Investigating Complementary Methods for Verb Sense Pruning

Hongyan Jing and Vasileios Hatzivassiloglou
and Rebecca Passonneau and Kathleen McKeown
Department of Computer Science
450 Computer Science Building
Columbia University
New York, N.Y. 10027
{hjing, vh, becky, kathy}@cs.columbia.edu

Abstract

We present an approach for tagging verb sense that combines a domain-independent method based on subcategorization and alternations with a domain-dependent method utilizing statistically extracted verb clusters. Initial results indicate that verb senses can be pruned for highly polysemous verbs by up to 74% by the first method and by up to 85% by the second method.

1 Introduction

Much work in natural language processing is predicated on the notion that linguistic usage varies sufficiently across different situations of language use that systems can be tailored to a particular sublanguage variety (Kittredge and Lehrberger, 1982). Biber (1993) presents evidence that a corpus restricted to one or two language registers would exclude "much of the English language" by narrowing the lexicon, verb tense and aspect, and syntactic complexity. Such observations inform the increasing trend towards analysis of homogeneous corpora to identify linguistic constraints for use in systems intended to understand or generate coherent discourse. Recent work in this vein includes identification of lexical constraints from textual tutorial dialogue (Moser and Moore, 1995), constraints on illocutionary act type from spoken task-oriented dialogue (Allen et al., 1995), prosodic constraints from spoken information-seeking monologues (Hirschberg and Nakatani, 1996), and constraints on referring expressions from spoken narrative monologue (Passonneau, 1996). Related work suggests that constraints of different types are interdependent (Biber, 1993; Passonneau and Litman, forthcoming), hence should be investigated together. Our ultimate goal is to develop methods to tag lexical semantic features in discourse corpora in order to enhance extraction of constraints of the sort just listed. Two types of investigations that would undoubtedly be enhanced are explorations of the interrelation of lexical cohesion and global discourse structure (Morris and Hirst, 1991; Hearst, 1994), and identification of lexicalization patterns for domain-specific concepts (Robin, 1994).

In this paper, we propose a two-pronged approach to an initial step in lexical semantic tagging, pruning the search space for polysemous verbs. Rather than attempting to identify unique word senses, we aim for the more realistic goal of pruning sense information. We will then incrementally evaluate the utility of tagging corpora with pruned sense sets for different types of discourse. We begin with verbs on the hypothesis that verb sense distinctions correlate with syntactic properties of verbs (Levin, 1993). Our initial results indicate that domain-independent syntactic information reduces potential verb senses for multiply polysemous verbs (five or more WordNet senses) by more than 50%. In Section 2, we outline our first method, based on domain-independent lexical knowledge, presenting results from an analysis of thousands of verbs. In the section following that, we present our complementary method, a technique utilizing verb clusters automatically computed from corpus data. In the conclusion, we discuss how the combination of the two methods increases the performance of our system and enhances the robustness of the final results.

2 Exploiting domain-independent syntactic clues

A given word may have n distinct senses and appear within m different syntactic contexts, but typically, not all $n \times m$ combinations are valid. The syntactic context can partly disambiguate the semantic content. For example, when the verb *question* has a that-clause complement, it cannot have the sense of "ask", but rather must have the sense of "challenge".

To identify such interacting syntactic and semantic constraints at the lexical level, we utilize three knowledge bases for verbs:

- The COMLEX database (Grishman et al., 1994; Macleod and Grishman, 1995), which includes detailed subcategorization information for each verb, and some adjectives and nouns.

- Levin's classification of verbs in terms of their allowed *alternations* (Levin, 1993). Alternations include syntactic transformations such as *there*-insertion (e.g., *A ship appeared on the horizon* → *There appeared a ship on the horizon*) and locative-inversion (e.g., → *On the horizon there appeared a ship*). Much in the same way as subcategorization frames, alternations are constrained by the sense of the word; for example, the verb *appear* allows *there*-insertion and locative-inversion in its senses of "come into being" or "become visible", but not in its senses of "come out" or "participate in a play".
- WordNet's (Miller et al., 1990) hierarchical semantic classification. WordNet supplies links between semantically related senses as encoded in synonym sets (*synsets*). Though many words are polysemous, Miller *et al.* (1990) argue that a set of synonymous or nearly synonymous words can serve to identify the single lexical concept they have in common. It also supplies limited subcategorization information, in the form of allowed sentential frames ("verb frames") for each sense.

WordNet contains the needed information on permissible combinations of syntactic context and semantic content, but its subcategorization information is limited. Thirty-five different subcategorization frames are used for all verbs in WordNet, and the frames supplied are partial. COMLEX provides more detailed specifications of the syntactic frames for each verb (92 distinct subcategorization types). The allowed alternations (which we encoded in machine-readable form from the detailed rules supplied in (Levin, 1993)) provide additional constraints. Mapping the more precise syntactic information in COMLEX to the verb frames of WordNet allows the construction of a more detailed syntactic entry for each word sense, and enables the association of alternation constraints with the senses in WordNet. In the future, it will also allow us to use corpora tagged with COMLEX subcategorization frames, e.g., (Macleod et al., 1996).

We have manually constructed a table that maps WordNet syntactic constraints to the ones used in COMLEX (and vice versa) and another that maps allowed alternations from (Levin, 1993) to COMLEX or WordNet syntactic frames. A program consults the three databases and the mapping tables and, for each word occurrence constructs a list of the senses that are compatible with the syntactic constraints. During this process, a detailed entry for the word is formed, containing both syntactic and semantic information. The resulting entries comprise a rich lexical resource that we plan to use for text generation and other applications (Jing et al., 1997).

For a specific example, consider the verb *appear*. The pertinent information in the three databases for this word is listed in parts (a)–(c) of Figure 1. For

```
(VERB  :ORTH "appear"
       :SUBC ((PP-TO-INF-RS :PVAL ("to"))
              (PP-PRED-RS
                 :PVAL ("to" "of"
                        "under" "against"
                        "in favor of"
                        "before" "at"))
              (EXTRAP-TO-NP-S)
              (INTRANS)
              (SEEM-S)
              (SEEM-TO-NP-S)
              (TO-INF-RS)
              (NP-PRED-RS)
              (ADJP-PRED-RS)
              (ADVP-PRED-RS)
              (AS-NP)))
```

(a) COMLEX entry for *appear*

```
INTRANS    THERE-V-SUBJ
           :ALT there-insertion
LOCPP      LOCPP-V-SUBJ
           :ALT locative-inversion
```

(b) Allowed alternations for *appear*

```
appear  Sense 1  (give an impression)
         *> Something ----s Adjective/Noun
         *> Somebody ----s Adjective
         *> Somebody ----s to INFINITIVE
        Sense 2  (become visible)
         *> Something ----s
         *> Somebody ----s
         *> Something is ----ing PP
         *> Somebody ----s PP
        ...
        Sense 8  (have an outward
                  expression)
         *> Something ----s Adjective/Noun
         *> Somebody ----s Adjective
```

(c) WordNet sense-syntax constraints for *appear*

Figure 1: Database information for the verb *appear*.

example, one of the subcategorization frames of *appear* in part (a), ADJP-PRED-RS, indicates a predicate adjective with subject raising, as in *He appeared confused*. Part (b) of Figure 1 lists no alternations that are applicable to this subcategorization frame, while part (c) shows only two WordNet synsets where *appear* takes an adjectival complement, senses S1 and S8. The complex entry of Figure 2 is produced automatically from these three types of lexical information. The resulting syntax-semantics restriction matrix for *appear* is shown in Table 1. When *appear* is encountered in a particular syntactic structure, the program consults the

```
(appear
 ((1 ((PP-TO-INF-RS :PVAL ("to")
                    :SO ((sb, -)))
      (TO-INF-RS :SO ((sb, -)))
      (NP-PRED-RS :SO ((sb, -) (sth, -)))
      (ADJP-PRED-RS :SO ((sb, -)
                         (sth, -)))))
      (ADVP-PRED-RS :SO ((sb, -)
                         (sth, -)))))
  (2 ((PP-TO-INF-RS :PVAL ("to")
                    :SO ((sb, -)
                         (sth, -)))
      (PP-PRED-RS :PVAL ("to" "of"
                         "under" "against"
                         "in favor of"
                         "before" "at")
                  :SO ((sb, -) (sth, -)))
      (INTRANS :SO ((sb, -) (sth, -)))
      (AS-NP :SO ((sb, -) (sth, -)))
      (LOCPP :SO ((sb, -) (sth, -)))
      (INTRANS THERE-V-SUBJ
          :ALT there-insertion
          :SO ((sb, -) (sth, -)))
      (LOCPP LOCPP-V-SUBJ
          :ALT locative-inversion
          :SO ((sb, -) (sth, -)))))
  ...
  (8 ((NP-PRED-RS :SO ((sth, -)))
      (ADJP-PRED-RS :SO ((sb, -)
                         (sth, -)))
      (ADVP-PRED-RS :SO ((sb, -)
                         (sth, -)))))))
```

Figure 2: Automatically synthesized lexicon entry for the verb *appear*.

restriction matrix to eliminate senses that can be excluded. In the case of *appear*, only 47 cells of the 8×23 matrix represent possible combinations of syntactic patterns with senses, corresponding to a 74.5% reduction in ambiguity.

Due to incompatibilities between the COMLEX and WordNet representations of syntactic information, and the differences in coverage, the process of linking the information sources can in some cases result in relatively underspecified rows of a restriction matrix, or to spurious cells. For example, the frame `ADVP-PRED-RS` in Table 1 occurs in COMLEX but does not correspond to any of the more general frames mentioned in WordNet. Rather than having no appropriate senses for this syntactic pattern, we map it to WordNet's verb frames "Something ——s Adjective/Noun" and "Somebody ——s Adjective" by analyzing experiment results regressively.

On the other hand, the entry for S2 in the `PP-TO-INF-RS` frame for *appear* represents a spurious entry: *appear* does not occur in the S2 meaning of "become visible" with a *to*-prepositional phrase and a subject-controlled infinitive. In a sentence with this syntactic structure, such as "The river appeared to the residents to be rising too rapidly", *appear* can take only senses S1 and S6 for animate subjects and senses S3 and S7 for inanimate subjects. Yet the cell for S2 $\times$ `PP-TO-INF-RS` is generated in our matrix because of the overly general specification of verb frames in WordNet. We have chosen to risk overgeneration in these cases at present, rather than accidentally eliminating a valid sense. Eliminating spurious cells by hand would be time-consuming and error-prone, but the automatic classification method we report in the next section may help prune them. Also, as reported elsewhere (Jing et al., 1997), we are extending our lexical resource with annotations of frequency information for each sense-subcategorization pair, derived from sense-tagged corpus data. As data is accumulated, zero frequency could be taken to represent less valid usages.

We have performed preliminary evaluation tests of our method for tagging verb occurrences with pruned word sense tags using the Brown corpus. The first step of the method is to identify the subcategorization pattern for a specific verb token. Here we rely on heuristics to identify the major constituents to the left and right of a verb token, as described in (Jing et al., 1997). After hypothesizing the subcategorization pattern for a specific verb token, we use our sense restriction matrices (as in Table 1) to tag the verb token with a pruned set of senses. We evaluate the resulting sense tag against the version of the Brown corpus that has been hand-tagged with WordNet senses (Miller et al., 1993). For *appear*, which we use as an example throughout this paper, we find 100 tokens in the Brown corpus. Of these, 46 are intransitive or have a locative prepositional phrase complement. Our method tags each of these tokens with two or three possible senses, and in all but one case, the sense tag includes the valid sense. Another 31 tokens are followed by *to* and a subject-controlled infinitive. In all these cases, our method makes a single, correct prediction out of the eight possible senses. For all 100 uses of *appear* in the corpus, the average number of possible senses predicted by our method is 1.99. We find a 75–76% reduction of possible senses (depending on whether we use the additional *something/somebody* selectional constraints), with only 2–3% of the tags being incorrect.

For the 5,676 verbs present in all three databases, the average reduction in ambiguity was 36.82% for words with two to four senses, 59.36% for words with five to ten senses, and 73.86% for words with more than ten senses; the overall average for all polysemous words was 47.91%. Figure 3 is a bar chart showing, for each number of senses from 1 to 41, how many verbs with that number of senses occur

Subcategorization/Alternation		Sense							
		S1	S2	S3	S4	S5	S6	S7	S8
PP-TO-INF-RS	(sb, −)	+	+				+		
	(sth, −)		+	+				+	
PP-PRED-RS	(sb, −)		+				+		
	(sth, −)		+	+				+	
EXTRAP-TO-NP-S					+				
INTRANS	(sb, −)		+			+			
	(sth, −)		+	+		+			
SEEM-S					+				
SEEM-TO-NP-S					+				
TO-INF-RS	(sb, −)	+							
	(sth, −)								
NP-PRED-RS	(sb, −)	+							
	(sth, −)	+							+
ADJP-PRED-RS	(sb, −)	+							+
	(sth, −)	+							+
ADVP-PRED-RS	(sb, −)	+							+
	(sth, −)	+							+
AS-NP	(sb, −)		+				+		
	(sth, −)		+	+				+	
LOCPP	(sb, −)		+			+			
	(sth, −)		+	+		+			
THERE-INSERTION			+	+		+			
LOCATIVE-INVERSION			+	+		+			

Table 1: Valid combinations of syntactic subcategorization frames/alternations and senses (marked with +) for the verb *appear*.

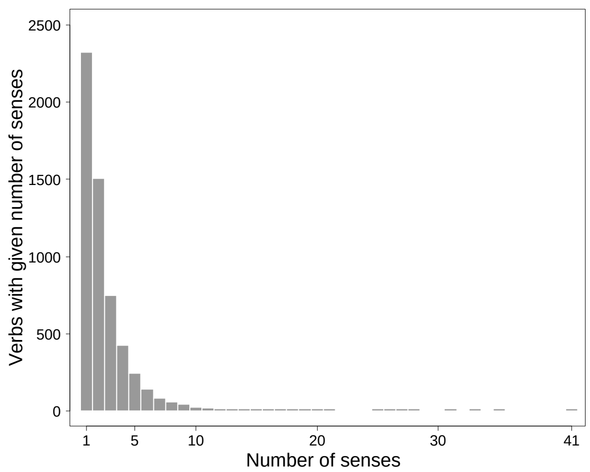

Figure 3: Distribution of verbs according to number of senses. Low frequencies are not drawn to scale; rather, the presence of a bar for a category corresponding to more than 10 senses indicates that at least one verb falls in that category.

in our databases. The most polysemous verb in our databases, *run*, is identified as having 41 senses.

About half the verbs have more than one sense, and 20% have more than two. Our method performs better on the more polysemous words, which are the most difficult to prune. This increased difficulty applies even to statistical methods because of the large number of alternatives and the likely closeness in meaning among them. Selecting a subset of almost synonymous verb senses is significantly harder than, for example, disambiguating *bank* between the "edge of river" and "financial institution" senses.

3 Using domain-dependent semantic classifications to identify predominant senses

The process outlined above has two significant advantages: first, it can be automatically applied, assuming a robust method for parsing the relevant verb phrase context (the experiments presented in (Pustejovsky et al., 1993) depend on the same type of information). Second, it reduces the ambiguity of a given word without sacrificing accuracy, insofar as the three input knowledge sources are accurate. To further restrict the size of the set of valid senses produced, we are currently exploring domain-dependent, automatically constructed semantic classifications.

Semantic classification programs (Brown et al., 1992; Hatzivassiloglou and McKeown, 1993; Pereira et al., 1993) use statistical information based on cooccurrence with appropriate marker words to partition a set of words into semantic groups or classes.

For example, using head nouns that occur with premodifying adjectives as one type of marker word, the adjective set {*blue*, *cold*, *green*, *hot*, *red*} can be partitioned into the subsets (lexical fields (Lehrer, 1974)) {*blue*, *green*, *red*} and {*cold*, *hot*}. Automatic classification programs can achieve high performance, near that of humans on the same task, when supplied with enough data and with appropriate syntactic constraints (see (Hatzivassiloglou, 1996) for a detailed evaluation). However, given that each word must be assigned to one class independently of context,[1] the problem of ambiguity is "solved" by placing each word in the class where it fits best; that is, in the class dictated by the predominant sense of the word in the training text.

While this might be a limitation of partitioning methods for lexicographical purposes, it offers an advantage for our task. By an indirect route, it allows the automatic identification of the predominant sense of a word in a given text or subject topic. It is indirect because the actual result is groups of word forms, but we presume each group to represent a relatively homogeneous semantic class. Thus we presume that the relevant sense of a given word form in a group is in the same lexical field as the senses of the other word forms in the same group. The process is highly domain-dependent, i.e., the same set of words will be partitioned in different ways when the domain changes. For example, when our word grouping system (Hatzivassiloglou and McKeown, 1993) classified about 280 frequent adjectives in stock market reports, it formed, among others, the cluster {*common*, *preferred*}. This cluster would look odd were not the domain considered.[2]

This information on predominant senses for each word form in a given corpus can be computed automatically, but remains implicit. To map the results onto word *sense* associations, and thus explicitly identify the predominant senses, we utilize the links between *senses* provided by WordNet. We note that while words like *question* and *ask* are ultimately connected in WordNet, the actual connections are only between some of the senses of the two words. Similarly, the words *question* and *dispute* are also connected, but through a different subset of senses. Thus, if the automatically induced semantic classification indicates that the predominant sense of *question* is associated with *dispute* rather than with *ask* (by placing *question* and *dispute* but not *ask* in the same group), we can infer which of the WordNet senses of *question* is the predominant one in this domain. The algorithm involves the following steps:

[1]Some systems produce "soft" clusters, where words can belong into more than one group. These can be converted to non-overlapping groups for the purposes of this discussion by assigning each word to the group for which it has the highest membership coefficient.

[2]In this domain, the two adjectives are complementaries, describing the two types of issued stock shares.

- Construct the domain-dependent word classification.
- For each word x, let $Y = \{Y_1, Y_2, \ldots\}$ be the set of other words placed in the same semantic group with x.
- For each $Y_i \in Y$, traverse the WordNet hierarchy and locate the (set of) senses of x, S_i, that are connected with *some* sense of Y_i. The distance and the types of links that can be traversed while still considering two senses "related" can be heuristically determined; alternatively, we can use a measure of semantic distance such as those proposed in (Resnik, 1995) or (Passonneau et al., 1996).
- Finally, the union of the sets S_i contains the predominant sense of x. While in the general case it is possible to have multiple links between word forms (corresponding to different sense pairings), typically each S_i will contain only one sense, and their union will contain a few elements. This set can be further reduced, e.g., by giving more weight to senses supported by more than one of the Y_i's or by unambiguous Y_i's.

For a concrete example, consider the verb *question*, which can have, among others, the senses of *dispute* (sense 1 in WordNet) or *inquire* (sense 3 in WordNet). If we consider a sense as linked with one of the senses of *question* if it is in the maximal subtree which includes that sense but no other senses of *question*, we find the following links between *question* and the verbs *ask*, *inquire*, *challenge*, and *dispute*: ($question_1$, ask_2), ($question_2$, ask_5), ($question_3$, ask_5), ($question_3$, $inquire_2$), and ($question_1$, $challenge_2$). Thus, if *question* is placed in the same semantic group with *ask* and *inquire*, the three senses {1, 2, 3} survive out of the five senses of *question*, with a preference for sense 3. If, on the other hand, *question* is classified with *challenge* and *dispute*, only sense 1 survives.

We performed an experiment analyzing a specific verb group produced by one semantic clustering program (McMahon and Smith, 1996). This group contains 19 verbs, all but one of them ambiguous, including *ask*, *call*, *charge*, *regard*, *say*, and *wish*. We measured for each sense of the 19 words how many of the other words have at least one sense linked with that sense in WordNet (in the same top-level verb sense tree). The results, part of which is shown in Table 2, indicate that some senses are much more strongly connected with the other words in the group, and so probably predominate in the corpus that was used to induce the group. For example, one of the senses of *ask*, "require" (as in *This job asks (for) long hours*) is not linked to any of the other 18 words in the cluster, and should therefore be removed. If, for each word w we analyze, we require that each of its probable senses be linked to at least a fixed percentage (e.g., one-third) of the total number of words linked to w, we can eliminate

Word	Number of senses	Number of other words in group linked with given sense												
		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
ask	5	9	9	9	0	9								
call	13	0	0	9	1	9	9	1	9	2	0	9	9	3
charge	9	3	2	0	0	2	3	0	2	3				
describe	4	1	9	3	0									
know	8	0	0	0	0	9	5	1	1					
regard	3	5	0	0										
say	9	2	0	1	0	9	0	9	1	9				
wish	7	2	2	1	2	2	9	9						

Table 2: Number of words in a semantic group linked with each sense of each word in it, and associated reduction in ambiguity. Eight of the 19 words are shown.

Verb	Number of senses in WordNet	Surviving senses after cluster-based method is applied	Reduction in ambiguity (types)	Occurrences in the corpus (tokens)	Wrongly tagged tokens	Error rate
show	13	9	30.77%	109	4	3.67%
describe	4	2	50.00%	32	1	3.12%
present	12	6	50.00%	8	1	12.50%
prove	9	4	55.56%	10	5	50.00%
introduce	10	4	60.00%	6	3	50.00%
Average	9.6	5	49.27%	33	2.8	8.48%

Table 3: Reduction in ambiguity and sense tagging error rate for the cluster-based method, as measured for five verbs on the J part of the Brown corpus.

many of the senses as improbable. The achieved reduction in ambiguity (for the 18 ambiguous words) ranges from 20% to 84.62% (including cases of full disambiguation), and its average for all 18 words is 55.89%.

In another experiment, we looked at a specific corpus, taking into account the frequency distribution of the verbs in it. We selected the J part of the Brown corpus, which focuses on learned knowledge (the Natural Sciences, Mathematics, Medicine, the Humanities, etc.) (Kučera and Francis, 1967). This part of the corpus is more homogeneous and contains a larger number of articles (80). The increased homogeneity makes it suitable for investigating our hypothesis of predominant verb senses.

We selected five verbs from this sub-corpus (*show*, *describe*, *present*, *prove*, and *introduce*), and applied our algorithm assuming that the predominant senses of these verbs are linked together and consequently, that the five verbs would be placed in the same group by the clustering program.

Under this assumption, we measured the reduction in ambiguity (number of possible senses) for each verb (*types*) as well as over all occurrences of the five verbs in the sub-corpus (*tokens*) when the cluster-based algorithm is applied. We also counted how many of the verbs receive a wrong tag, i.e., a set of senses that does not include the hand-assigned one. The results of these experiments are shown in Table 3. We observe that the cluster-based method achieves a 49.27% reduction in the number of senses when measured on types. When the distribution of the words is factored in, the corresponding measure on tokens (which better describes the applicability of the method in practice) is 38.00%. The average error rate is 8.48%; this average is driven up by the inclusion of *present*, *prove*, and *introduce* in our test set. The relatively high error rate for these verbs may be due to their low frequency in our corpus, or may indicate that their predominant senses are not associated with the predominant senses of *show* and *describe* as we hypothesized.

4 Combining the two methods

While the syntactic constraints method almost always produces a semantic tag that includes the correct sense for a verb,[3] it has no capability to further distinguish the surviving senses in the tag. The semantic link-based method, on the other hand, can eliminate some senses from this tag. By applying the two methods in tandem and intersecting the sense sets produced by them, we can reduce the size of the final tag. Using the verb "show" of the experiment described in the previous section as an illustration, we note that whenever the verb takes only a direct object, the syntactic method eliminates three of the thirteen possible senses while always retaining the

[3] Assuming no gaps in the subcategorization information for this verb in COMLEX and WordNet.

correct sense in the produced tag (error rate 0%). For the same verb and subcategorization pattern, the cluster-based method rejects four of the thirteen senses with error rate 5% (i.e, 3 out of 58 occurrences in the J part of the Brown Corpus will be assigned wrong tags). The intersection of the two methods increases the number of rejected senses to five. It reduces the ambiguity by 38% but has the combined error rate of both methods, in this case 5%.

As we see from this experiment, the integration of the two methods can improve the reduction rate of ambiguity, but may slightly increase the error rate. We are investigating ways to stratify the application of the cluster-based method on appropriate groups of tokens identified by the syntactic method, by separately clustering tokens of the same verb that appear in different syntactic frames. We expect that this will partly alleviate the increase in the error rate.

5 Discussion

Our method for using detailed knowledge about verb subcategorizations and alternations to prune verb senses is domain independent. It also prunes senses without loss of correctness. By intersecting the resulting sense sets with the output of our cluster-based method, verb senses can be pruned further. In using the clustering method's output, we make two further assumptions. Previous work has shown that within a given discourse (Gale et al., 1992), or with respect to a given collocation (Yarowsky, 1993), a word appears in only one sense. By extrapolation, we will assume that words appear in only one sense within a homogeneous corpus,[4] except for certain high frequency verbs or for semantically empty support verbs. We will assign this predominant sense to all non-disambiguated occurrences of a verb. While this provides a reasonable default, the resulting semantic tag has to be considered provisional, and validated independently. Also, we currently assume that words placed in the same group will share relatively few links (connecting pairs of competing senses) in WordNet. This is supported by our initial experiments, but is an issue we will continue to investigate.

Above we gave some preliminary evaluation results; we plan to carry out a more complete evaluation of our system by continuing to use the hand-tagged (with WordNet senses) Brown corpus (Miller et al., 1993) as the initial evaluation standard. Each stage will be separately measured, as well as their combined effectiveness in pruning senses. We anticipate that the use of multiple methods to investigate sense pruning will lead to more robust results. In addition, we believe that the two methods can be interleaved in the following manner: Both methods rely on recognizing features of the local syntactic context of a verb occurrence; the look-up method uses the local syntactic context to identify the likely subcategorization pattern while the automatic classification method uses the local syntactic context to extract marker words. The look-up method can tag distinct tokens of the same verb with distinct senses if the subcategorization patterns are distinct and correlate with distinct senses. The automatic classification method could be extended to classify sense sets, using as its input corpus the output of the syntactic constraints look-up method, where verb tokens have been tagged with a subset of the full collection of senses. In principle, this would make it possible to use the automatic classification method on a more heterogeneous corpus, i.e., where the same verb occurs frequently with two distinct senses.

[4]Or a few predominant senses, that can perhaps be disambiguated using syntactic constraints as we discuss below.

References

James F. Allen, Lenhart K. Schubert, George Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel G. Martin, Bradford W. Miller, Massimo Poesion, and David R. Traum. 1995. The TRAINS project: A case study in defining a conversational planning agent. *Journal of Experimental and Theoretical AI*, **7**:7–48.

Douglas Biber. 1993. Using register-diversified corpora for general language studies. *Computational Linguistics*, **19**(2):219–242, June. Special Issue on Using Large Corpora: II.

Peter F. Brown, Vincent J. della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, **18**(4):467–479.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, February.

Ralph Grishman, Catherine Macleod, and Adam Meyers. 1994. COMLEX syntax: Building a computational lexicon. In *Proceedings of COLING-94*, Kyoto, Japan, August.

Vasileios Hatzivassiloglou and Kathleen McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 172–182, Columbus, Ohio, June.

Vasileios Hatzivassiloglou. 1996. Do we need linguistics when we have statistics? A comparative analysis of the contributions of linguistic cues to a statistical word grouping system. In Judith L. Klavans and Philip S. Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 67–94. The MIT Press, Cambridge, Massachusetts.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, New Mexico.

Julia Hirschberg and Christine H. Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 286–293, Santa Cruz, California, June.

Hongyan Jing, Kathleen McKeown, and Rebecca Passonneau. 1997. Building a rich large-scale lexical base for generation. Submitted to the 35th Annual Meeting of the Association for Computational Linguistics.

R. Kittredge and J. Lehrberger, editors. 1982. *Sublanguage: Studies of Language in Restricted Semantic Domains*. De Gruyter, Berlin.

Henry Kučera and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, Rhode Island.

Adrienne Lehrer. 1974. *Semantic Fields and Lexical Structure*. North Holland, Amsterdam and New York.

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

Catherine Macleod and Ralph Grishman, 1995. *COMLEX Syntax Reference Manual*. Proteus Project, New York University.

Catherine Macleod, Adam Meyers, and Ralph Grishman. 1996. The influence of tagging on the classification of lexical complements. In *Proceedings of COLING-96*, Copenhagen, Denmark.

John G. McMahon and Francis J. Smith. 1996. Improving statistical language model performance with automatically generated word hierarchies. *Computational Linguistics*, **22**(2):217–247, June.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography (special issue)*, **3**(4):235–312.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. Cognitive Science Laboratory, Princeton University.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, **17**(1):21–48.

Megan Moser and Johanna D. Moore. 1995. Investigating cue selection and placement in tutorial discourse. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 130–135, Cambridge, Massachusetts, June.

Rebecca J. Passonneau and Diane J. Litman. Forthcoming. Combining multiple knowledge sources for discourse segmentation. *Computational Linguistics*. Special Issue on Empirical Studies in Discourse Interpretation and Generation.

Rebecca J. Passonneau, Karen K. Kukich, Jacques Robin, Vasileios Hatzivassiloglou, Larry Lefkowitz, and Hongyan Jing. 1996. Generating summaries of work flow diagrams. In *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*, New Brunswick, Canada, June. University of Moncton.

Rebecca J. Passonneau. 1996. Using centering to relax informational constraints on discourse anaphoric noun phrases. *Language and Speech*, **39**(2–3):229–264, April–September. Special Double Issue on Discourse and Syntax.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, Ohio, June.

James Pustejovsky, Sabine Bergler, and Peter Anick. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, **19**(2):331–359, June. Special Issue on Using Large Corpora: II.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, volume 1, pages 448–453, Montréal, Québec, Canada, August. Morgan Kaufmann, San Mateo, California.

Jacques Robin. 1994. *Revision-Based Generation of Natural Language Summaries Providing Historical Background: Corpus-Based Analysis, Design, Implementation, and Evaluation*. Ph.D. thesis, Department of Computer Science, Columbia University, New York. Also Technical Report CU-CS-034-94.

David Yarowsky. 1993. One sense per collocation. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 266–271, Plainsboro, New Jersey, March. ARPA Software and Intelligent Systems Technology Office, Morgan Kaufmann, San Francisco, California.