
NP1 V NP *for* NP3	66.4%	98.2%
NP1 V *at* NP2	2.4%	85.7%
NP1 V NP2	41.0%	84.6%
Combined	56.7%	93.9%

model can be used to recover this information.

4 Experiment 2: Using Corpus Distributions to Derive Verb Class Preferences

4.1 Method

We evaluated the performance of our model on 67 genuinely ambiguous verbs, i.e., verbs which inhabit a single frame and can be members of more than one semantic class (e.g., *write*). These verbs were listed in Levin (1993) and undergo the dative, benefactive, conative, and possessor object alternations. As in Experiment 1, we considered verbs with the double object frame (3.27 average class ambiguity), verbs with the frame 'NP1 V NP2 *to* NP3' (2.94 average class ambiguity), verbs with the frame 'NP1 V NP2 *for* NP3' (2.42 average class ambiguity), verbs with the frame 'NP1 V *at* NP3' (2.71 average class ambiguity), and transitive verbs (2.77 average class ambiguity). The model's predictions were compared against manually annotated data which was used only for testing purposes. The model was trained without access to a disambiguated corpus. More specifically, corpus tokens characteristic of the verb and frame in question were randomly sampled from the BNC and annotated with class information so as to derive the *true* distribution of the verb's classes in a particular frame. We describe the verb selection procedure as follows.

Given the restriction that these verbs are semantically ambiguous in a specific syntactic frame we could not simply sample from the entire BNC, since this would decrease the chances of finding the verb in the frame we are interested in. Instead, a stratified sample was used: for all class ambiguous verbs, tokens were randomly sampled from the parsed data used for the acquisition of verb frame frequencies. The model was evaluated on verbs for which a reliable sample could be obtained. This meant that verbs had to have a frame frequency larger than 50. For verbs exceeding this threshold 100 tokens were randomly selected and annotated with verb class information. For verbs with frame frequency less than 100 and more than 50, no sampling took place, the entire set of tokens was manually annotated. This selection procedure resulted in 14 verbs with the double object frame, 16 verbs with the frame 'NP1 V NP2 *to* NP3', two verbs with the frame 'NP1 V NP2 *for* NP3', one verb with the frame 'NP1 V *at* NP3', and 80 verbs with the transitive frame. From the transitive verbs we further randomly selected 34 verbs; these were manually annotated and used for evaluating the model's performance.[4]

The selected tokens were annotated with class information by two judges, both lin-

4 Although the model can yield predictions for any number of verbs, evaluation could not be performed for all 80 verbs for which our judges would have to annotate 8,000 corpus tokens.

Table 10
Model accuracy using equal distribution of verb frequencies for the estimation of $P(c)$

Frame	Baseline	Model
NP1 V NP2 NP3	50.0%	78.6%
NP1 V NP *to* NP3	43.8%	68.8%
NP1 V NP *for* NP3	00.0%	100.0%
NP1 V *at* NP2	100.0%	100.0%
NP1 V NP2	47.1%	73.5%
Combined	46.2%	74.6%

Table 11
Model accuracy using unequal distribution of verb frequencies for the estimation of $P(c)$

Frame	Baseline	Model
NP1 V NP2 NP3	50.0%	78.6%
NP1 V NP *to* NP3	43.8%	75.0%
NP1 V NP *for* NP3	00.0%	100.0%
NP1 V *at* NP2	100.0%	100.0%
NP1 V NP2	47.1%	67.6%
Combined	46.2%	73.1%

guistics graduates. The classes were taken from Levin (1993) and augmented with the class OTHER which was reserved for corpus tokens which either had the wrong frame or for which the classes in question were not applicable. The judges were given annotation guidelines (for each verb) but no prior training (for details on the annotation study see Lapata (2001)). The annotation provided a gold standard for evaluating the model's performance and enabled us to test whether humans agree on the class annotation task. We measured the judges' agreement on the annotation task using the Kappa coefficient (Cohen, 1960). In general, the agreement on the class annotation task was good with Kappa values ranging from .66 to 1.00 (the mean Kappa was .80, StdDev = .09).

4.2 Results

We counted the performance of our model as correct if it agreed with the "most preferred", i.e., most frequent verb class, as determined in the manually annotated corpus sample by taking the average of the responses of both judges. To give an example consider the verb *feed* which in the double object frame is ambiguous between the classes FEED and GIVE. According to the model FEED is the most likely class for *feed*. Out of 100 instances of the verb *feed* in the double object frame, 61 were manually assigned the FEED class, 32 were assigned the GIVE class, and 6 were parsing mistakes (and therefore assigned the class OTHER). In this case the model's outcome is considered correct given that the corpus tokens also reveal a preference for the FEED (i.e., the FEED instances outnumber the GIVE ones).

As in Experiment 1, we explored the influence of the parameter $P(c)$ on the model's performance by obtaining two sets of results corresponding to the two estimation schemes discussed in Section 2. The model's accuracy is shown in Tables 10 and 11. The results in Table 11 were obtained using the estimation scheme for $P(c)$ which relies on the even distribution of a verb's frequency across its semantic classes (see equation (23)). The results in Table 10 were obtained using a scheme which distributes verb frequency unequally among verb classes by taking class size into account (see equation (24)). As in Experiment 1, the results were compared to a simple baseline which defaults to the most likely class without taking verb frame information into account (see equations (20), (23), (24), and (25) in Section 2).

The model achieved an accuracy of 74.6% using the estimation scheme of equal distribution and a accuracy of 73.1% using the estimation scheme of unequal distribution. The difference between the two estimation schemes is not statistically significant (using the χ^2 statistic $p = .84$, $N = 67$). Table 12 gives the distribution of classes for 12 polysemous verbs taking the double object frame as obtained from the manual annotation of corpus tokens together with inter-annotator agreement (K). We also give the (log-

Table 12
Semantic preferences for verbs with the double object frame

Verb		Class				K
call	√	DUB 93 −7.59	GET 3 −8.12	OTHER 4		.82
cook		BUILD 28 −11.68	PREPARE 33 −11.50	OTHER 1		1.00
declare	√	DECLARE 35 −10.51	REF. APPEAR. 18 −12.18	OTHER 5		.89
feed	√	FEED 61 −10.63	GIVE 32 −12.16	OTHER 6		.73
find	√	DECLARE 36 −7.69	GET 47 −7.43	OTHER 17		.70
leave	√	GET 6 −7.91	FULFILL 14 −10.40	F. HAVE 56 −7.66	OTHER 23	.67
make	√	BUILD 21 −7.25	DUB 66 −6.13	OTHER 13		.79
pass	√	GIVE 81 −8.84	SEND 0 −8.96	THROW 0 −9.98	OTHER 19	.93
save	√	BILL 24 −9.74	GET 62 −9.59	OTHER 14		.74
shoot		THROW 91 −10.94	GET 0 −9.99	OTHER 5		1.00
take		BRING-TAKE 15 −7.02	PERFORM 40 −7.38	OTHER 45		.77
write	√	MSG. TRANS. 54 −8.79	PERFORM 19 −9.05	OTHER 18		.85

transformed) probabilities of these classes as derived by the model.[5] The presence of the symbol √ indicates that the model's class preference for a given verb agrees with its distribution in the corpus. The absence of √ indicates disagreement. For the comparison shown in Table 12 model class preferences were derived using the equal distribution estimation scheme for $P(c)$ (see equation (23)).

As shown in Tables 12 the model's predictions are generally borne out in the corpus data. Misclassifications are due mainly to the fact that the model does not take verb-class dependencies into account. Consider for example the verb *cook*. According to the model the most likely class for *cook* is BUILD. Although it may generally be the case that BUILD verbs (e.g., *make*, *assemble*, *build*) are more frequent than PREPARE verbs (e.g., *bake*, *roast*, *boil*) the situation is reversed for *cook*. The same is true for the verb *shoot* which when attested in the double object frame is more likely to be a THROW verb (*Jamie shot Mary a glance*) rather than a GET verb (*I will shoot you two birds*).

Notice that our model is not context-sensitive, i.e., it does not derive class rankings tailored to specific verbs, primarily because this information is not readily available in the corpus as explained in Section 2. However, we have effectively built a prior model of the joint distribution of verbs, their classes and their syntactic frames which can be useful for disambiguating polysemous verbs in context. We describe our class disambiguation experiments as follows.

5 No probabilities are given for the OTHER class; this is not a Levin class, however it was used by the annotators, mainly to indicate parsing errors.

5 Class Disambiguation

In the previous sections we focused on deriving a model of the distribution of Levin classes without relying on annotated data and showed that this model infers that right class for genuinely ambiguous verbs 74.6% of the time without taking the local context of their occurrence into account. An obvious question is whether this information is useful for disambiguating tokens rather than types. In the following we report on a disambiguation experiment that takes advantage of this prior information.

Word sense disambiguation is often cast as a problem in supervised learning, where a disambiguator is induced from a corpus of manually sense-tagged text. The context within which the ambiguous word occurs is typically represented by a set of linguistically motivated features from which a learning algorithm induces a representative model that performs the disambiguation. A variety of classifiers have been employed for this task (see (Mooney, 1996; Ide and Véronis, 1998) for overviews), the most popular being decision lists (Yarowsky, 1994; Yarowsky, 1995) and naive Bayesian classifiers (Pedersen, 2000; Ng, 1997; Pedersen and Bruce, 1998; Mooney, 1996; Cucerzan and Yarowsky, 2002). We employed a naive Bayesian classifier (Duda and Hart, 1973) for our experiments as it is a very convenient framework for incorporating prior knowledge and studying its influence on the classification task. In Section 5.1 we describe a basic naive Bayes classifier and show how it can be extended with informative priors. In Section 5.2 we discuss the types of contextual features we use and report on our experimental results in Section 6.2.

5.1 Naive Bayes Classification

A naive Bayesian classifier assumes that all the feature variables representing a problem are conditionally independent given the value of the classification variable. In word sense disambiguation, the features $(a_1, a_2, \cdots, a_n)$ represent the context surrounding the ambiguous word, and the classification variable c is the sense (Levin class in our case) of the ambiguous word in this particular context. Within a naive Bayes approach the probability of the class c given its context can be expressed as:

$$P(c|a_i) = \frac{P(c)\prod_{i=1}^{n} P(a_i|c)}{P(a_i)} \tag{30}$$

where $P(a_i|c)$ is the probability that a test example is of class c given the contextual features a_i. Since the denominator $P(a_i)$ is constant for all classes c, the problem reduces to finding the class c with the maximum value for the numerator:

$$P(c|a_i) \approx P(c)\prod_{i=1}^{n} P(a_i|c) \tag{31}$$

If we choose the prior $P(c)$ to be uniform ($P(c) = \frac{1}{|c|}$ for all $c \in C$), (31) can be further simplified to:

$$P(c|a) \approx \prod_{i=1}^{n} P(a_i|c) \tag{32}$$

Assuming a uniform prior, a basic naive Bayes classifier is as follows:

$$(33) \qquad *c = \prod_{i=1}^{n} P(a_i|c)$$

Note, however, that we have developed in the previous section two types of non-uniform prior models. The first model derives $P(c)$ heuristically from the BNC ignoring the identity of the polysemous verb and its subcategorization profile, while the second model estimates the class distribution $P(c, v, f)$ by taking the frame distribution into account. So, the naive Bayes classifier in (33) can be extended with a non-uniform prior as shown below:

$$(34) \qquad *c = P(c) \prod_{i=1}^{n} P(a_i|c)$$

$$(35) \qquad *c = P(c, v, f) \prod_{i=1}^{n} P(a_i|c, f, v)$$

where $P(c)$ is estimated as shown in (17)–(20) and $P(c, v, f)$, the prior for each class c corresponding to verb v in frame f, is estimated as explained in Section 2, (see (13)). As before, a_i are the contextual features. The probabilities $P(a_i|c)$ can be estimated from the training data simply by counting the co-occurrence of feature a_i with class c (for (34)) or the co-occurrence of a_i with class c, verb v, and frame f (for (35)). For features which have zero counts, we use add-k smoothing (Johnson, 1932), where k is a small number less than one.

5.2 Feature Space

As common in word sense disambiguation studies we experimented with two types of context representations, collocations and co-occurrences. Co-occurrences simply indicate if a given word occurs within some number of words to the left or right of the ambiguous word. In this case the contextual features are binary and represent the presence or absence of a particular word in the current or preceding sentence. We used four types of context in our experiments: left context (i.e., words occurring to the left of the ambiguous word), right context (i.e., words occurring to the right of the ambiguous word), the current sentence (i.e., words surrounding the ambiguous word), and the current sentence together with its immediately preceding sentence. Punctuation and capitalization were removed from the windows of context, non-content words were included. The context words were represented as lemmas or parts-of-speech.

Collocations are words that are frequently adjacent to the word to be disambiguated. We considered 12 types of collocations. Examples of collocations for the verb *write* are illustrated in Table 13. The L columns indicate the number of words to the left of the ambiguous words, and the R columns the number of words to the right. So for example, the collocation 1L3R represents one word to the left and three words to the right of the ambiguous word. Collocations again were represented as lemmas (see Table 13) or parts-of-speech.

6 Experiment 3: Disambiguating Polysemous Verbs

6.1 Method

We tested the performance of our naive Bayes classifiers on the 67 genuinely ambiguous verbs on which the prior models were tested. Recall that these models were trained

Table 13
Features for collocations

L	R	Example
0	1	**write** you
1	0	can **write**
0	2	**write** you a
2	0	I can **write**
0	3	**write** you a story
3	0	perhaps I can **write**

L	R	Example
1	1	can **write** you
1	2	can **write** you a
2	1	I can **write** you
1	3	can **write** you a story
3	1	perhaps I can **write** you
2	4	I can **write** you a story sunshine

without access to a disambiguated corpus. The latter was only used to determine for a given verb and its frame its most likely meaning overall (i.e., across the corpus) instead of focusing on the meaning of individual corpus tokens. The same corpus was used for the disambiguation of individual tokens, excluding tokens assigned the class OTHER. The naive Bayes classifiers were trained and tested using 10-fold cross-validation on a set of 5,002 examples. These were representative of the frames 'NP1 V NP2', 'NP1 V NP2 NP3', 'NP1 V NP2 *to* NP3', and 'NP1 V NP2 *for* NP3'. The frame 'NP1 V *at* NP2' was excluded from our disambiguation experiments as it was represented solely by the verb *kick* (50 instances).

In this study we compare a naive Bayes classifier that relies on a uniform prior (see (32)) against two classifiers that make use of non-uniform prior models: the classifier in (34) effectively uses as prior the baseline model $P(c)$ from Section 2, whereas the classifier in (34) relies on the more informative model $P(c, f, v)$. As a baseline for the disambiguation task, we simply assign the most common class in the training data to every instance in the test data, ignoring context and any form of prior information (Pedersen, 2001; Gale, Church, and Yarowsky, 1992a). We also report an upper bound on disambiguation performance by measuring how well human judges agree with one another (percentage agreement) on the class assignment task. Recall from Section 4.1 that our corpus was annotated by two judges with Levin compatible verb classes (see Section 4.1).

6.2 Results

The results of our class disambiguation experiments are summarized in Figures 2–5. In order to investigate differences among different frames, we show how the naive Bayes classifiers perform for each frame individually. Figures 2–5 (x-axis) also reveal the influence of collocational features of different sizes (see Table 13) on the classification task. The (b) figures present the classifiers' accuracy when the collocational features are encoded as lemmas; in the (c) figures, the context is represented as parts-of speech, whereas in the (a) figures the context is represented by both lemmas and parts-of-speech.

As can be seen the naive Bayes classifier with our informative prior ($P(c, f, v)$, IPrior in Figures 2–5) generally outperforms the baseline prior ($P(c)$, BPrior in Figures 2–5), the uniform prior (UPrior in Figures 2–5), and the baseline (Baseline in Figures 2–5) for all frames. Good performances are attained with lemmas, parts-of speech and their combination. The naive Bayes classifier (IPrior) reaches the upper bound (UpBound in Figures 2–5) for the ditransitive frames NP1 V NP2 NP3, NP1 V *to* NP2 NP3, and NP1 V *for* NP2 NP3.

The best accuracy (87.8%) for the transitive frame is achieved with the collocational features 0L2R, 1L2R, and 1L3R (see Figures 2a–c). For the double object frame the high-

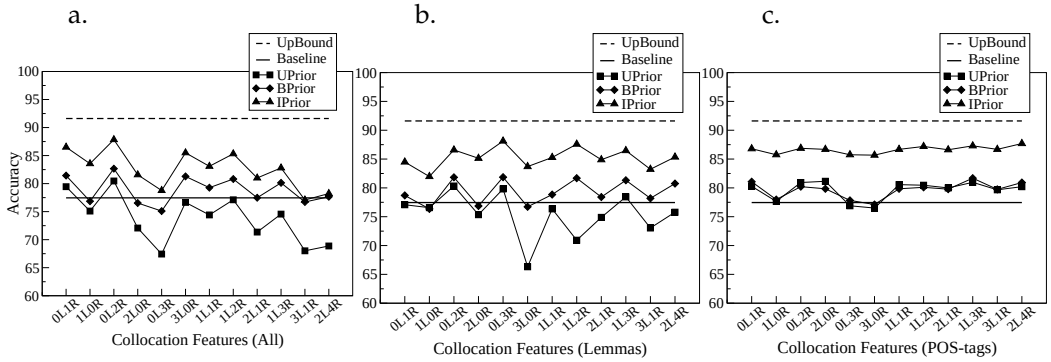

Figure 2
Word Sense Disambiguation accuracy for NP1 V NP2 frame

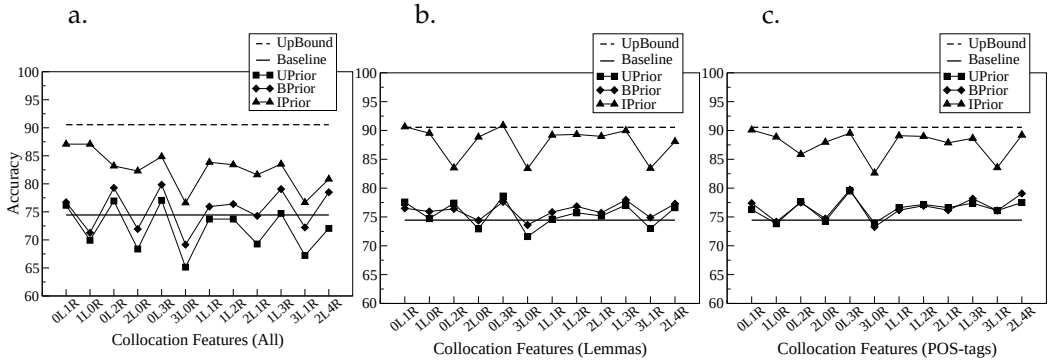

Figure 3
Word Sense Disambiguation accuracy for NP1 V NP2 NP3 frame

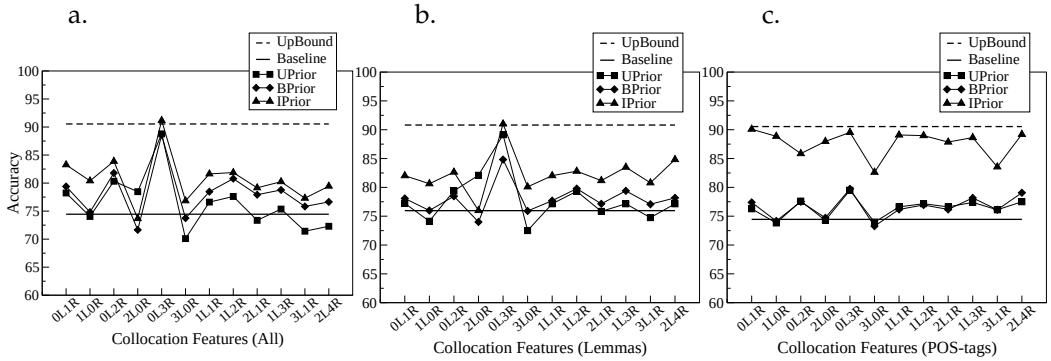

Figure 4
Word Sense Disambiguation accuracy for NP1 V NP2 *to* NP3 frame

est accuracy (90.8%) is obtained with features 0LR1 and 0L3R (see Figures 3b and 3c). Similarly, for the ditransitive 'NP1 V NP2 *to* NP3' frame, the features 0L3R and 0L1R yield the best accuracies (88.8%, see Figures 4a–c). Finally, for the 'NP1 V *for*' accuracy (94.4%) is generally good for most features when an informative prior is used. In fact, neither the uniform prior nor the baseline $P(c)$ outperform the baseline for this frame.

The context encoding (lemmas vs. parts-of-speech) does not seem to have a great influence on the disambiguation performance. Good accuracies are obtained with either

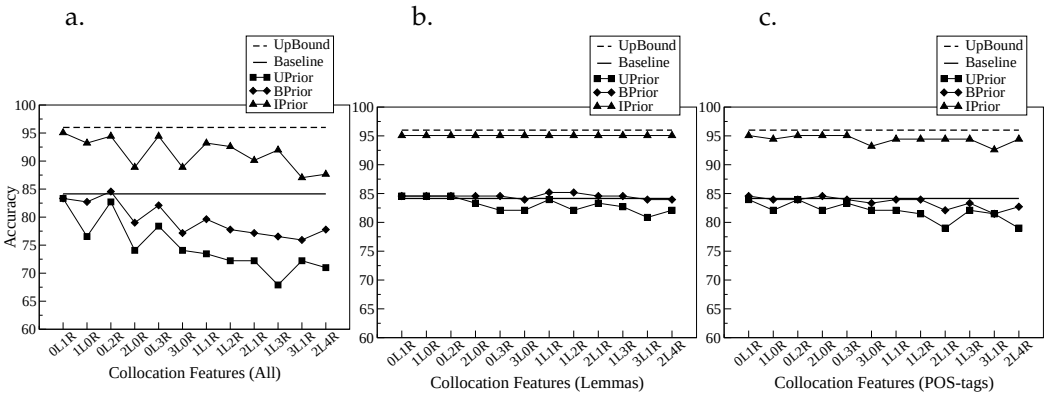

Figure 5
Word Sense Disambiguation accuracy for NP1 V NP2 *for* NP3 frame

parts-of-speech or lemmas; their combination does not yield better results.

The classifier with the informative prior $P(c, f, v)$ outperforms the baseline prior $P(c)$ and the uniform prior also when co-occurences are used. However, the co-occurrences never outperform the collocational features, for all four types of context. The classifiers (regardless of the type of prior being used) never beat the baseline for the frames 'NP1 V NP2' and 'NP1 V *to* N2'. Accuracies above the baseline are achieved for the frames NP1 V NP2 NP3' and 'NP1 V *for* N2' when an informative prior is used. Detailed results are summarized in Appendix A. Co-occurrences and windows of large sizes traditionally work well for topical distinctions (Gale, Church, and Yarowsky, 1992b). Levin classes, however, typically capture differences in argument structure, i.e., the types of objects or subjects that verbs select for. Argument structure is approximated by our collocational features. For example, a verb often taking a reflexive pronoun as its object is more likely to be a REFLEXIVE VERB OF APPEARANCE than a verb which never subcategorizes for a reflexive object. There is not enough variability among the wider contexts surrounding a polysemous verb to inform the class disambiguation task, as the Levin classes often do not cross topical boundaries.

7 Discussion

In this paper we presented a probabilistic model of verb class ambiguity based on Levin's (1993) semantic classification. Our results show that subcategorization information acquired automatically from corpora provides important cues for verb class disambiguation (Experiment 1). In the absence of subcategorization cues, corpus-based distributions and quantitative approximations of linguistic concepts can be used to derive a preference ordering on the set of verbal meanings (Experiment 2). The semantic preferences which we generate can be thought of as default semantic knowledge, to be used in the absence of any explicit contextual or lexical semantic information to the contrary (see Tables 12). We also show that these preferences are useful for disambiguating polysemous verbs within their local contexts of occurrence (Experiment 3).

The approach is promising in that it achieves satisfactory results with a simple model which has a straightforward interpretation in a Bayesian framework and does not rely on the availability of annotated data. The model's parameters are estimated using simple distributions that can be easily extracted from corpora. Our model achieved an accuracy of 93.9% (over a baseline of 56.7%) on the class disambiguation task (Experiment 1) and an accuracy of 74.6% (over a baseline of 46.2%) on task of deriving

dominant verb classes (Experiment 2). Our disambiguation experiments reveal that this default semantic knowledge when incorporated as a prior in a naive Bayes classifier outperforms the uniform prior, and the baseline of always defaulting to the most frequent class (Experiment 3). In fact, for three out of the four frames under study, our classifier with the informative prior achieved upper bound performance.

Although our results are promising, it remains to be shown that they generalize across frames and alternations. Four types of alternations were investigated in this study. However, Levin lists 79 alternations and approximately 200 classes. Although distributions for different class/frame combinations can be easily derived automatically, it remains to be shown that these distributions are useful for all verbs, frames, and classes. Also note that the models described in the previous sections crucially rely on the acquisition of relatively accurate frames from the corpus. It is a matter of future work to examine how the quality of the acquired frames influences the disambiguation task. Also, the assumption that the semantic class determines the subcategorization patterns of its class members independently of their identity may not be harmless for all classes and frames.

Although our original aim was to develop a probabilistic framework which exploits Levin's (1993) linguistic classification and the systematic correspondence between syntax and semantics, a limitation of the model is that it cannot infer class information for verbs not listed in Levin. For these verbs $P(c)$, and hence $P(c, f, v)$, will be zero. Recent work in computational linguistics (e.g., Schütze (1998)) and cognitive psychology (e.g., Landauer and Dumais (1997)) has shown that large corpora implicitly contain semantic information, which can be extracted and manipulated in the form of co-occurrence vectors. One possible approach would be to compute the centroid (geometric mean) of the vectors of all members of a semantic class. Given an unknown verb (i.e., a verb not listed in Levin) we can decide its semantic class by comparing its semantic vector to the centroids of all semantic classes. For example, we could determine class membership on the basis of the closest distance to the centroid representing a semantic class (see Patel, Bullinaria, and Levy (1998) for a proposal similar in spirit). Another approach put forward by Dorr and Jones (1996) utilizes WordNet (Miller and Charles, 1991) to find similarities (via synonymy) between unknown verbs and verbs listed in Levin. Once we have chosen a class for an unknown verb, we are entitled to assume that it will share the broad syntactic and semantic properties of that class.

8 Related Work

Levin's (1993) seminal study on diathesis alternations and verb semantic classes has recently influenced work in dictionary creation (Dorr, 1997; Dang et al., 1998; Dorr and Jones, 1996) and notably lexicon acquisition on the basis of the assumption that verbal meaning can be gleaned from corpora using cues pertaining to syntactic structure (Merlo and Stevenson, 2001; Schulte im Walde, 2000; Lapata, 1999; McCarthy, 2000). Previous work in word sense disambiguation has not tackled explicitly the ambiguity problems arising from Levin's classification although methods for deriving informative priors in an unsupervised manner have been proposed by Ciaramita and Johnson (2000) and Chao and Dyer (2000) within the context of noun and adjective sense disambiguation, respectively. In this section we review related work on classification and lexicon acquisition and compare it to our own work.

Dang et al. (1998) observe that verbs in Levin's (1993) database are listed in more than one class. The precise meaning of this ambiguity is left open to interpretation in Levin as it may indicate that the verb has more than one sense or that one sense (i.e., class) is primary and the alternations for this class should take precedence over

the alternations for the other classes for which the verb is listed. They augment Levin's semantic classes with a set of "intersective" classes which are created by grouping together sets of existing classes which share a minimum of three overlapping members. Intersective classes are more fine-grained than the original Levin classes and exhibit more coherent sets of syntactic frames and associated semantic components. Dang et al. further argue that intersective classes are more compatible with WordNet than the broader Levin classes and thus make it possible to attribute the semantic components and associated sets of syntactic frames to specific WordNet senses as well, thus enriching the WordNet representation, and providing explicit criteria for word sense disambiguation.

Most statistical approaches, including ours, treat verbal meaning assignment as a semantic classification task. The underlying question is the following: how can corpus information be exploited in deriving the semantic class for a given verb? Despite the unifying theme of using corpora and corpus distributions for the acquisition task, the approaches differ in the inventory of classes they employ, in the methodology used for inferring semantic classes and the specific assumptions concerning the verbs to be classified (i.e., can they be polysemous or not).

Merlo and Stevenson (2001) use grammatical features (acquired from corpora) to classify verbs into three semantic classes: unergative, unaccusative, and object-drop. These classes are abstractions of Levin's (1993) classes and as a result yield a coarser classification. For example, object-drop verbs comprise a variety of Levin classes such as GESTURE verbs, CARING verbs, LOAD verbs, PUSH-PULL verbs, MEET verbs, SOCIAL INTERACTION verbs, AMUSE verbs, etc. Unergative, unaccusative, and object-drop verbs have identical subcategorization patterns (i.e., they alternate between the transitive and intransitive frame), yet distinct argument structures and therefore differ in the thematic roles they assign to their arguments. For example, when attested in the intransitive frame the subject of an object-drop verb is an Agent, whereas the subject of an unaccusative verb is a Theme. Under the assumption that differences in thematic role assignment uniquely identify semantic classes, numeric approximations of argument structure are derived from corpora and used in a machine learning paradigm to classify verbs in their semantic classes. The approach is evaluated on 59 verbs manually selected from Levin (20 unergatives, 20 object-drop, and 19 unaccusatives). It is assumed that these verbs are monosemous, i.e., they can be either ergative, unergative or object-drop. A decision-tree learner achieves a accuracy of 69.8% on the classification task over a chance baseline of 34%.

Schulte im Walde (2000) uses subcategorization information and selectional restrictions to cluster verbs into Levin (1993) compatible semantic classes. Subcategorization frames are induced from the BNC using a robust statistical parser (Carroll and Rooth, 1998). The selectional restrictions are acquired using Resnik's (1993) information-theoretic measure of selectional association which combines distributional and taxonomic information (e.g., WordNet) in order to formalize how well a predicate associates with a given argument. Two sets of experiments are run to evaluate the contribution of selectional restrictions using two types of clustering algorithms, iterative clustering and latent class clustering (see Schulte im Walde, 2000 for details). The approach is evaluated on 153 verbs taken from Levin, 53 of which are polysemous (i.e., belong to more than one class). The size of the derived clusters is restricted to four verbs and compared to Levin: verbs are classified correctly if they are members of a non-singleton cluster which is a subset of a Levin class. Polysemous verbs can be assigned to distinct clusters only using the latent class clustering method. The best results achieve a recall of 36% and a precision of 61% (over a baseline of 5%, calculated as the number of randomly created clusters which are subsets of a Levin class) using subcategorization information only

and iterative clustering. Inclusion of information about selectional restrictions yields a lower accuracy of 38% (with a recall of 20%), again using iterative clustering.

Dorr and Jones (1996) use Levin's (1993) classification to show that there is a predictable relationship between verbal meaning and syntactic behavior. They create a database of Levin verb classes and the sentences exemplifying them (including both positive and negative examples, i.e., examples marked with asterisks). A parser is used to extract basic syntactic patterns for each semantic class. These patterns form the syntactic signature of the class. 97.9% of the semantic classes are identified uniquely by their syntactic signatures. Grouping verbs (instead of classes) with identical signatures to form a semantic class yields a 6.3% overlap with Levin classes. Their results are somewhat difficult to interpret since in practice information about a verb and its syntactic signature is not available, and it is precisely this information that is crucial for classifying verbs into Levin classes. Schulte im Walde's study and our own study shows that acquisition of syntactic signatures (i.e., subcategorization frames) from corpora is feasible, however these acquired signatures are not necessarily compatible with Levin and in most cases will depart from those derived by Dorr and Jones as negative examples are not available in real corpora.

Ciaramita and Johnson (2000) propose an unsupervised Bayesian model for disambiguating verbal objects that uses WordNet's inventory of senses. For each verb the model creates a Bayesian network whose architecture is determined by Wordnet's hierarchy and whose parameters are estimated from a list of verb-object pairs found in a corpus. A common problem for unsupervised models trained on verb-object tuples is that the objects can belong to more than one semantic class. The class ambiguity problem is commonly resolved by considering each observation of an object as evidence for each of the classes the word belongs to. The formalization of the problem in terms of Bayesian networks, allows to weight the contribution of different senses via *explaining away* (Pearl, 1988): if A is a hyponym of B and C is a hyponym of B, and B is true, then finding that C is true makes A less likely.

Prior knowledge about the likelihoods of concepts is hand-coded in the network according to the following principles: (a) it is unlikely that any given class will be *a priori* selected for, (b) if a class is selected, then its hyponyms are also likely to be selected, (c) a word is likely as the object of a verb, if at least one of its classes is selected for. Likely and unlikely here correspond to numbers that sum up to to one. Ciaramita and Johnson show that their model outperforms other word sense disambiguation approaches that do not make use of prior knowledge.

Chao and Dyer (2000) propose a method for the disambiguation of polysemous adjectives in adjective-noun combinations which also uses Bayesian networks and WordNet's taxonomic information. Prior knowledge about the likelihood of different senses or semantic classes is derived heuristically by submitting queries (e.g., *great hurricane*) to the Altavista search engine and extrapolating from the number of returned documents the frequency of the adjective-noun pair (see (Mihalcea and Moldovan, 1998) for details of this technique). Fore each polysemous adjective-noun combination, the synonyms representative of each sense are retrieved from WordNet (e.g., {*great*, *large*, *big*} vs. {*great*, *neat*, *good*}). Queries are submitted to Altavista for each synonym-noun pair; the number of documents returned is used then as an estimate of how likely the different adjective senses are. Chao and Dyer obtain better results when prior knowledge is factored into their Bayesian network.

Our work focuses on the ambiguity inherently present in Levin's (1993) classification. The problem is ignored by Merlo and Stevenson (2001) who focus only on monosemous verbs. Polysemous verbs are included in Schulte im Walde's (2000) experiments: the clustering approach can go so far as to identify more than one class for a given verb

without, however, providing information about its dominant class. We recast Levin's classification in a statistical framework and show in agreement with Merlo and Stevenson and Schulte im Walde that corpus-based distributions provide important information for semantic classification, especially in the case of polysemous verbs whose meaning cannot be easily inferred from the immediate surrounding context (i.e., subcategorization). We additionally show that the derived model is not only useful for determining the most likely overall class for a given verb (i.e., across the corpus) but also for disambiguating polysemous verb tokens in context.

Like Schulte im Walde (2000), our approach relies on subcategorization frames extracted from the BNC (although using a different methodology). We employ Levin's inventory of semantic classes arriving at a finer grained classification than Merlo and Stevenson (2001). In contrast to Schulte im Walde we do not attempt to discover Levin classes from corpora; instead, we exploit Levin's classification and corpus frequencies in order to derive a distribution of verbs, classes and their frames, that is not known a priori but is approximated using simplifications. Our approach is not particularly tied to Levin's exact classification. We presented a general framework which could be extended to related classifications such as the semantic hierarchy proposed by Dang et al. (1998). In fact the latter may be more appropriate for our disambiguation experiments as it is based on a tighter correspondence between syntactic frames and semantic components and contains links to the WordNet taxonomy.

Prior knowledge with regard to the likelihood of polysemous verb classes is acquired automatically in an unsupervised manner by combining corpus frequencies estimated from the BNC and information inherent in Levin. The models proposed by Chao and Dyer (2000) and Ciaramita and Johnson (2000) are not directly applicable to Levin's classification as the latter is not a hierarchy (and therefore not a DAG) and cannot be straightforwardly mapped into a Bayesian network. However, in agreement with Chao and Dyer and Ciaramita and Johnson we show that prior knowledge about class preferences improves word sense disambiguation performance.

Unlike Schulte im Walde (2000) and Merlo and Stevenson (2001), we ignore information about the arguments of a given verb either in the form of selectional restrictions or argument structure while building our prior models. The latter information is however indirectly taken into account in our disambiguation experiments: the verbs' arguments are features for our naive Bayes classifiers. Such information can be also incorporated into the prior model in the form of conditional probabilities where the verb is, for example, conditioned on the thematic role of its arguments if this is known (see (Gildea and Jurafsky, 2000) for a method that automatically labels thematic roles). Unlike Stevenson and Merlo, Schulte im Walde, and Dorr and Jones (1996) we provide a general probabilistic model which assigns a probability to each class of a given verb by calculating the probability of a complex expression in terms of the probability of simpler expressions that compose it. We further show that this model is useful for disambiguating polysemous verbs in context.

Acknowledgments

Mirella Lapata was supported by ESRC grant number R000237772. Thanks to Frank Keller, Alex Lascarides, Katja Markert, Paola Merlo, Sabine Schulte im Walde, Stacey Bailey, Markus Dickinson,Anna Feldman and Anton Rytting, and two anonymous reviewers for valuable comments.

References

Boguraev, Branimir K. and Ted Briscoe. 1989. Utilising the LDOCE grammar codes. In Ted Briscoe and Branimir K. Boguraev, editors, *Computational Lexicography for Natural Language Processing*. Longman, London, pages 85–116.

Briscoe, Ted and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th*

Conference on Applied Natural Language Processing, pages 356–363, Washington, DC.

Briscoe, Ted and Ann Copestake. 1999. Lexical rules in constraint-based grammar. *Computational Linguistics*, 25(4):487–526.

Burnard, Lou, 1995. *Users Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Service.

Carroll, Glenn and Mats Rooth. 1998. Valence induction with a head-lexicalized PCFG. In Nancy Ide and Atro Voutilainen, editors, *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, pages 36–45, Granada, Spain.

Chao, Gerald and Michael G. Dyer. 2000. Word sense disambiguation of adjectives using probabilistic networks. In COLING (COLING, 2000), pages 152–158.

Ciaramita, Massimiliano and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional restrictions with Bayesian networks. In COLING (COLING, 2000), pages 187–193.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

2000. *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken, Germany.

1998. *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montréal, Canada.

Corley, Steffan, Martin Corley, Frank Keller, Matthew W. Crocker, and Shari Trewin. 2001. Finding syntactic structure in unparsed corpora: The Gsearch corpus query system. *Computers and the Humanities*, 35(2):81–94.

Cucerzan, Silviu and David Yarowsky. 2002. Augmented mixture models for lexical disambiguation. In Jan Hajič and Yuji Matsumoto, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 33–40, Philadelphia, PA.

Dang, Hoa Trang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig. 1998. Investigating regular sense extensions based on intersective Levin classes. In COLING/ACL (COLING/ACL, 1998), pages 293–299.

Dang, Hoa Trang, Joseph Rosenzweig, and Martha Palmer. 1997. Associating semantic components with intersective Levin classes. In *Proceedings of the 1st AMTA SIG-IL Workshop on Interlinguas*, pages 1–8, San Diego, CA.

Dorr, Bonnie J. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):371–322.

Dorr, Bonnie J. and Doug Jones. 1996. Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 322–327, Copenhagen, Denmark.

Duda, Richard O. and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis*. Wiley, NY.

Fillmore, Charles. 1965. *Indirect Object Constructions and the Ordering of Transformations*. Mouton, The Hague.

Gale, William, Kenneth Ward Church, and David Yarowsky. 1992a. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 249–256, Columbus, OH.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992b. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5–6):415–439.

Gildea, Daniel and Daniel Jurafsky. 2000. Automatic labelling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages ??–??, Hong Kong.

Goldberg, Adele. 1995. *Constructions*. Chicago University Press, Chicago.

Green, Georgia. 1974. *Semantics and Syntactic Regularity*. Indiana University Press, Bloomington.

Gropen, J, S Pinker, M Hollander, M Goldberg, and R Wilson. 1989. The learnability and acquisition of the dative alternation. *Language*, 65(2):203–257.

Hindle, Donald and Mats Rooth. 1993. Structural ambiguity and lexical relations. *omputational Linguistics*, 19(1):103–120.

Ide, Nancy and Jean Véronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

Jackendoff, Ray. 1983. *Semantic and Cognition*. The MIT Press, Cambridge, MA.

Johnson, W. E. 1932. Probability: The deductive and inductive problems. *Mind*, 49:409–423.

Kipper, Karin, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of

a verb lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pages 691–696, Austin, TX.

Klavans, Judith and Min-Yen Kan. 1998. Role of verbs in document analysis. In COLING/ACL (COLING/ACL, 1998), pages 680–688.

Kupiec, Julian. 1992. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, 6(3):225–242.

Landauer, Thomas K. and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Lapata, Maria. 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 397–404, College Park, MD.

Lapata, Maria. 2001. *The Acquisition and Modeling of Lexical Knowledge: A Corpus-based Investigation of Systematic Polysemy*. Ph.D. thesis, University of Edinburgh.

Lauer, Mark. 1995. *Designing Statistical Language Learners: Experiments on Compound Nouns*. Ph.D. thesis, Macquarie University.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Levow, Gina-Anne, Bonnie Dorr, and Dekang Lin. 2000. Construction of chinese-english semantic hierarchy for information retrieval. Technical report, University of Maryland, College Park.

McCarthy, Diana. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In North American Annual Meeting of the Association for Computational Linguistics (North American Annual Meeting of the Association for Computational Linguistics, 2000), pages 256–263.

Merlo, Paola and Susanne Stevenson. 2001. Automatic verb classification based on statistical distribution of argument structure. *Computational Linguistics*, 27(3):373–408.

Mihalcea, Rada and Dan Moldovan. 1998. Word sense disambiguation based on semantic density. In Sanda Harabagiu, editor, *Proceedings of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing*, pages 16–22, Montréal, Canada.

Miller, George A. and William G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Mooney, Raymond J. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In Eric Brill and Kenneth Church, editors, *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing*, pages 82–91, Philadelphia, PA.

Ng, Hwee Tou. 1997. Exemplar-based word sense disambiguation: Some recent hnprovements. In Claire Cardie and Ralph Weischedel, editors, *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, pages 208–216, Providence, Rhodes Island.

2000. *Proceedings of the 1st North American Annual Meeting of the Association for Computational Linguistics*, Seattle, WA.

Palmer, Martha. 2000. Consistent criteria for sense distinctions. *Computers and the Humanities*, 34(1–2):217–222.

Palmer, Martha and Zhibiao Wu. 1995. Verb semantics for english-chinese translation. *Machine Translation*, 10:59–92.

Patel, Malti, John A. Bullinaria, and Joseph P. Levy. 1998. Extracting semantic representations from large text corpora. In John A. Bullinaria, D. W. Glasspool, and G. Houghton, editors, *In Proceedings of the 4th Workshop on Neural Computation and Psychology*, pages 199–212. Springer, Berlin.

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pedersen, Ted. 2000. A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. In North American Annual Meeting of the Association for Computational Linguistics (North American Annual Meeting of the Association for Computational Linguistics, 2000), pages 63–69.

Pedersen, Ted. 2001. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the 2nd North American Annual Meeting of the Association for Computational Linguistics*, pages 63–69, Pittsburgh, PA.

Pedersen, Ted and Rebecca Bruce. 1998. Knowledge lean word-sense disambiguation. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pages 800–805, Madison, WI.

Pinker, Steven. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. The MIT Press, Cambridge, MA.

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

Schulte im Walde, Sabine. 2000. Clustering verbs semantically according to their alternation behaviour. In COLING (COLING, 2000), pages 747–753.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

Stede, Manfred. 1998. A generative perspective on verb alternations. *Computational Linguistics*, 24(3):401–430.

Talmy, L. 1985. Lexicalisation patterns: Semantic structure in lexical forms. In T Shopen, editor, *Language Typology and Syntactic Description III: Grammatical Categories and the Lexicon*. Cambrige University Press, Cambridge, pages 57–149.

Yarowsky, David. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 88–95, Las Cruces, NM.

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, MA.

Appendix A. Disambiguation Results with Co-occurrences

Figures 6–9 show the performances of our naive Bayes classifier when co-occurrences are used as features. We experimented with four types of context: left context (Left), right context (Right), sentential context (Sentence), and the sentence within which the ambiguous verb is found together with its immediately preceding sentence (PSentence). The context was encoded as lemmas or parts-of-speech.

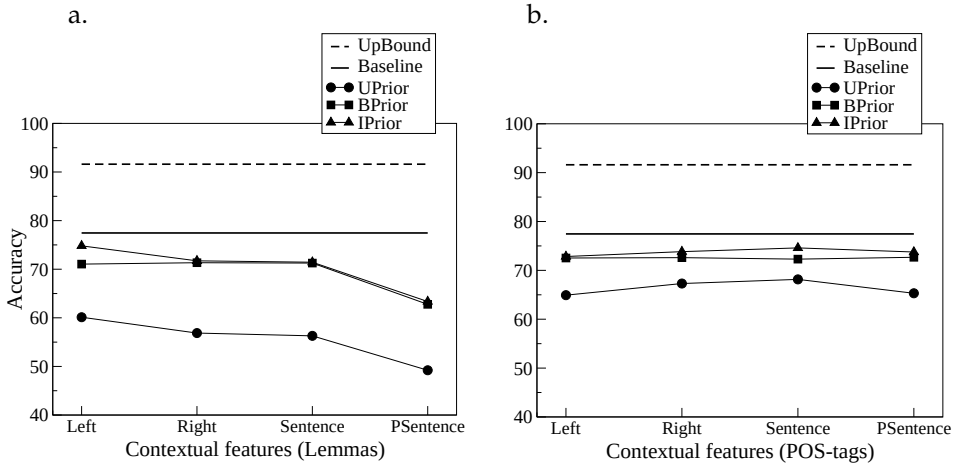

Figure 6
Word Sense Disambiguation accuracy for NP1 V NP2 frame

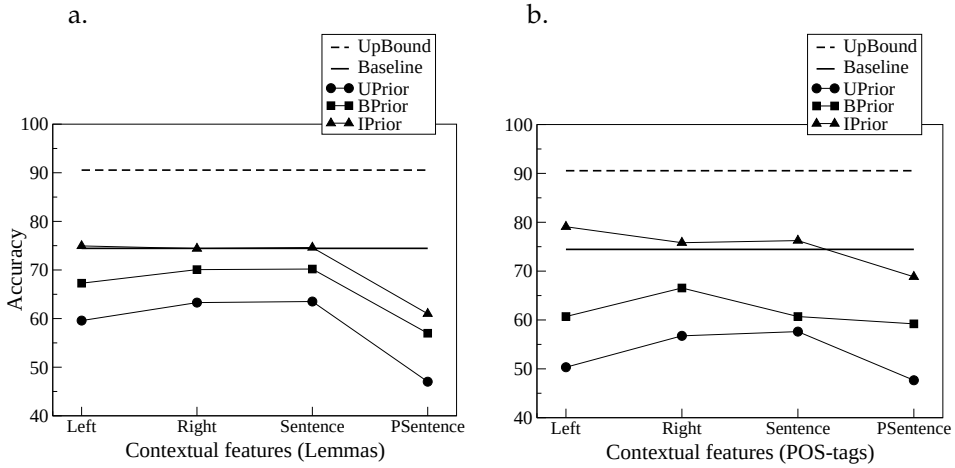

Figure 7
Word Sense Disambiguation accuracy for NP1 V NP2 NP3 frame

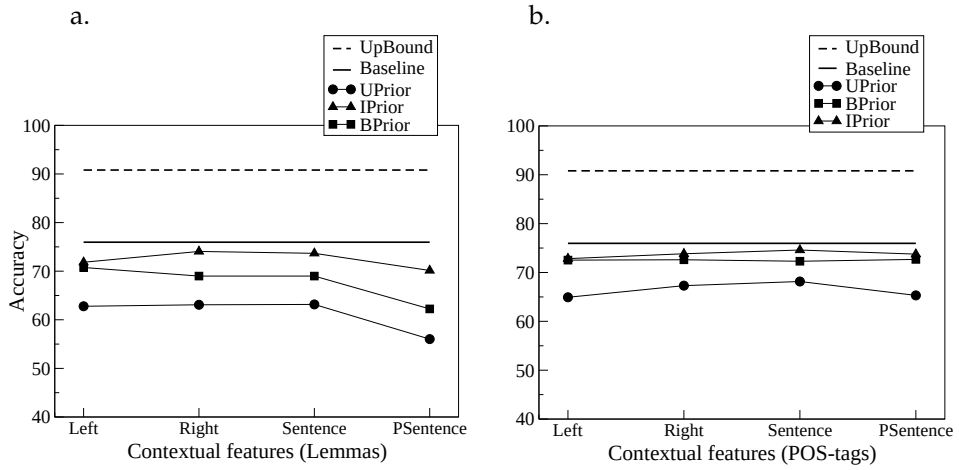

Figure 8
Word Sense Disambiguation accuracy for NP1 V *to* NP2 NP3 frame

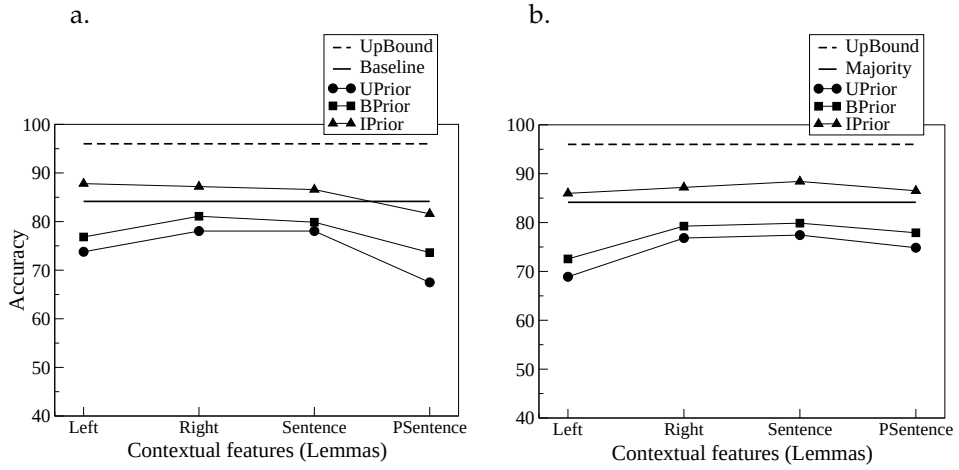

Figure 9
Word Sense Disambiguation accuracy for NP1 V *for* NP2 NP3 frame