WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery

Philip Resnik*
Department of Computer and Information Science
University of Pennsylvania, Philadelphia, PA 19104, USA
resnik@linc.cis.upenn.edu

Introduction

It has become common in statistical studies of natural language data to use measures of lexical association, such as the information-theoretic measure of mutual information, to extract useful relationships between words (e.g. [Church *et al.*, 1989; Church and Hanks, 1989; Hindle, 1990]). For example, [Hindle, 1990] uses an estimate of mutual information to calculate what nouns a verb can take as its subjects and objects, based on distributions found within a large corpus of naturally occurring text.

Lexical association has its limits, however, since oftentimes either the data are insufficient to provide reliable word/word correspondences, or a task requires more abstraction than word/word correspondences permit. In this paper I present a generalization of lexical association techniques that addresses these limitations by facilitating statistical discovery of facts involving word *classes* rather than individual words. Although defining association measures over classes (as sets of words) is straightforward in theory, making direct use of such a definition is impractical because there are simply too many classes to consider. Rather than considering all possible classes, I propose constraining the set of possible word classes by using WordNet [Beckwith *et al.*, 1991; Miller, 1990], a broad-coverage lexical/conceptual hierarchy.

Section 1 very briefly reviews mutual information as it is commonly used to discover lexical associations, and Section 2 discusses some limitations of this approach. In Section 3 I apply mutual information to word/class relationships, and propose a knowledge-based interpretation of "class" that takes advantage of the WordNet taxonomy. In Section 4 I apply the technique to a problem involving verb argument relationships, and in Sections 5 and 6 I mention some related work and suggest further applications for which such class-based investigations might prove useful.

*I would like to acknowledge the support of an IBM graduate fellowship, and to thank Eric Brill, Aravind Joshi, David Magerman, Christine Nakatani, and Michael Niv for their helpful comments. This research was also supported in part by the following grants: ARO DAAL 03-89-C-0031, DARPA N00014-90-J-1863, NSF IRI 90-16592, and Ben Franklin 91S.3078C-1.

Word/Word Relationships

Mutual information is an information-theoretic measure of association frequently used with natural language data to gauge the "relatedness" between two words x and y. The mutual information of x and y is defined as follows:

$$\mathrm{I}(x;y) = \log \frac{\Pr(x,y)}{\Pr(x)\Pr(y)} \quad (1)$$

Intuitively, the probability of seeing x and y together, $\Pr(x,y)$, gives some idea as to how related they are. However, if x and y are both very common, then it is likely that they appear together frequently simply by chance and not as a result of any relationship between them. In order to correct for this possibility, $\Pr(x,y)$ is divided by $\Pr(x)\Pr(y)$, which is the probability that x and y would have of appearing together by chance if they were independent. Taking the logarithm of this ratio gives mutual information some desirable properties; for example, its value is respectively positive, zero, or negative according to whether x and y appear together more frequently, as frequently, or less frequently than one would expect if they were independent.[1]

Let us consider two examples of how mutual information can be used. [Church and Hanks, 1989] define the *association ratio* between two words as a variation on mutual information in which word y must follow word x within a window of w words. The association ratio is then used to discover "interesting associations" within a large corpus of naturally occurring text — in this case, the 1987 Associated Press corpus. Table 1 gives some of the word pairs found in this manner. An application of such information is the discovery of interesting lexico-syntactic regularities. Church and Hanks write:

[1] The definition of mutual information is also related to other information-theoretic notions such as relative entropy; see [Cover and Thomas, 1991] for a clear discussion of these and related topics. See [Church *et al.*, 1991] for a discussion of mutual information and other statistics as applied to lexical analysis.

Association ratio	x	y
11.3	honorary	doctor
11.3	doctors	dentists
10.7	doctors	nurses
9.4	doctors	treating
9.0	examined	doctor
8.9	doctors	treat
8.7	doctor	bills

Table 1: *Some interesting associations with "Doctor" (data from Church and Hanks 1989)*

Co-occurrence score	verb	object
11.75	drink	tea
11.75	drink	Pepsi
11.75	drink	champagne
10.53	drink	liquid
10.20	drink	beer
9.34	drink	wine

Table 2: *High-scoring verb/object pairs for* drink *(This is a portion of Hindle 1990, Table 2.)*

> . . . We believe the association ratio can also be used to search for interesting lexico-syntactic relationships between verbs and typical arguments/adjuncts. . . . [For example, i]t happens that *set ... off* was found 177 times in the 1987 AP Corpus of approximately 15 million words . . . Quantitatively, $I(set, off) = 5.9982$, indicating that the probability of *set ... off* is almost 64 greater than chance.

As a second example, consider Hindle's [1990] application of mutual information to the discovery of predicate argument relations. Unlike Church and Hanks, who restrict themselves to surface distributions of words, Hindle investigates word co-occurrences as mediated by syntactic structure. A six-million-word sample of Associated Press news stories was parsed in order to construct a collection of subject/verb/object instances. On the basis of these data, Hindle calculates a *co-occurrence score* (an estimate of mutual information) for verb/object pairs and verb/subject pairs. Table 2 shows some of the verb/object pairs for the verb *drink* that occurred more than once, ranked by co-occurrence score, "in effect giving the answer to the question 'what can you drink?' " [Hindle, 1990, p. 270].

Word/Word Limitations

The application of mutual information at the level of word pairs suffers from two obvious limitations:

- the corpus may fail to provide sufficient information about relevant word/word relationships, and
- word/word relationships, even those supported by the data, may not be the appropriate relationships to look at for some tasks.

Let us consider these limitations in turn.

Co-occurrence score	verb	object	count
5.8470	open	closet	2
5.8470	open	chapter	2
5.7799	open	mouth	7
5.5840	open	window	5
5.4320	open	store	2
5.4045	open	door	26
5.0170	open	scene	3
4.6246	open	season	2
4.2621	open	church	2
3.6246	open	minute	2
3.2621	open	eye	7
3.1101	open	statement	2
1.5991	open	time	2
1.3305	open	way	3

Table 3: *Verb/object pairs for* open *(with count > 1)*

First, as in all statistical applications, it must be possible to estimate probabilities accurately. Although larger and larger corpora are increasingly available, the specific task under consideration often can restrict the choice of corpus to one that provides a smaller sample than necessary to discover all the lexical relationships of interest. This can lead some lexical relationships to go unnoticed.

For example, the Brown Corpus [Francis and Kucera, 1982] has the attractive (and for some tasks, necessary) property of providing a sample that is balanced across many genres. In an experiment using the Brown Corpus, modelled after Hindle's [1990] investigation of verb/object relationships, I calculated the mutual information between verbs and the nouns that appeared as their objects.[2] Table 3 shows objects of the verb *open*; as in Table 2, the listing includes only verb/object pairs that were encountered more than once.

Attention to the verb/object pairs that occurred *only* once, however, led to an interesting observation. Included among the "discarded" object nouns was the following set: *discourse*, *engagement*, *reply*, *program*, *conference*, and *session*. Although each of these appeared as the object of *open* only once — certainly too infrequently to provide reliable estimates — this set, *considered as a whole*, reveals an interesting fact about some kinds of things that can be opened, roughly captured by the notion of *communications*. More generally, several pieces of statistically unreliable information at the lexical level may nonetheless capture a useful statistical regularity when combined. This observation motivates an approach to lexical association that makes such combinations possible.

A second limitation of the word/word relationship is simply this: some tasks are not amenable to treatment using lexical relationships alone. An ex-

[2] The experiment differed from that described in [Hindle, 1990] primarily in the way the direct object of the verb was identified; see Section 4 for details.

ample is the automatic discovery of verb selectional preferences for natural language systems. Here, the relationship of interest holds not between a verb and a noun, but between the verb and a *class* of nouns (e.g. between *eat* and nouns that are [+EDIBLE]). Given a table built using lexical statistics, such as Table 2 or 3, no single lexical item necessarily stands out as the "preferred" object of the verb — the selectional restriction on the object of *drink* should be something like "beverages," not *tea*. Once again, the limitation seems to be one that can be addressed by considering sets or classes of nouns, rather than individual lexical items.

Word/Class Relationships

A Measure of Association

In this section, I present a method for investigating *word/class* relationships in text corpora on the basis of mutual information, using for illustration the problem of finding "prototypical" object classes for verbs. As will become evident, the generalization from word/word relationships to word/class relationships addresses the two limitations discussed in the previous section.

Let

$$\begin{aligned} \mathcal{V} &= \{v_1, v_2, \ldots, v_l\} \\ \mathcal{N} &= \{n_1, n_2, \ldots, n_m\} \end{aligned}$$

be the sets of all verbs and all nouns, respectively, and let

$$\mathcal{C} = \{c | c \subseteq \mathcal{N}\}$$

be the set of noun classes; that is, the power set of $\mathcal{N}$. Since the relationship being investigated holds between verbs and classes of their objects, the elementary events of interest are members of $\mathcal{V} \times \mathcal{C}$. The joint probability of a verb and a class is estimated as

$$\Pr(v, c) \approx \frac{\sum_{n \in c} count(v, n)}{\sum_{v' \in \mathcal{V}} \sum_{n' \in \mathcal{N}} count(v', n')}. \quad (2)$$

Given $v \in \mathcal{V}$, $c \in \mathcal{C}$, define the *association score*

$$A(v, c) \triangleq \Pr(c|v) \log \frac{\Pr(v, c)}{\Pr(v)\Pr(c)} \quad (3)$$

$$= \Pr(c|v) \mathrm{I}(v; c) \quad (4)$$

The association score takes the mutual information between the verb and a class, and scales it according to the likelihood that a member of the class will actually appear as the object of the verb.[3]

[3]Scaling the pointwise mutual information score in this fashion is relatively common in practice; cf. the score used by [Gale *et al.*, 1992] and the "expected utility" of [Rosenfeld and Huang, 1992].

Coherent Classes

The search for a verb's most typical object nouns requires at most $|\mathcal{N}|$ evaluations of the association function, and can thus be done exhaustively. An exhaustive search among object *classes* is impractical, however, since the number of classes is exponential. Clearly some way to constrain the search is needed.

Let us consider the possibility of restricting the search by imposing some requirement of *coherence* upon the classes to be considered. For example, considering objects of *open*, the class $\{closet, locker\}$ is more coherent than $\{closet, discourse\}$ on intuitive grounds: every noun in the former class describes a repository of some kind, whereas the latter class has no such obvious interpretation.

There are two ways to arrive at a catalogue of coherent noun classes. First, one can take a "knowledge-free" approach, deriving a set of classes or a class hierarchy on the basis of distributional analysis. [Hindle, 1990] describes one distributional approach to classifying nouns, based upon the verbs for which they tend to appear as arguments. [Brown *et al.*, 1990] and [Brill *et al.*, 1990] propose algorithms for building word class hierarchies using n-gram distributions.[4] Approaches such as these share some of the common advantages and disadvantages of statistical methods. On the positive side, they fine-tune themselves to fit the data at hand: the classes discovered by these techniques are exactly those for which there is evidence in the particular corpus under consideration. On the negative side, there may be useful classes for which there is nonetheless too little evidence in the data — this may result from low counts, or it may be that what makes the class coherent (e.g., some shared semantic feature) is not reflected sufficiently by lexical distributions.

A second possibility is to take a "knowledge-based" approach, constraining possible noun classes to be those that appear within a hand-crafted lexical knowledge base or semantic hierarchy. Here, too, there are known advantages and disadvantages associated with knowledge-based systems. On the positive side, a knowledge base built by hand captures regularities that may not be easily recoverable statistically, and it can be organized according to principles and generalizations that seem appropriate to the researcher regardless of what evidence there is in any particular corpus. On the negative side, the decisions made in building a knowledge base may be based upon intuitions (or theoretical biases) that are not fully supported by the data; worse, few knowledge bases, regardless of quality, are broad enough in scope to support work based upon corpora of unconstrained text.

Clearly there is no foolproof way to choose between the knowledge-free and knowledge-based ap-

[4]Noun classification has even been investigated within a connectionist framework [Elman, 1990], although to my knowledge no one has attempted a practical demonstration using a non-trivial corpus.

proaches. For largely practical reasons, I have chosen to adopt a knowledge-based approach, using the WordNet lexical database [Beckwith *et al.*, 1991; Miller, 1990]. WordNet is a lexical/conceptual database constructed as robustly as possible using psycholinguistic principles. Miller and colleagues write:

> . . . How the leading psycholinguistic theories should be exploited for this project was not always obvious. Unfortunately, most research of interest for psycholexicology has dealt with relatively small samples of the English lexicon . . . All too often, an interesting hypothesis is put forward, fifty or a hundred words illustrating it are considered, and extension to the rest of the lexicon is left as an exercise for the reader. One motive for developing WordNet was to expose such hypotheses to the full range of the common vocabulary. WordNet presently contains approximately 54,000 different word forms . . . and only the most robust hypotheses have survived. [Miller *et al.*, 1990], p. 2.

Although I cannot judge how well WordNet fares with regard to its psycholinguistic aims, its noun taxonomy appears to have many of the qualities needed if it is to provide basic taxonomic knowledge for the purpose of corpus-based research in English.[5] As noted above, its coverage is remarkably broad (though proper names represent a notable exception). In addition, most easily-identifiable senses of words are distinguished; for example, *buck* has distinct senses corresponding to both the unit of currency and the physical object, among others, and *newspaper* is identified both as a form of paper and as a mass medium of communication.

Given the WordNet noun hierarchy, the definition of "coherent class" adopted here is straightforward. Let words(w) be the set of nouns associated with a WordNet class w.[6]

Definition. A noun class $c \in \mathcal{C}$ is *coherent* iff there is a WordNet class w such that words(w)$\cap$ $\mathcal{N} = c$.

As a consequence of this definition, noun classes that are "too small" or "too large" to be coherent are excluded. For example, suppose that *water*, *tea*, *beer*, *whiskey*, and *happiness* are among the elements of $\mathcal{N}$. The noun class $\{water, tea, beer\}$ turns out to be incoherent, because there is no WordNet class that includes its three elements but fails to include *whiskey*. Conversely, the noun class $\{water, tea, happiness\}$ is excluded because there is no WordNet class that includes all three elements: the WordNet noun hierarchy has no unique top element, so even the most general WordNet class that includes *water* and *tea* (the class ⟨thing, [entity, thing]⟩) fails to include *happiness*, which has ⟨state, [state]⟩, ⟨psychol_feature, [psychol_feature]⟩, and ⟨abstraction, [abstraction]⟩ as its most general superordinate categories.[7]

Experimental Results

An experiment was performed in order to discover the "prototypical" object classes for a set of common English verbs. The verbs considered are frequently-occurring verbs found in a corpus of parental speech.[8]

The counts of equation (2) were calculated by collecting a sample of verb/object pairs from the Brown Corpus, which contains approximately one million words of English text, balanced across diverse genres such as newspaper articles, periodicals, humor, fiction, and government documents.[9] The object of the verb was identified in a fashion similar to that of [Hindle, 1990], except that a set of heuristics was used to extract surface objects only from surface strings, rather than using a fragmentary parse to extract both deep and surface objects. In addition, verb inflections were mapped down to a base form and plural nouns were mapped down to their singular forms.[10] For example, the sentence *John ate two shiny red apples* would yield the pair $(eat, apple)$. The sentence *These are the apples that John ate* would not provide a pair for *eat*, since *apple* does not appear as its *surface* object.

Given each verb, v, the verb's "prototypical" object class was found by conducting a search upwards in the WordNet noun hierarchy, starting with WordNet classes containing synonym-set members that appeared as objects of the verb. Each WordNet class w considered was evaluated by calculating $A(v, \mathcal{N} \cap \text{words}(w))$. (For details see Figure 1.) Classes having too low a count (fewer than five occurrences with the verb) were excluded from consideration, as were a set of classes at or near the top of the hierarchy that described concepts such as *entity*, *psychological_feature*, *abstraction*, *state*, *event*, and so forth.[11]

The results of this experiment are encouraging.

[5] WordNet also contains taxonomic information about verbs and adjectives, and the taxonomy goes well beyond subordinate/superordinate relations. See [Beckwith *et al.*, 1991].

[6] Strictly speaking, WordNet as described by [Miller *et al.*, 1990] does not have classes, but rather lexical groupings called synonym sets. By "WordNet class" I mean a pair ⟨word, synonym-set⟩. For example, the WordNet class ⟨iron, [iron, branding_iron]⟩ is a subclass of ⟨implement, [implement]⟩.

[7] A related lexical representation scheme being investigated independently by Paul Kogut (personal communication) is to assign to each noun and verb a vector of feature/value pairs based upon the word's classification in the WordNet hierarchy, and to classify nouns on the basis of their feature-value correspondences.

[8] I am grateful to Annie Lederer for providing these data.

[9] The version of the Brown corpus used was the tagged corpus found as part of the Penn Treebank [Brill *et al.*, 1990].

[10] Nouns outside the scope of WordNet that were tagged as proper names were mapped to the token *pname*, a subclass of classes ⟨someone, [person]⟩ and

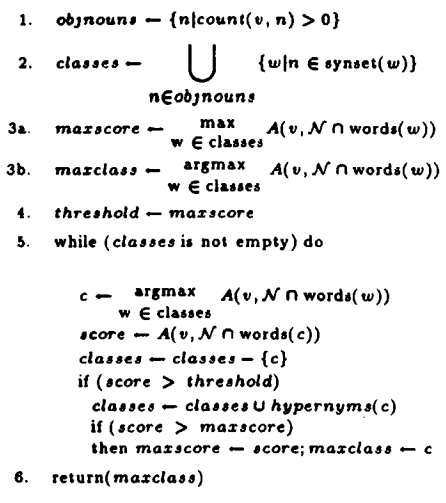

Figure 1: *Search algorithm for identifying a verb's "prototypical" object class*

A(v,c)	object class
3.58	⟨beverage, [drink,...]⟩
2.05	⟨alchoholic_beverage, [intoxicant,...]⟩

Table 4: *Object classes for* drink

Table 4 shows the object classes discovered for the verb ***drink***, which can be compared to the corresponding results using lexical statistics in Table 2. Table 5 shows the highest-scoring object classes for a selection of fifteen verbs chosen randomly from the sample (these are taken to be the "prototypical" object classes) and Table 6 identifies which subset of the objects of the verb fell into that "prototypical" class.

Tables 5 and 6 can be evaluated on the basis of how reasonably the class answers the question "What do you typically —?" for each verb. However, the search process also provides additional information about the kinds of object that appear with the verb. For example, although ⟨door, [door]⟩ is the top-scoring class for *open*, other classes with positive scores include

- ⟨entrance, [entrance]⟩: (door)
- ⟨mouth, [mouth]⟩: (mouth)
- ⟨repository, [depository, ...]⟩: (store, closet, locker, trunk)

⟨location, [location]⟩.

[11]The categories at the top of the hierarchy were excluded primarily for computational reasons. Since instances of such high-level classes appear extremely frequently, one would expect them not to have high mutual information with any given verb. The experiment has recently been replicated using parental speech from the CHILDES corpus [MacWhinney, 1991], without excluding any classes *a priori*, with comparable results.

A(v,c)	verb	object class
0.16	call	⟨someone, [person]⟩
0.30	catch	⟨looking_at, [look]⟩
2.39	climb	⟨stair, [step]⟩
1.15	close	⟨movable_barrier, [...]⟩
3.64	cook	⟨meal, [repast]⟩
0.27	draw	⟨cord, [cord]⟩
1.76	eat	⟨nutrient, [food]⟩
0.45	forget	⟨conception, [concept]⟩
0.81	ignore	⟨question, [problem]⟩
2.29	mix	⟨intoxicant, [alcohol]⟩
0.26	move	⟨article_of_commerce, [...]⟩
0.39	need	⟨helping, [aid]⟩
0.66	stop	⟨vehicle, [vehicle]⟩
0.34	take	⟨spatial_property, [...]⟩
0.45	work	⟨change_of_place, [...]⟩

Table 5: *Some prototypical object classes*

verb	object subset
call	(pname,man,...)
catch	(eye)
climb	(step,stair)
close	(door)
cook	(meal,supper,dinner)
draw	(line,thread,yarn)
eat	(egg,cereal,meal,mussel,celery,chicken,...)
forget	(possibility,name,word,rule,order)
ignore	(problem,question)
mix	(martini,liquor)
move	(comb,arm,driver,switch,bit,furniture,...)
need	(assistance,help,support,service,...)
stop	(engine,wheel,car,tractor,machine,bus)
take	(position,attitude,place,shape,form,...)
work	(way,shift)

Table 6: *Objects of the verb appearing in the "prototypical" class.*

- ⟨container, [container, ...]⟩: (bottle, bag, trunk, locker, can, box, hamper)
- ⟨time_period, [time_period, ...]⟩: (tour, round, season, spring, session, week, evening, morning, saturday)
- ⟨oral_communication, [speech, ...]⟩: (discourse, engagement, relation, reply, mouth, program, conference, session)
- ⟨writing, [writing]⟩: (scene, book, program, statement, bible, paragraph, chapter)

Thus, although this experiment suggests that one "prototypically" opens doors, the other object classes for *open* suggest a number of plausible distinctions in the way the verb is used. Some of the distinctions, such as that between the classes "repository" and "container," are not at all obvious by simple inspection of the verb's object nouns, even when ranked by mutual information.[12]

The careful reader may have noted a potential objection to this approach among the various classes listed above. Although *mouth* (saucy or disrespectful language) and *relation* (the act of telling or recounting) are both instances of oral communication, it is likely that when the pairs (*open*, *mouth*) and (*open*, *relation*) appeared in the corpus, the objects were not being used in that sense. Nonetheless, those instances counted as evidence for the purpose of calculating the association score for *open* and ⟨oral_communication, [speech, ...]⟩. In addition, this problem surfaces in an obvious way for classes containing just a single word; for example, the singleton set {*mouth*} was also associated with WordNet classes ⟨impertinence, [impudence]⟩ and ⟨riposte, [rejoinder]⟩.

Although inappropriate word senses add noise to the data, it is not clear that they seriously undermine efforts to extract useful associations. Previous efforts using word/word associations (e.g. [Church *et al.*, 1989; Hindle, 1990]) appear to have relied on the strength of large numbers to wash out inappropriate senses. Presumably (*play*, *record*) (as in *John played a record on his stereo*) and (*beat*, *record*) (as in *John beat the school record for absenteeism*) are both considered instances of *record*, and the word is effectively counted in all senses appropriate to the verbs with which it appears. The approach taken here is similar: so long as the appropriate class is *among* the classes to which a word belongs, the word's appearance in other classes should not make a significant difference. (An exception pointed out by Hindle [1990] is the case of consistent or systematic errors in parsing, or, in this case, in the heuristics used to identify the direct object.)

[12] According to the American Heritage Dictionary, a repository is "a place where things may be put for safekeeping," whereas a container is "something, as a box or barrel, in which material is held or carried." In WordNet neither class subsumes the other.

In summary, the experiment described in this section demonstrates the viability of using a hand-constructed conceptual hierarchy such as WordNet as a basis for computing class-based statistics. Exploring the relationships between words and classes in this fashion, rather than relationships between words, permits even low-frequency word-pairs to contribute to statistical regularities, and provides a a suitable level of abstraction for tasks such as this one that go beyond purely lexical associations.

Related Work

[Yarowsky, 1992] has independently investigated an approach similar to the one described here and applied it to the problem of word sense disambiguation. He demonstrates that word classes within a hand-constructed taxonomy — Roget's Thesaurus — can successfully be used as word sense discriminators by calculating statistics similar to those of equations (2) and (4).

One interesting difference between the two approaches has to do with use of levels within the taxonomy. Yarowsky restricts the possible word-sense labels to Roget's *numbered* classes — in effect, designating *a priori* which horizontal slice across the hierarchy will contain the relevant distinctions. In contrast, the search procedure defined here relies on the statistical association score to automatically find a proper level in the hierarchy: too low, and the next higher class may bring in additional relevant words; too high, and the class may have expanded too much. The advantages and disadvantages of *a priori* restrictions within the hierarchy warrant further consideration.

Recently, [Basili *et al.*, 1992] have independently pointed out the limitations discussed in Section 2, writing that "the major problem with word-pairs collections is that reliable results are obtained only for a small subset of high-frequency words on very large corpora" and that "an analysis based simply on surface [i.e. word-pair] distribution may produce data at a level of granularity too fine" (p. 97). They propose using high-level semantic tags in order to perform a statistical analysis similar to the one presented here, where the tags are assigned by hand from a very small, domain-specific set of categories. Their technique is applied to the problem of acquiring selectional restrictions across a variety of syntactic relations (subject/verb, verb/object, noun/adjective, etc.). As in the present study, only surface syntactic relationships are considered.

Were an equivalent of the WordNet taxonomy available for Italian, it would be interesting to apply the technique proposed here to the same set of data considered by [Basili *et al.*, 1992]. The most significant difference between the two approaches, as was the case for [Yarowsky, 1992], is in how nouns are grouped into classes. Although Basili *et al.* argue in favor of domain-dependent classification, they point out the obvious drawback of having to reclassify the corpus for new application domains.

An alternative approach might be to use a broad-coverage taxonomy such as WordNet rather than hand-defined, hand-assigned semantic tags. Such an approach would represent a compromise between the overly fine-grained distinctions found when considering individual words and the extremely high-level classification required if semantic tagging is to be done manually. Should the classification still prove too fine grained, or unsuitable for the application domain, one could (a) *exclude* from consideration all taxonomic classes other than those considered relevant to the task, and (b) *construct*, as disjunctions of taxonomic classes, a small set of domain-specific classes not found directly in the taxonomy.

[Grefenstette and Hearst., 1992] have recently presented an interesting combination of knowledge-based and statistical methods concerning noun classes. They explore the possibility of combining a corpus-based lexical discovery technique (using pattern-matching to identify hyponymy relations) with a statistically-based word similarity measure, in order to discover new instances of lexical relations for extending lexical hierarchies such as WordNet .

Further Applications

In this section, I briefly consider three areas in statistical natural language processing for which a class-based approach of the kind proposed here might prove useful. Of these, only the last represents work in progress; the first two are as yet unexplored.

Computational Lexicography and Lexicology

For the purposes of lexicography, the work presented here can be considered a generalization of [Church and Hanks, 1989], which describes the use of a lexical association score in the identification of semantic classes and the analysis of concordances. The discussion of the object classes discovered for *open* in Section 4 suggests that corpus-based statistics and broad-coverage lexical taxonomies represent a powerful combination, despite the fact that neither the corpus data nor the taxonomy are free of error. To the extent that WordNet respects the lexicographer's intuitions about taxonomy, the classes uncovered as statistically relevant will provide important information about usage that might otherwise be overlooked.

In addition, the experience of applying WordNet to large corpora may be useful to researchers in psycholexicology. WordNet was designed to "expose [psycholinguistic] hypotheses to the full range of the common vocabulary," and it is likely that class-based statistics of the kind described here, calculated using large quantities of naturally occurring text, can contribute to that enterprise by uncovering places in the hierarchy where the linguist's view of lexical taxonomy diverges from common usage.

Statistical Language Modelling

A central problem in statistical language modelling — as used in speech recognition, for example — is accurate prediction of the next word on the basis of its prior context. Abstractly, the goal is to closely approximate $\Pr(X_n|X_1^{n-1})$, where X_1^{n-1} is an abbreviation for $X_1, \ldots, X_{n-1}$ and each X_i ranges over words in the vocabulary. In practice, these estimates are frequently calculated by computing

$$\Pr(X_n|X_1^{n-1}) \approx \Pr(X_n|f(X_1^{n-1})),$$

where f is a function that partitions the possible prior contexts into equivalence classes. For the common trigram language model, for example, $f(x_1^{n-1}) = x_{n-2}x_{n-1}$, so that

$$\Pr(X_n|X_1^{n-1}) \approx \Pr(X_n|X_{n-2}X_{n-1})). \quad (5)$$

In addition, probability estimates often make use of information from several different equivalence classes either by "backing off" or by smoothing; for example, the smoothing technique of interpolated estimation [Bahl *et al.*, 1983; Jelinek and Mercer., 1980] estimates

$$\Pr(X_n|X_1^{n-1}) \approx \sum_i \lambda_i \Pr(X_n|f_i(X_1^{n-1})),$$

where each f_i determines a different equivalence class, and the λ_i's weight the equivalence classes' contributions.

Several recent approaches to language modelling have extended the notion of equivalence class to include more than just the classes formed by "forgetting" all but the previous few words (e.g. equation (5)). [Bahl *et al.*, 1989] proposes using a binary decision tree to classify prior contexts: since each node of a decision tree consitutes an equivalence class, interpolated estimation can be used to combine each value of f_i computed along the path from any leaf node to the root of the tree. [Brown *et al.*, 1990] suggests a hierarchical partitioning of the vocabulary into equivalence classes on the basis of similar bigram distributions: trigram models can then be estimated by defining $f(x_1^{n-1}) = c_{n-2}c_{n-1}$, where each c_j is the equivalence class of x_j.

These approaches are strictly "knowledge-free," in the sense that they eschew hand-constructed information about lexical classes. However, the results of Section 4 suggest that, by using a broad-coverage lexical database like WordNet , statistical techniques can take advantage of a conceptual space structured by a hand-constructed taxonomy. A disadvantage of the purely statistical classification methods proposed in [Bahl *et al.*, 1989] and [Brown *et al.*, 1990] is their difficulty in classifying a word not encountered in the training data; in addition, such methods have no remedy if a word's distribution in the training data differs substantially from its distribution in the test data. These difficulties might be alleviated by an integrated approach that takes advantage of taxonomic *as well as* purely distributional information, treating concepts in the taxonomy as just another

kind of equivalence class. The method of interpolated estimation provides one straightforward way to effect the combination: the contribution of the taxonomic component will be large or small, compared to other kinds of equivalence class, according to how much its predictions contribute to the accuracy of the language model as a whole.

Linguistic Investigation

Diathesis alternations are variations in the way that a verb syntactically expresses its arguments. For example, (1) shows an instance of the *indefinite object alternation*,

(1)a. John ate the meal.
b. John ate.
(2) *The meal ate.

and example (3) shows an instance of the *causative/inchoative alternation* [Levin, 1989].

(3)a. John opened the door.
b. The door opened.
(4) *John opened.

Such phenomena are of particular interest in the study of how children learn the semantic and syntactic properties of verbs, because they stand at the border of syntax and lexical semantics. There are numerous possible explanations for why verbs can and cannot appear in particular alternations, ranging from shared semantic properties of verbs within a class, to pragmatic factors, to "lexical idiosyncracy." Statistical techniques like the one described in this paper may be useful in investigating relationships between verbs and their arguments, with the goal of contributing data to the study of diathesis alternations, and, ideally, in constructing a computational model of verb acquisition.

One interesting outcome of the experiment described in Section 4 is that the verbs participating in "implicit nominal object" diathesis alternations[13] associate more strongly with their "prototypical" object classes than do verbs for which implicit objects are disallowed. Preliminary results, in fact, show a statistically significant difference between the two groups. This suggests the following interesting question: might shared *information-theoretic* properties of verbs play a role in their acquisition, in the same way that shared semantic properties do?

Conclusions

In this paper, I have suggested a method for investigating statistical relationships between words and word classes. The approach comprises two parts: a (straightforward) application of word-based mutual information to sets of words, and a proposal for organizing those sets according to a knowledge-based conceptual taxonomy. The experiment described here demonstrates the viability of the technique, and shows how the class-based approach is an improvement on statistics that use simple lexical association.

[13]The indefinite object alternation [Levin, 1989] and the specified object alternation [Cote, 1992] — roughly, verbs for which the direct object is understood when omitted.

References

Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. A maximum likelihood approach to continuous speech recognition. In IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 179–190. PAMI-5(2), March 1983.

Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer. A tree-based statistical language model for natural language speech recognition. In IEEE Transactions on Acoustics, Speech and Signal Processing, pages 37:1001–1008, July 1989.

Roberto Basili, Teresa Pazienza, and Paola Velardi. Computational lexicons: the neat examples and the odd exemplars. In Third Conference on Applied Natural Language Processing, pages 96–103. Association for Computational Linguistics, March 1992.

Richard Beckwith, Christiane Fellbaum, Derek Gross, and George Miller. WordNet: A lexical database organized on psycholinguistic principles. In Uri Zernik, editor, Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, pages 211–232. Erlbaum, 1991.

Eric Brill, D. Magerman, M. Marcus, and B. Santorini. Deducing linguistic structure from the statistics of large corpora. In DARPA Speech and Natural Language Workshop, 1990.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, and Robert L. Class-based n-gram models of natural language. In Proceedings of the IBM Natural Language ITL, pages 283–298, Paris, France, March 1990.

Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In Proceedings of the 27th Annual Meeting of the ACL, 1989.

K. Church, W. Gale, P. Hanks, and D. Hindle. Parsing, word associations and typical predicate-argument relations. In Proceedings of the International Workshop on Parsing Technologies, August 1989.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. Using statistics in lexical analysis. In Uri Zernik, editor, Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, pages 116–164. Erlbaum, 1991.

Sharon Cote. Discourse functions of two types of null objects in English, January 1992. Presented at the 66th Annual Meeting of the Linguistic Society of America, Philadelphia, PA.

Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. John Wiley, 1991.

Jeffrey Elman. Finding structure in time. Cognitive Science, 14:179–211, 1990.

W. Francis and H. Kucera. Frequency Analysis of English Usage. Houghton Mifflin Co., New York, 1982.

William Gale, Kenneth Church, and David Yarowsky. A method for disambiguating word senses in a large corpus. Technical report, AT&T Bell Laboratories, 1992.

Gregory Grefenstette and Marti Hearst. A method for refining automatically-discovered lexical relations:

Combining weak techniques for stronger results. In AAAI Workshop on Statistically-based NLP Techniques, San Jose, July 1992.

Donald Hindle. Noun classification from predicate-argument structures. In Proceedings of the 28th Annual Meeting of the ACL, 1990.

Frederick Jelinek and Robert L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In Proceedings of the Workshop on Pattern Recognition in Practice, Amsterdam, The Netherlands, May 1980. North-Holland.

Beth Levin. Towards a lexical organization of English verbs. Technical report, Dept. of Linguistics, Northwestern University, November 1989.

Brian MacWhinney. The CHILDES project : tools for analyzing talk. Erlbaum, 1991.

George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Five papers on WordNet. CSL Report 43, Cognitive Science Laboratory, Princeton University, July 1990.

George Miller. Wordnet: An on-line lexical database. International Journal of Lexicography, 3(4), 1990. Special Issue.

Ronald Rosenfeld and Xuedong Huang. Improvements in stochastic language modelling. In Mitch Marcus, editor, Fifth DARPA Workshop on Speech and Natural Language, Arden House Conference Center, Harriman NY, February 1992.

David Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In Proceedings of COLING-92, pages 454–460, Nantes, France, July 1992.