Inducing Frame Semantic Verb Classes from WordNet and LDOCE

Rebecca Green,[*†‡] **Bonnie J. Dorr,**[*†] and **Philip Resnik**[*†]
[*]Institute for Advanced Computer Studies
[†]Department of Computer Science
[‡]College of Information Studies
University of Maryland
College Park, MD 20742 USA
{rgreen, bonnie, resnik}@umiacs.umd.edu

Abstract

This paper presents SemFrame, a system that induces frame semantic verb classes from WordNet and LDOCE. Semantic frames are thought to have significant potential in resolving the paraphrase problem challenging many language-based applications.

When compared to the handcrafted FrameNet, SemFrame achieves its best recall-precision balance with 83.2% recall (based on SemFrame's coverage of FrameNet frames) and 73.8% precision (based on SemFrame verbs' semantic relatedness to frame-evoking verbs). The next best performing semantic verb classes achieve 56.9% recall and 55.0% precision.

1 Introduction

Semantic content can almost always be expressed in a variety of ways. Lexical synonymy (*She esteemed him highly* vs. *She respected him greatly*), syntactic variation (*John paid the bill* vs. *The bill was paid by John*), overlapping meanings (*Anna turned at Elm* vs. *Anna rounded the corner at Elm*), and other phenomena interact to produce a broad range of choices for most language generation tasks (Hirst, 2003; Rinaldi et al., 2003; Kozlowski et al., 2003). At the same time, natural language understanding must recognize what remains constant across paraphrases.

The paraphrase phenomenon affects many computational linguistic applications, including information retrieval, information extraction, question-answering, and machine translation. For example, documents that express the same content using different linguistic means should typically be retrieved for the same queries. Information sought to answer a question needs to be recognized no matter how it is expressed.

Semantic frames (Fillmore, 1982; Fillmore and Atkins, 1992) address the paraphrase problem through their slot-and-filler templates, representing frequently occurring, structured experiences. Semantic frame types of an intermediate granularity have the potential to fulfill an interlingua role within a solution to the paraphrase problem.

Until now, semantic frames have been generated by hand (as in Fillmore and Atkins, 1992), based on native speaker intuition; the FrameNet project (http://www.icsi.berkeley.edu/~framenet; Johnson et al., 2002) now couples this generation with empirical validation. Only recently has this project begun to achieve relative breadth in its inventory of semantic frames. To have a comprehensive inventory of semantic frames, however, we need the capacity to generate semantic frames semi-automatically (the need for manual post-editing is assumed).

To address these challenges, we have developed SemFrame, a system that induces semantic frames automatically. Overall, the system performs two primary functions: (1) identification of sets of verb senses that evoke a common semantic frame (in the sense that lexical units call forth corresponding conceptual structures); and (2) identification of the conceptual structure of semantic frames. This paper explores the first task of identifying frame semantic verb classes. These classes have several types of uses. First, they are the basis for identifying the internal structure of the frame proper, as set forth in Green and Dorr, 2004. Second, they may be used to extend FrameNet. Third, they support applications needing access to sets of semantically related words, for example, text segmentation and word sense disambiguation, as explored to a limited degree in Green, 2004.

Section 2 presents related research efforts on developing semantic verb classes. Section 3 summarizes the features of WordNet (http://www.cogsci.princeton.edu/~wn) and LDOCE (Procter, 1978) that support the

automatic induction of semantic verb classes, while Section 4 sets forth the approach taken by SemFrame to accomplish this task. Section 5 presents a brief synopsis of SemFrame's results, while Section 6 presents an evaluation of SemFrame's ability to identify semantic verb classes of a FrameNet-like nature. Section 7 summarizes our work and motivates directions for further development of SemFrame.

2 Previous Work

The EAGLES (1998) report on semantic encoding differentiates between two approaches to the development of semantic verb classes: those based on syntactic behavior and those based on semantic criteria.

Levin (1993) groups verbs based on an analysis of their syntactic properties, especially their ability to be expressed in diathesis alternations; her approach reflects the assumption that the syntactic behavior of a verb is determined in large part by its meaning. Verb classes at the bottom of Levin's shallow network group together (quasi-) synonyms, hierarchically related verbs, and antonyms, alongside verbs with looser semantic relationships.

The verb categories based on Pantel and Lin (2002) and Lin and Pantel (2001) are induced automatically from a large corpus, using an unsupervised clustering algorithm, based on syntactic dependency features. The resulting clusters contain synonyms, hierarchically related verbs, and antonyms, as well as verbs more loosely related from the perspective of paraphrase.

The handcrafted WordNet (Fellbaum, 1998a) uses the hyperonymy/hyponymy relationship to structure the English verb lexicon into a semantic network. Each collection of a top-level node supplemented by its descendants may be seen as a semantic verb class.

In all fairness, resolution of the paraphrase problem is not the explicit goal of most efforts to build semantic verb classes. However, they can process some paraphrases through lexical synonymy, hierarchically related terms, and antonymy.

3 Resources Used in SemFrame

We adopt an approach that relies heavily on pre-existing lexical resources. Such resources have several advantages over corpus data in identifying semantic frames. First, both definitions and example sentences often mention their participants using semantic-type-like nouns, thus mapping easily to the corresponding frame element. Corpus data, however, are more likely to include instantiated participants, which may not generalize to the frame element. Second, lexical resources provide a consistent amount of data for word senses, while the amount of data in a corpus for word senses is likely to vary widely. Third, lexical resources provide their data in a more systematic fashion than do corpora.

Most centrally, the syntactic arguments of the verbs used in a definition often correspond to the semantic arguments of the verb being defined. For example, Table 1 gives the definitions of several verb senses in LDOCE that evoke the COMMERCIAL TRANSACTION frame, which includes as its semantic arguments a Buyer, a Seller, some Merchandise, and Money. Words corresponding to the Money (*money, value*), the Merchandise (*property, goods*), and the Buyer (*buyer, buyers*) are present in, and to some extent shared across, the definitions; however, no words corresponding to the Seller are present.

Verb sense	LDOCE Definition
buy 1	to obtain (something) by giving *money* (or something else of *value*)
buy 2	to obtain in exchange for something, often something of great *value*
buy 3	to be exchangeable for
purchase 1	to gain (something) at the cost of effort, suffering, or loss of something of *value*
sell 1	to give up (*property* or *goods*) to another for money or other *value*
sell 2	to offer (*goods*) for sale
sell 3	to be bought; get a *buyer* or *buyers*; gain a sale

Table 1. LDOCE Definitions for Verbs Evoking the COMMERCIAL TRANSACTION Frame

Of available machine-readable dictionaries, LDOCE appears especially useful for this research. It uses a restricted vocabulary of about 2000 words in its definitions and example sentences, thus increasing the likelihood that words with closely related meanings will use

the same words in their definitions and support the pattern of discovery envisioned. LDOCE's subject field codes also accomplish some of the same type of grouping as semantic frames.

WordNet is a machine-readable lexico-semantic database whose primary organizational structure is the synset—a set of synonymous word senses. A limited number of relationship types (e.g., antonymy, hyponymy, meronymy, troponymy, entailment) also relate synsets within a part of speech. (Version 1.7.1 was used.)

Fellbaum (1998b) suggests that relationships in WordNet "reflect some of the structure of frame semantics" (p. 5). Through the relational structure of WordNet, *buy, purchase, sell*, and *pay* are related together: *buy* and *purchase* comprise one synset; they entail *pay*ing and are opposed to *sell.*

The relationship of *buy, purchase, sell*, and *pay* to other COMMERCIAL TRANSACTION verbs—for example, *cost, price,* and the demand payment sense of *charge*—is not made explicit in WordNet, however. Further, as Roger Chaffin has noted, the specialized vocabulary of, for example, tennis (e.g. *racket, court, lob*) is not co-located, but is dispersed across different branches of the noun network (Miller, 1998, p. 34).

4 SemFrame Approach

SemFrame gathers evidence about frame semantic relatedness between verb senses by analyzing LDOCE and WordNet data from a variety of perspectives. The overall approach used is shown in Figure 1. The first stage of processing extracts pairs of LDOCE and WordNet verb senses that potentially evoke the same frame. By exploiting many different clues to semantic relatedness, we overgenerate these pairs, favoring recall; subsequent stages improve the precision of the resulting data.

Figures 2 and 3 give details of the algorithms for extracting verb pairs based on different types of evidence. These include: clustering LDOCE verb senses/WordNet synsets on the basis of words in their definitions and example sentences (fig. 2); relating LDOCE verb senses defined in terms of the same verb (fig. 3a); relating LDOCE verb senses that share a common stem (fig. 3b); extracting explicit sense-linking relationships in LDOCE (fig. 3c); relating verb senses that share general or specific subject field codes in LDOCE (fig. 3d); and extracting (direct or extended) semantic relationships in WordNet (fig. 3e).

In the second stage, mapping between WordNet verb synsets and LDOCE verb senses relies on finding matches between the data available for the verb senses in each resource (e.g., other words in the synset; words in definitions and example sentences; words closely related to these words; and stems of these words). The similarity measure used is the average of the proportion of words on each side of the comparison that are matched in the other. This mapping is used both to relate LDOCE verb senses, that map to the same WordNet synset (fig. 3f) and to translate previously paired WordNet verb synsets into LDOCE verb sense pairs.

In the third stage, the resulting verb sense pairs are merged into a single data set, retaining only those pairs whose cumulative support exceeds thresholds for either the number of supporting data sources or strength of support, thus achieving higher precision in the merged data set than in the input data sets. Then, the graph formed by the verb sense pairs in the merged data set is analyzed to find the fully connected components.

Finally, these groups of verb senses become input to a clustering operation (Voorhees, 1986). Those groups whose similarity (due to overlap in membership) exceed a threshold are merged together, thus reducing the number of verb sense groups. The verb senses within each resulting group are hypothesized to evoke the same semantic frame and constitute a *frameset*.

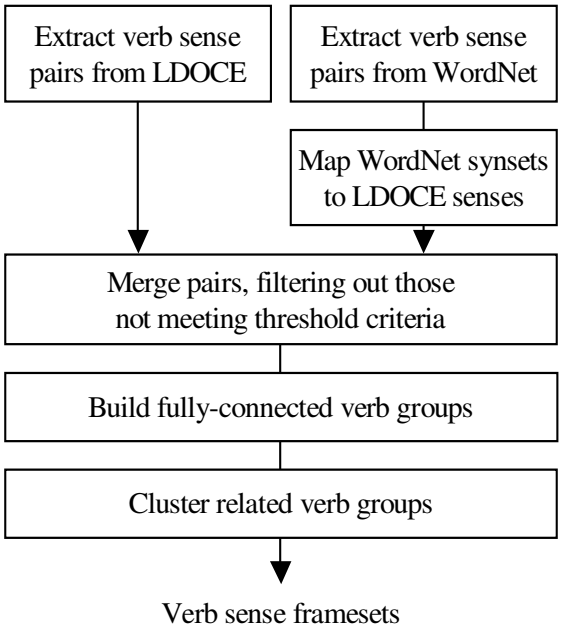

Figure 1. Approach for Building Frame Semantic Verb Classes

Input. *SW*, a set of stop words; *M*, a set of (*word, stem*) pairs; *F*, a set of (*word, frequency*) pairs; *DE*, a set of (*verb_sense_id, def+ex*) pairs, where $def+ex_d$ = the set of words in the definitions and example sentences of $verb\_sense\_id_d$

Step 1. forall $d \in DE$, append to $def+ex_d$: $verb\_sense\_id_d$ and remove from $def+ex_d$ any word $w \in SW$

Step 2. forall $d \in DE$
forall $m \in M$
if $word_m$ exists in $def+ex_d$, substitute $stem_m$ for $word_m$

Step 3. forall $f \in F$
if $frequency_f > 1$,

$$wgt_{word_f} \leftarrow \frac{1}{frequency_f},$$

else if $frequency_f == 1$,

$$wgt_{word_f} \leftarrow .01$$

Step 4. $O \leftarrow$ Voorhees' average link clustering algorithm applied to *DE*, with initial weights forall *t* in *def+ex* set to wgt_t

Step 5. forall $o \in$ O
return all combinations of two members from *o*

Figure 2. Algorithm for Generating Clustering-based Verb Pairs

5 Results

We explored a range of thresholds in the final stage of the algorithm.[1] In general, the lower the threshold, the looser the verb grouping. The number of verb senses retained (out of 12,663 non-phrasal verb senses in LDOCE) and the verb sense groups produced by using these thresholds are recorded in Table 2.

Threshold	Num verb senses	Num groups
0.5	6461	1338
1.0	6414	1759
1.5	5607	1421
2.0	5604	1563

Table 2. Results of Frame Clustering Process

6 Evaluation

One of our goals is to produce sets of verb senses capable of extending FrameNet's coverage while requiring reasonably little post-editing. This goal has two subgoals: identifying new frames and identifying additional lexical units that evoke previously recognized frames. We use the hand-crafted FrameNet, which is of reliably high precision, as a gold standard[2] for the initial evaluation of SemFrame's ability to achieve these subgoals. For the first, we evaluate SemFrame's ability to generate frames that correspond to FrameNet's frames, reasoning that the system must be able to identify a large proportion of known frames if the quality of its output is good enough to identify new frames. (At this stage we do not measure the quality of new frames.) For the second subgoal we can be more concrete: For frames identified by both systems, we measure the degree to which the verbs identified by SemFrame can be shown to evoke those frames, even if FrameNet has not identified them as frame-evoking verbs.

FrameNet includes hierarchically organized frames of varying levels of generality: Some semantic areas are covered by a general frame, some by a combination of specific frames, and some by a mix of general and specific frames. Because of this variation we determined the degree to which SemFrame and FrameNet overlap by automatically finding and comparing corresponding frames instead of fully equivalent frames. Frames *correspond* if the semantic scope of one frame is included within the semantic

[1] For the clustering algorithm used, the clustering threshold range is open-ended. The values investigated in the evaluation are fairly low.

[2] Certain constraints imposed by FrameNet's development strategy restrict its use as a full-fledged gold standard for evaluating semantic frame induction. (1) As of summer 2003, only 382 frames had been identified within the FrameNet project. (2) Low recall affects not only the set of semantic frames identified by FrameNet, but also the sets of frame-evoking units listed for each frame. No verbs are listed for 38.5% of FrameNet's frames, while another 13.1% of them list only 1 or 2 verbs. The comparison here is limited to the 197 FrameNet frames for which at least one verb is listed with a counterpart in LDOCE. (3) Some of FrameNet's frames are more syntactically than semantically motivated (e.g., EXPERIENCER-OBJECT, EXPERIENCER-SUBJECT).

a. Relates LDOCE verb senses that are defined in terms of the same verb

Input. *D*, a set of (*verb_sense_id, def_verb*) pairs, where def_verb_d = the verb in terms of which $verb\_sense\_id_d$ is defined

Step 1. forall *v* that exist as *def_verb* in *D*, form $DV_v \subset D$, by extracting all (*verb_sense_id, def_verb*) pairs where *v* = *def_verb*

Step 2. remove all DV_v for which $| DV_v | > 40$

Step 3. forall *v* that exist as *def_verb* in *D*, *return* all combinations of two members from DV_v

---

b. Relates LDOCE verb senses that share a common stem

Input. *D*, a set of (*verb_sense_id, verb_stem*) pairs, where $verb\_stem_d$ = the stem for the verb on which $verb\_sense\_id_d$ is based

Step 1. forall *m* that exist as *verb_stem* in *D*, form $DV_m \subset D$, by extracting all (*verb_sense_id, verb_stem*) pairs where *m* = *verb_stem*

Step 2. forall *m* that exist as *verb_stem* in *D*, *return* all combinations of two members from DV_v

---

c. Extracts explicit sense-linking relationships in LDOCE

Input. *D*, a set of (*verb_sense_id, def*) pairs, where def_d = the definition for $verb\_sense\_id_d$

Step 1. forall $d \in D$, if def_d contains *compare* or *opposite* note, extract *related_verb* from note; generate ($verb\_sense\_id_d$, $related\_verb_d$) pair

Step 2. forall $d \in D$, if def_d defines $verb\_sense\_id_d$ in terms of a related standalone verb (in BLOCK CAPS), extract *related_verb* from definition; generate ($verb\_sense\_id_d$, $related\_verb_d$) pair

Step 3. forall ($verb\_sense\_id_d$, $related\_verb_d$) pairs, if there is only one sense of $related\_verb_d$, choose it and *return* ($verb\_sense\_id_d$, $related\_verb\_sense\_id_d$), else apply generalized mapping algorithm to *return* ($verb\_sense\_id_d$, $related\_verb\_sense\_id_d$) pairs where overlap occurs in the glosses of $verb\_sense\_id_d$ and $related\_verb\_sense\_id_d$

---

d. Relates verb senses that share general or specific subject field codes in LDOCE

Input. *D*, a set of (*verb_sense_id, subject_code*) pairs, where $subject\_code_d$ = any 2- or 4-character subject field code assigned to *verb_sense_id*

Step 1. forall *c* that exist as *subject_code* in *D*, form $DV_c \subset D$, by extracting all (*verb_sense_id, subject_code*) pairs where *c* = *subject_code*

Step 2. forall *c* that exist as *subject_code* in *D*,
return all combinations of two members from DV_v

---

e. Extracts (direct or extended) semantic relationships in WordNet

Input. WordNet data file for verb synsets

Step 1. forall synset lines in input file
return (synset, related_synset) pairs for all synsets directly related through hyponymy, antonymy, entailment, or cause_to relationships in WordNet
(for extended relationship pairs, also *return* (synset, related_synset) pairs for all synsets within hyponymy tree, i.e., no matter how many levels removed)

---

f. Relates LDOCE verb senses that map to the same WordNet synset

Input. mapping of LDOCE verb senses to WordNet synsets

Step 1. forall lines in input file
return all combinations of two LDOCE verb senses mapped to the same WordNet synset

Figure 3. Algorithms for Generating Non-clustering-based Verb Pairs

scope of the other frame or if the semantic scopes of the two frames have significant overlap. Since FrameNet lists evoking words, without specification of word sense, the comparison was done on the word level rather than on the word sense level, as if LDOCE verb senses were not specified in SemFrame. However, it is clearly specific word senses that evoke frames, and SemFrame's verb classes list specific LDOCE verb senses. In extending FrameNet, verbs from SemFrame would be word-sense-disambiguated in the same way that FrameNet verbs currently are, through the correspondence of lexeme and frame.

Incompleteness in the listing of evoking verbs in FrameNet and SemFrame precludes a straight-

forward detection of correspondences between their frames. Instead, correspondence between FrameNet and SemFrame frames is established using either of two somewhat indirect approaches.

In the first approach, a SemFrame frame is deemed to correspond to a FrameNet frame if the two frames meet both a *minimal-overlap criterion* (i.e., there is some, perhaps small, overlap between the FrameNet and SemFrame framesets) and a *frame-name-relatedness criterion*. The minimal-overlap criterion is met if either of two conditions is met: (1) If the FrameNet frame lists four or fewer verbs (true of over one-third of the FrameNet frames that list associated verbs), minimal overlap occurs when any one verb associated with the FrameNet frame matches a verb associated with a SemFrame frame. (2) If the FrameNet frame lists five or more verbs, minimal overlap occurs when two or more verbs in the FrameNet frame are matched by verbs in the SemFrame frame.

The looseness of the minimal overlap criterion is tightened by also requiring that the names of the FrameNet and SemFrame frames be closely related. Establishing this frame-name relatedness involves identifying individual components of each frame name[3] and augmenting this set with morphological variants from CatVar (Habash and Dorr 2003). The resulting set for each FrameNet and SemFrame frame name is then searched in both the noun and verb WordNet networks to find all the synsets that might correspond to the frame name. To these sets are also added all synsets directly related to the synsets corresponding to the frame names. If the resulting set of synsets gathered for a FrameNet frame name intersects with the set of synsets gathered for a SemFrame frame name, the two frame names are deemed to be semantically related.

For example, the FrameNet ADORNING frame contains 17 verbs: *adorn, blanket, cloak, coat, cover, deck, decorate, dot, encircle, envelop, festoon, fill, film, line, pave, stud,* and *wreathe*. The SemFrame ORNAMENTATION frame contains 12 verbs: *adorn, caparison, decorate, embellish, embroider, garland, garnish, gild, grace, hang, incrust,* and *ornament*. Two of the verbs—*adorn* and *decorate*—are shared. In addition, the frame names are semantically related through a WordNet synset consisting of *decorate, adorn* (which CatVar relates to ADORNING), *grace, ornament* (which CatVar relates to ORNAMENTATION), *embellish,* and *beautify.* The two frames are therefore designated as corresponding frames by meeting both the minimal-overlap and the frame-name relatedness criteria.

In the second approach, a SemFrame frame is deemed to correspond to a FrameNet frame if the two frames meet either of two relatively stringent verb overlap criteria, the *majority-match criterion* or the *majority-related criterion,* in which case examination of frame names is unnecessary.

The majority-match criterion is met if the set of verbs shared by FrameNet and SemFrame framesets account for half or more of the verbs in either frameset. For example, the APPLY_HEAT frame in FrameNet includes 22 verbs: *bake, blanch, boil, braise, broil, brown, char, coddle, cook, fry, grill, microwave, parboil, poach, roast, saute, scald, simmer, steam, steep, stew,* and *toast*, while the BOILING frame in SemFrame includes 7 verbs: *boil, coddle, jug, parboil, poach, seethe,* and *simmer.* Five of these verbs—*boil, coddle, parboil, poach,* and *simmer*—are shared across the two frames and constitute over half of the SemFrame frameset. Therefore the two frames are deemed to correspond by meeting the majority-match criterion.

The majority-related criterion is met if half or more of the verbs from the SemFrame frame are semantically related to verbs from the FrameNet frame (that is, if the precision of the SemFrame verb set is at least 0.5). To evaluate this criterion, each FrameNet and SemFrame verb is associated with the WordNet verb synsets it occurs in, augmented by the synsets to which the initial sets of synsets are directly related. If the sets of synsets corresponding to two verbs share one or more synsets, the two verbs are deemed to be semantically related. This process is extended one further level, such that a SemFrame verb found by this process to be semantically related to a SemFrame verb, whose semantic relationship to a FrameNet verb has already been established, will also be designated a frame-evoking verb. If half or more of the verbs listed for a SemFrame frame are established as evoking the same frame as the list of WordNet verbs, then the FrameNet

[3] All SemFrame frame names are nouns. (See Green and Dorr, 2004 for an explanation of their selection.) FrameNet frame names (e.g., ABUNDANCE, ACTIVITY_START, CAUSE_TO_BE_WET, INCHOATIVE_ATTACHING), however, exhibit considerable variation.

and SemFrame frames are hypothesized to correspond through the majority-related criterion.

For example, the FrameNet ABUNDANCE frame includes 4 verbs: *crawl, swarm, teem,* and *throng*. The SemFrame FLOW frame likewise includes 4 verbs: *pour, teem, stream,* and *pullulate*. Only one verb—*teem*—is shared, so the majority-match criterion is not met, nor is the related-frame-name criterion met, as the frame names are not semantically related. The majority-related criterion, however, is met through a WordNet verb synset that includes *pour, swarm, stream,* teem, and *pullulate*.

Of the 197 FrameNet frames that include at least one LDOCE verb, 175 were found to have a corresponding SemFrame frame. But this 88.8% recall level should be balanced against the precision ratio of SemFrame verb framesets. After all, we could get 100% recall by listing all verbs in every SemFrame frame.

The majority-related function computes the precision ratio of the SemFrame frame for each pair of FrameNet and SemFrame frames being compared. By modifying the minimum precision threshold, the balance between recall and precision, as measured using F-score, can be investigated. The best balance for the SemFrame version is based on a clustering threshold of 2.0 and a minimum precision threshold of 0.4, which yields a recall of 83.2% and overall precision of 73.8%.

To interpret these results meaningfully, one would like to know if SemFrame achieves more FrameNet-like results than do other available verb category data, more specifically the 258 verb classes from Levin, the 357 semantic verb classes of WordNet 1.7.1, or the 272 verb clusters of Lin and Pantel, as described in Section 2.

For purposes of comparison with FrameNet, Levin's verb class names have been hand-edited to isolate the word that best captures the semantic sense of the class; the name of a WordNet-based frame is taken from the words for the root-level synset; and the name of each Lin and Pantel cluster is taken to be the first verb in the cluster.[4]

Evaluation results for the best balance between recall and precision (i.e., the maximum F-score) of the four comparisons are summarized in Table 3. FrameNet itself constitutes the upper bound on the task, i.e., 100% recall and 100% precision. The Lin & Pantel results are here a lower bound for automatically induced semantic verb classes and probably reflect the limitations of using only corpus data. Among efforts to develop semantic verb classes, SemFrame's results correspond more closely to semantic frames than do others.

Semantic verb classes	Precision threshold at max F-score	Recall	Precision
SemFrame	0.40	0.832	0.738
Levin	0.20	0.569	0.550
WordNet	0.15	0.528	0.466
Lin & Pantel	0.15	0.472	0.407

Table 3. Best Recall-Precision Balance When Compared with FrameNet

7 Conclusions and Future Work

We have demonstrated that sets of verbs evoking a common semantic frame can be induced from existing lexical tools. In a head-to-head comparison with frames in FrameNet, the frame semantic verb classes developed by the SemFrame approach achieve a recall of 83.2% and the verbs listed for frames achieve a precision of 73.8%; these results far outpace those of other semantic verb classes. On a practical level, a large number of frame semantic verb classes have been identified. Associated with clustering threshold 1.5 are 1421 verb classes, averaging 14.1 WordNet verb synsets. Associated with clustering threshold 2.0 are 1563 verb classes, averaging 6.6 WordNet verb synsets.

Despite these promising results, we are limited by the scope of our input data set. While LDOCE and WordNet data are generally of high quality, the relative sparseness of these resources has an adverse impact on recall. In addition, the mapping technique used for picking out corresponding word senses in WordNet and LDOCE is shallow, thus constraining the recall and precision of SemFrame outputs. Finally, the multi-step process of merging smaller verb groups into verb groups that are intended to correspond to frames sometimes fails to achieve an appropriate degree of correspondence (all the verb classes discovered are not distinct).

[4]Lin and Pantel have taken a similar approach, "naming" their verb clusters by the first three verbs listed for a cluster, i.e., the three most similar verbs.

In our future work, we will experiment with the more recent release of WordNet (2.0). This version provides derivational morphology links between nouns and verbs, which will promote far greater precision in the linking of verb senses based on morphology than was possible in our initial implementation. Another significant addition to WordNet 2.0 is the inclusion of category domains, which co-locate words pertaining to a subject and perform the same function as LDOCE's subject field codes.

Finally, data sparseness issues may be addressed by supplementing the use of the lexical resources used here with access to, for example, the British National Corpus, with its broad coverage and carefully-checked parse trees.

Acknowledgments

This research has been supported in part by a National Science Foundation Graduate Research Fellowship NSF ITR grant #IIS-0326553, and NSF CISE Research Infrastructure Award EIA0130422.

References

Boguraev, Bran and Ted Briscoe. 1989. Introduction. In B. Boguraev and T. Briscoe (Eds.), *Computational Lexicography for Natural Language Processing*, 1-40. London: Longman.

EAGLES Lexicon Interest Group. 1998. *EAGLES Preliminary Recommendations on Semantic Encoding: Interim Report*, <http:// www.ilc.cnr.it/EAGLES96/rep2/ rep2.html>.

Fellbaum, Christiane (Ed.). 1998a. *WordNet: An Electronic Lexical Database.* Cambridge, MA: The MIT Press.

Fellbaum, Christiane. 1998b. Introduction. In C. Fellbaum, 1998a, 1-17.

Fillmore, Charles J. 1982. Frame semantics. In *Linguistics in the Morning Calm*, 111-137. Seoul: Hanshin.

Fillmore, Charles J. and B. T. S. Atkins. 1992. Towards a frame-based lexicon: The semantics of RISK and its neighbors. In A. Lehrer and E. F. Kittay (Eds.), *Frames, Fields, and Contrasts*, 75-102. Hillsdale, NJ: Erlbaum.

Green, Rebecca. 2004. Inducing Semantic Frames from Lexical Resources. Ph.D. dissertation, University of Maryland.

Green, Rebecca and Bonnie J. Dorr. 2004. Inducing A Semantic Frame Lexicon from WordNet Data. In *Proceedings of the 2nd Workshop on Text Meaning and Interpretation* (ACL 2004).

Habash, Nizar and Bonnie Dorr. 2003. A categorial variation database for English. In *Proceedings of North American Association for Computational Linguistics*, 96-102.

Hirst, Graeme. 2003. Paraphrasing paraphrased. Keynote address for *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications*, ACL 2003, <http://nlp.nagaokaut.ac.jp/IWP2003/pdf/ Hirst-slides.pdf>.

Johnson, Christopher R., Charles J. Fillmore, Miriam R. L. Petruck, Collin F. Baker, Michael Ellsworth, Josef Ruppenhofer, and Esther J. Wood. 2002. *FrameNet: Theory and Practice*, version 1.0, <http://www.icsi.berkeley.edu/ ~framenet/book/book.html>.

Kozlowski, Raymond, Kathleen F. McCoy, and K. Vijay-Shanker. 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003),* ACL 2003, 1-8.

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

Lin, Dekang and Patrick Pantel. 2001. Induction of semantic classes from natural language text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 317-322.

Litkowski, Ken. 2004. Senseval-3 task: Word-sense disambiguation of WordNet glosses, <http://www.clres.com/SensWNDisamb.html>.

Miller, George A. 1998. Nouns in WordNet. In C. Fellbaum, 1998a, 23-67.

Pantel, Patrick and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 613-619.

Procter, Paul (Ed.). 1978. *Longman Dictionary of Contemporary English*. Longman Group Ltd., Essex, UK.

Rinaldi, Fabio, James Dowdall, Kaarel Kaljurand, Michael Hess, and Diego Mollá. 2003. Exploiting paraphrases in a question answering system. In *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003),* ACL 2003, 25-32.

Voorhees, Ellen. 1986. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing & Management* 22/6: 465-476.