Frame Semantic Enhancement of Lexical-Semantic Resources

Rebecca Green
Institute of Advanced Computer Studies
College of Information Studies
University of Maryland
College Park, MD 20742, USA
rgreen@umd.edu

Bonnie J. Dorr
Institute of Advanced Computer Studies
Department of Computer Science
University of Maryland
College Park, MD 20742, USA
bonnie@umiacs.umd.edu

Abstract

SemFrame generates FrameNet-like frames, complete with semantic roles and evoking lexical units. This output can enhance FrameNet by suggesting new frames, as well as additional lexical units that evoke existing frames. SemFrame output can also support the addition of frame semantic relationships to WordNet.

1 Introduction

The intuition that semantic analysis can make a positive contribution to language-based applications has motivated the development of a number of lexical-semantic resources. Prominent among them are WordNet,[1] PropBank,[2] and FrameNet.[3] The potential contribution of these resources is constrained by the information they contain and the level of effort involved in their development.

For example, semantic annotation tasks (Baker et al., 2004) typically assign semantic roles to the arguments of predicates. The benefit of the semantic annotation is constrained by the presence and quality of semantic roles in the lexical-semantic resource(s) used. Gildea and Jurafsky (2002) suggest that the availability of semantic annotation of this sort is useful for information extraction, word sense disambiguation, machine translation, text summarization, text mining, and speech recognition.

[1] http://www.cogsci.princeton.edu/~wn
[2] http://www.cis.upenn.edu/~ace
[3] http://framenet.icsi.berkeley.edu

Other tasks rely on the identification of semantic relationships to recognize lexical chains (sets of semantically related words that enable a text to be cohesive) (Morris and Hirst, 1991). The success of this work is constrained by the set of semantic relationship types and instantiations underlying the recognition of lexical chains. As Stokes's dissertation (2004) notes, lexical cohesion has been used in discourse analysis, text segmentation, word sense disambiguation, text summarization, topic detection and tracking, and question answering.

Unfortunately, most lexical-semantic resources, including those previously mentioned, are the product of considerable ongoing human effort. Given the high development costs associated with these resources, the possibility of enhancing them on the basis of complementary resources that are produced automatically is welcome.

This paper demonstrates several of the characteristics and benefits of SemFrame (Green et al., 2004; Green and Dorr, 2004), a system that produces such a resource.

1. SemFrame generates semantic frames in a form like those of FrameNet, the ostensible gold standard for semantic frames.

2. Some SemFrame frames correspond to FrameNet frames. When SemFrame identifies additional lexical units that evoke the frame, it bolsters the use of semantic frames for identifying lexical chains.

3. Some SemFrame frames cover semantic space not yet investigated in FrameNet, which, be-

Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition, pages 57–66,
Ann Arbor, June 2005.

cause of the labor-intensive nature of its development, is incomplete. The identification of new frames thus helps fill in gaps in FrameNet.

4. In addition to complementing FrameNet, SemFrame could be used as a more systematic source of semantic roles for PropBank or could serve as the basis for adding frame semantic relationships to WordNet.

The rest of the paper is organized as follows: Section 2 discusses lexical-semantic resources that could be enhanced by using SemFrame's output. Section 3 sets out how SemFrame works, with Subsections 3.1 and 3.2 explaining, respectively, the identification of lexical units that evoke shared semantic frames and the generation of the internal structure of those frames. Section 4 discusses how we evaluate SemFrame's output. Finally, Section 5 summarizes SemFrame's contributions and sketches future directions in its development.

2 Lexical-Semantic Resources

Lexical-semantic resources, such as FrameNet and PropBank, which involve semantic frames and/or semantic roles, are one kind of resource that SemFrame's output can enhance. SemFrame could also benefit a resource like WordNet that captures different kinds of semantic relationships. Here we discuss characteristics of these resources that make them amenable to enhancement through SemFrame.

2.1 FrameNet

FrameNet documents the semantic and syntactic behavior of words with respect to frames. A frame characterizes a conventional conceptual structure, for instance, a situation involving risk, a hitting event, a commercial transaction. Lexical units are said to evoke a frame. For example, use of the literal sense of *buy* introduces into a discourse an expectation that some object or service (the Goods) passes from one person (the Seller) to another (the Buyer) in exchange for something of (presumably equivalent) value (typically Money).

A significant contribution of the FrameNet project is the creation of frames, which involves the enumeration both of participant roles in the frame (a.k.a, frame elements, frame slots) and of lexical units that evoke the frame. As of May 2005, 657 frames have been defined in FrameNet; approximately 8600 lexical unit/frame associations have been made.

FrameNet's approach to identifying frames is "opportunistic" and driven by the corpus data being annotated. Thus the FrameNet team does not expect to have a full inventory of frames until a substantial proportion of the general-purpose vocabulary of English has been analyzed. As the development of FrameNet is labor-intensive, supplementing FrameNet's frames and evoking lexical units using data from SemFrame would be beneficial.

2.2 PropBank

Like FrameNet, PropBank (Kingsbury et al., 2002) is a project aimed at semantic annotation, in this case of the Penn English Treebank.[4] The intent of PropBank is to provide for "automatic extraction of relational data" on the basis of consistent labeling of predicate argument relationships. Typically the labels/semantic roles are verb-specific (but are often standardized across synonyms). For example, the set of semantic arguments for *promise, pledge,* etc. (its 'roleset') includes the promiser, the person promised to, and the promised thing or action. These correspond respectively to FrameNet's Speaker, Addressee, and Message elements within the Commitment frame.

The more general labels used in FrameNet and SemFrame give evidence of a more systematic approach to semantic argument structure, more easily promoting the discovery of relationships among frames. It can be seen from the terminology used that PropBank is more focused on the individual arguments of the semantic argument structure, while FrameNet and SemFrame are more focused on the overall gestalt of the argument structure, that is, the frame. The use of FrameNet and SemFrame to suggest more generic (that is, frame-relevant) roleset labels would help move PropBank toward greater systematicity.

[4]The semantic annotation tasks in the FrameNet and PropBank projects enable them to link semantic roles and syntactic behavior. Enhancing and stabilizing its semantic frame inventory must precede the inclusion of such linkage in SemFrame.

2.3 WordNet

WordNet is a lexical database for English nouns, verbs, adjectives, and adverbs. Fine-grained sense distinctions are recognized and organized into synonym sets ('synsets'), WordNet's basic unit of analysis; each synset has a characterizing gloss, and most are exemplified through one or more phrases or sentences.

In addition to the synonymy relationship at the heart of WordNet, other semantic relationships are referenced, including, among others, antonymy, hyponymy, troponymy, partonomy, entailment, and cause-to. On the basis of these relationships, Fellbaum (1998) noted that WordNet reflected the structure of frame semantics to a degree, but suggested that its organization by part of speech would preclude a full frame-semantic approach.

With release 2.0, WordNet added morphological and topical category relationships that cross over part-of-speech boundaries. This development relates to incorporating a full frame-semantic approach in WordNet in two ways.

First, since the lexical units that evoke a frame are not restricted to a single part of speech, the ability to create links between parts of speech is required in order to encode frame semantic relationships.

Second, topical categories (e.g., slang, meat, navy, Arthurian legend, celestial body, historical linguistics, Mafia) have a kinship with semantic frames, but are not the same. While topical category domains map between categories and lexical items—as do semantic frames—it is often not clear what internal structure might be posited for a category domain. What, for example, would the participant structure of 'meat' look like?

Should WordNet choose to adopt a full frame-semantic approach, FrameNet and SemFrame are natural starting points for identifying frame-semantic relationships between synsets. The most beneficial enhancement would involve WordNet's incorporating FrameNet and/or SemFrame frames as a separate resource, with a mapping between WordNet's synsets and the semantic frame inventory. SemFrame has the extra advantage that its lexical units are already identified as WordNet synsets.

3 Development of SemFrame

There are two main processing stages in producing SemFrame output: The first establishes verb classes, while the second generates semantic frames. The next two subsections describe these stages.

3.1 Establishing Verb Classes

SemFrame adopts a multistep approach to identifying sets of frame-semantically related verb senses. The basic steps involved in the current version[5] of SemFrame are:

1. Building a graph with WordNet verb synsets as vertices and semantic relationships as edges
2. Identifying for each vertex a maximal highly connected component (HCC) (i.e., a highly interconnected subgraph that the vertex is part of)
3. Eliminating HCC's with undesirable qualities
4. Forming preliminary verb semantic classes by supplementing HCC's with reliable semantic relationships
5. Merging verb semantic classes with a high degree of overlap

Building the Relationships Graph

WordNet 2.0 includes a vast array of semantic relationships between synsets of the same part of speech and has now been enhanced with relationships linking synsets of different parts of speech. Some of these relationships are almost guaranteed to link synsets that evoke the same frame, while others operate within the bounds of a semantic frame on some occasions, but not others. Among the relationship types in WordNet most fruitful for identifying verb synsets within the same frame semantic verb class are: synonymy (e.g., *buy, purchase*, as collocated within synsets), antonymy (e.g., *buy,*

[5]The process of establishing verb classes has been redesigned. All that has been carried over from the previous/initial version of SemFrame is the use of some of the same WordNet relationships. New in the current version are: the use of relationship types first implemented in WordNet 2.0, the predominant and exclusive use of WordNet as the source of data (the previous version used WordNet as a source secondary to the Longman Dictionary of Contemporary English), and modeling the identification of classes of related verbs as a graph, specifically through the use of highly connected components.

sell), cause-to (e.g., *transfer, change hands*), entailment (e.g., *buy, pay*), verb group (e.g., different commercial senses of *buy*, morphological derivation (e.g., *buy, buyer*),[6] and "see also" (e.g., *buy, buy out*). Instances of these relationship types for all verb synsets in WordNet 2.0 are represented as edges within the graph.

Additional edges are inserted between any two synsets/vertices related by two or more of the following: clustering of synsets based on the occurrence of word stems in their glosses and example sentences;[7] hyperonymy/hyponymy relationships; and category domain relationships. These three relationship types are too noisy to be used on their own for identifying frame semantic relationships among synsets, but when a relationship is verified by two or more of these relationships, the likelihood that the related synsets evoke the same frame is considerably higher. Table 1 summarizes the number of edges in the graph supported by each relationship type.

Relationship Type	Count
Antonymy	502
Cause-to	218
Entailment	409
Verb group	874
Morphological derivation	8,986
See also	539
Two of:	2,223
Clustering	54,298
Hyperonymy/hyponymy	12,985
Category domain	18,482
Total	13,751

Table 1: Relationship Counts in WordNet 2.0

Identifying Highly Connected Components (HCC's)

Step 1 constructs a graph interconnecting thousands of WordNet verb synsets. Identifying sets of verb synsets likely to evoke the same semantic frame requires identifying subgraphs with a high degree of interconnectivity. Empirical investigation has shown that "highly connected components" (Hartuv and Shamir, 2000)—induced subgraphs of size k in which every vertex's connectivity exceeds $\frac{k}{2}$ vertices—identify such sets of verb synsets.[8] For example, in a 5-vertex highly connected component, each vertex is related to at least 3 other vertices. Figure 1 shows a portion of the original graph in which relationship arcs constituting an HCC are given as solid lines, while those that fail the interconnectivity threshold are given as dotted lines.

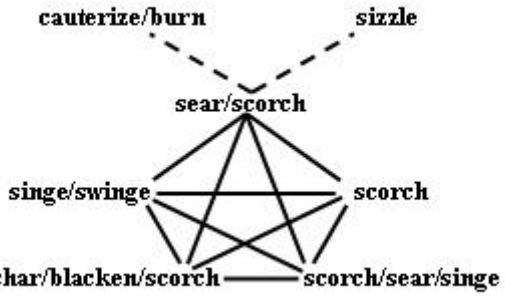

Figure 1: Relationships Subgraph with HCC

Given an undirected graph, the Hartuv-Shamir algorithm for identifying HCC's returns zero or more non-overlapping subgraphs (including zero or more singleton vertices). But it is inaccurate to assume that verb synsets evoke only a single frame, as is suggested by non-overlapping subgraphs.[9] For this reason, we have modified the Hartuv-Shamir algorithm to identify a maximal HCC, if one exists, for (i.e., that includes) each vertex of the graph. This modification reduces the effort involved in identifying any single HCC: Since the diameter of a HCC is no greater than two, only those vertices who are neighbors of the source vertex, or neighbors of those neighbors, need to be examined.

[6] SemFrame relates verb synsets with a morphological derivation relationship to a common noun synset. This includes verbs related to different members of the shared noun synset.

[7] Voorhees' (1986) hierarchical agglomerative clustering algorithm was implemented.

[8] The algorithm for computing HCC's first finds the minimum cut for a (sub)graph. If the graph meets the highly connected component criterion, the graph is returned, else the algorithm is called recursively on each of the subgraphs created by the cut. The Stoer-Wagner (1997) algorithm has been implemented for finding the minimum cut.

[9] Semantic frames can be defined at varying levels of generality; thus, a given synset may evoke a set of hierarchically related frames. Words/Synsets may also evoke multiple, unrelated frames simultaneously; *criticize*, for example, evokes both a Judging frame and a Communication frame.

Eliminating Duplicates

Because HCC's were generated for each vertex in the relationships graph, considerable duplication and overlap existed in the output. The output of step 2 was cleaned up using three filters. First, duplicate HCC's were eliminated. Second, any HCC wholly included within another HCC was deleted.[10] Third, any HCC based only on morphological derivation relationships was deleted. In SemFrame, all verb synsets morphologically derived from the same noun synset were related to each other. Thus all verb synsets derived from a common noun synset are guaranteed to generate an HCC. If only such relationships support an HCC, the likelihood that all of the interrelated verb synsets evoke the same semantic frame is much lower than if other types of relationships also provide evidence for their interrelationship.

Supplementing HCC's

The HCC's generated in step 2 that survived the filters implemented in step 3 form the basis of verb *framesets*, that is, sets of verb senses that evoke the same semantic frame. Specifically, all the synsets represented by vertices in a single HCC form a frameset.

The connectivity threshold imposed by HCC's helps maintain reasonably high precision of the resulting framesets, but is too strict for high recall. Some types of relationships known to operate within frame-semantic boundaries generally do not survive the connectivity threshold cutoff. For example, for frames of a certain level of generality, if a specific verb evokes that frame, it is also the case that its antonym evokes the frame, as antonyms operate against the backdrop of the same situational context; that is, they share participant structure.[11] However, since antonymy is (only) a *lexical* relationship between two word senses, A and B, the tight coupling of A and B is unlikely to be reflected in A's being directly related to other synsets that are related to B and vice-versa. Thus, antonyms are unlikely to be highly connected through WordNet to other words/synsets that evoke the frame and thus fail the HCC connectivity threshold. The same argument can be made for causatively related verbs. A post-processing step was required therefore to add to a frameset any verb synsets related through WordNet's antonymy or cause-to relationships to a member of the frameset. Similarly, any verb synset entailed by a member of a verb frameset was added to the frameset.

Other verb synsets fail to survive the connectivity threshold cutoff because they enter into few relationships of any kind. If a verb synset is related to only one other verb synset, the assumption is made that it evokes the same frame as that one other synset; it is then added to the corresponding frameset.

Lastly, if a synset is related to two or more members of a frameset, the likelihood that it evokes the same semantic frame is reasonably high. Such verb synsets were added to the frameset if not already present.

At the end of this phase, any framesets wholly included within another frameset were again deleted.

Merging Overlapping Verb Classes

The preceding processes produced many framesets with a significant degree of overlap. For any two framesets, if at least half of the verb synsets in both framesets were also members of the other, the two framesets were merged into a single frameset.

Summary of Stage 1 Results

The above steps generated 1434 framesets, varying in size from 2 to 25 synsets (see Table 2). Small framesets dominate the results, with over 60% of the framesets including only 2 or 3 synsets.

Representative examples of these framesets are given in Appendix A, where members of each synset appear in parentheses, followed by the synset's gloss. (Examples are ordered by frameset size.) Smaller and medium-sized framesets generally enjoy high precision, but many of the largest framesets would be better split into two or more framesets.

3.2 Generating Semantic Frames

Generating frames from verb framesets relies on the insight that the semantic arguments of a frame are largely drawn from nouns associated with verb

[10] Given the interest in generating semantic frames of varying levels of generality, this filter may itself be eliminated in the future.

[11] Identifying antonyms is especially helpful in the case of conversives, as with *buy* and *sell*; the inclusion of both in the frameset promotes discovery of all relevant frame participants, in this case, both buyer and seller.

Frameset Size	Count
2	536
3	346
4-5	309
6-8	169
9-12	54
13-25	20
Total	1434

Table 2: Count of Frameset Sizes

synsets in the frameset. In SemFrame's processing, these include nouns in the gloss of a verb synset or in the gloss of its corresponding LDOCE verb sense(s), as well as nouns (that is, noun synsets) to which a verb synset is morphologically related and those naming the category domain to which a verb synset belongs. In the latter two cases, the nouns come disambiguated within WordNet, but nouns from glosses must undergo disambiguation. The set of noun senses associated with a verb frameset is then analyzed against the WordNet noun hierarchy, using an adaptation of Agirre and Rigau's (1995) conceptual density measure. This analysis identifies a frame name and a set of frame participants, all of which correspond to nodes in the WordNet noun hierarchy.

Disambiguating Nouns from Glosses

First we consider how nouns from WordNet and LDOCE verb glosses are disambiguated.[12] This step involves looking for matches between the stems of words in the glosses of WordNet noun synsets that include the noun needing to be disambiguated, on the one hand, and the stems of words in the glosses of all WordNet verb synsets (and corresponding LDOCE verb senses) in the frameset, on the other hand.

A similarity score is computed by dividing the match count by the number of non-stop-word stems in the senses under consideration. SemFrame favors predominant senses by examining word senses in frequency order. Any sense with a non-zero similarity score that is the highest score yet seen is chosen as an appropriate word sense.

The various nodes within WordNet's noun network that correspond to a verb frameset—either through morphological derivation or category domain relationships in WordNet or through the disambiguation of nouns from the glosses of verbs in the frameset—constitute 'evidence synsets' for the participant structure of the corresponding semantic frame and form the input for the conceptual density calculation.

In preparation for use in calculating conceptual density, evidence synsets are given weights that take into account the source and basis of the disambiguation. In the current implementation, noun synsets related to the frameset through morphological derivation or shared category domain are given a weight of 4.0 (the nouns are guaranteed to be related to the verbs, and disambiguation of the nouns is built into the fact that relationships are given between synsets); disambiguated noun synsets coming from WordNet verb synsets receive a weight of 2.0 (since the original framesets contain WordNet synsets, and the disambiguation strategy is fairly conservative); non-disambiguated nouns coming from LDOCE verbs related to the frameset have a weight of 0.5 (LDOCE verbs are a step removed from the original framesets, and the nouns have not been disambiguated); all other nouns receive a weight of 1.0. The weight for non-disambiguated nouns is ultimately distributed across the noun's senses, with higher proportions of the weight being assigned to more frequent senses.

Computing Conceptual Density

The overall idea behind transforming the list of evidence synsets into a list of participants involves using the relationship structure of WordNet to identify an appropriately small set of concepts (i.e., synsets) within WordNet that account for (i.e., are superordinate to) as many of the evidence synsets as possible; such synsets will be referred to as 'covering synsets'.

This task relies on the hypothesis that a frame's evidence synsets will not be randomly distributed across WordNet, but will be clustered in various subtrees within the hierarchy. Intuitively, when evidence synsets cluster together, the subtrees in which they occur will be more dense than those subtrees where few or no evidence synsets occur. It is hypothesized that the WordNet subtrees with the high-

[12] Identification of LDOCE verb senses that correspond to WordNet verb synsets is carried out using a similar strategy.

est density are the most likely to correspond to frame slots. Thus, the task is to identify such clusters/subtrees and then to designate the nodes at the roots of the subtrees as covering synsets (subject to certain constraints).

The conceptual density measure we have used has been inspired by the measure of the same name in Agirre and Rigau (1995). The conceptual density, CD(n), of a node *n* is computed as follows:

$$CD(n) = \frac{\sum_{i \epsilon descendants_n} (wgt_i * treesize_i)}{treesize_n}$$

Both frame names and frame slots are identified on the basis of this conceptual density measure, with the frame name being taken from the node with the highest conceptual density from a specified group of subnetworks within the WordNet noun network (including abstractions, actions, events, phenomena, psychological features, and states). Frame slots are subject to a density threshold (based on mean density and variance), an evidence-synset-support threshold, and a constraint on the number of possible slots to be taken from specific subnetworks within WordNet. Further details on the computation and interpretation of conceptual density are given in (Green and Dorr, 2004).

Frame names and frame structures for the framesets in Appendix A are given in Appendix B. The full set of SemFrame's frames (including ca. 30,000 lexical unit/frame associations) is publicly available at: http://www.cs.umd.edu/~rgreen/semframe2.tar.gz.

The correspondence between frameset sizes and the number of slots generated for the frame is worth noting, since we have independent evidence about the number of slots that should be generated. Frames in FrameNet generally have from 1 to 5 slots (occasionally more). Over 70% of SemFrame's frames contain from 1 to 5 frame slots. Of course, generating an appropriate number of frame slots is not the same as generating the right frame slots, a determination that requires empirical investigation.

4 Evaluation

Three student judges evaluated SemFrame's results, with 200 frames each assessed by two judges, and 1234 frames each assessed by one judge.

In evaluating a frame, judges began by examining the set of verb synsets deemed to evoke a common frame and identified from among them the largest subset of synsets they considered to evoke the same frame. This frame—designated the 'target frame'—was simply a mental construct in the judge's mind. For only 9% of the frame judgments were the judges unable to identify a target frame.

If a target frame was discerned, judges were then asked to evaluate whether the WordNet verb synsets and LDOCE verb senses listed by SemFrame could be used to communicate about the frame the judge had in mind. This evaluation step applied to 6147 WordNet verb synsets and 7148 LDOCE verb senses; in the judges' views, 78% of the synsets and 68% of the verb senses evoke the target frame.

Judges were asked how well the frame names generated by SemFrame capture the overall target frame. Some 53% of the names were perceived to be satisfactory (good or excellent), with another 25% of the names in the right hierarchy. Only 11% of the names were deemed to be only mediocre and 9% to be unrelated.

Judges were also asked how well the frame element names generated by SemFrame named a participant or attribute of the target frame. Here 46% of the names were found satisfactory, with another 18% of the names consistent with a target frame participant, but either too general or too narrow. Another 5% of the names were regarded as mediocre and 30% as unrelated.

Lastly, judges were asked to look for correspondences between target frames and FrameNet frames. While only 17% of the target frames were considered equivalent to a FrameNet frame, many were judged to be hierarchically related; 51% of the FrameNet frames were judged more general than the corresponding SemFrame frame, while 8% were judged more specific. This reflects the need to combine some number of SemFrame frames. For 23% of the SemFrame frames, even the best FrameNet match was considered only mediocre. These may represent viable frames not yet recognized by FrameNet. Judges also found 3668 verbs in SemFrame that could be appropriately listed for a corresponding frame in FrameNet, but were not.

These results reveal SemFrame's strengths in in-

ducing frames by enumerating sets of verbs that evoke a shared frame and in naming such frames. SemFrame's ability to postulate names for the elements of a frame is less robust, although results in this area are still noteworthy.

5 Conclusion and Future Work

SemFrame's output can be used to enhance lexical-semantic resources in various ways. For example, WordNet has recently incorporated new relationship types, some of which touch on frame semantic relationships. But frame semantic relationships are as yet only implicit in WordNet; not all morphological derivation relationships, for example, operate within a frame. Should WordNet choose to reflect frame semantic relationships, SemFrame would provide a useful point of departure, since the verb framesets, frame names, and frame slots are all already expressed as WordNet synsets.

SemFrame can also add to FrameNet. The extensive human effort that has gone into FrameNet is overwhelmingly evident in the quality of its frame structures (and attendant annotations). SemFrame is unlikely ever to compete with FrameNet on this score. However, SemFrame has identified frames not recognized in FrameNet, e.g., SemFrame's SOILING frame. SemFrame has likewise identified lexical units appropriate to FrameNet frames that have not yet been incorporated into FrameNet, e.g., *stick to*, *stick with*, and *abide by* in the COMPLIANCE / CONFORMITY frame. These contributions would add as well to the semantic representations in PropBank. Since identifying frames and their evoking lexical units from scratch requires more effort than assessing the general quality of proposed frames and lexical units—indeed, since there is currently no other systematic way in which to identify either a universal set of semantic frames or the set of lexical items that evoke a frame—SemFrame's ability to propose new frames and new evoking lexical units constitutes a major contribution to the development of lexical-semantic resources.

SemFrame's current results might themselves be enhanced by considering data from other parts of speech. For instance, at present SemFrame bases all its frames on verb framesets, but some FrameNet frames list only adjectives as evoking lexical units. At the same time, potentially more can be done in associating verb synsets with frames: Only one-third of WordNet's verb synsets are now included in SemFrame's output. Some of those not now included evoke none of SemFrame's current frames, but some do and have not yet been recognized. Ways of establishing hierarchical and compositional relationships among frames should also be investigated.

The above suggestions for enhancing SemFrame notwithstanding, major progress in improving SemFrame awaits incorporation of corpus data. Relying on data from lexical resources has contributed to SemFrame's precision, but the data sparseness bottleneck that SemFrame faces is nonetheless real. On the basis of the lexical resource data used, verb synsets are related on average to only 5 nouns, many of which closely reflect the participant structure of the corresponding frame. However, it is not uncommon for specific elements of the participant structure to go unrepresented, and any nouns in the dataset that are not particularly reflective of the participant structure carry far too much weight amidst such a paucity of data.

In contrast, the number of nouns that co-occur with a verb in a corpus may be orders of magnitude greater.[13] But the nouns in a corpus are less likely to reflect closely the participant structure of the corresponding frame; many more nouns are thus likely to be needed. Furthermore, word sense disambiguation will be required to assign to a frame only those nouns corresponding to an appropriate sense of the verb.[14] We are optimistic, however, that the presence of additional corpus data will help fill in frame element gaps arising from the sparseness of lexical resource data and can also be used to help reduce the impact of nouns from lexical resource data that are not representative of a frame's participant structure.

Coupled with subject-specific resources, the analysis of corpus data may then lead to the development

[13] We are investigating two levels of noun-verb co-occurrence. The first counts co-occurrences of all nouns and verbs appearing within the same paragraph of newswire texts. The second counts only those nouns related to verbs as their subjects, direct objects, indirect objects, or as objects of prepositional phrases that modify the verb.

[14] We make the simplifying assumption that if a noun occurs with some reasonable percentage of the verbs within a frameset, the desired verb sense is in play.

of subject-specific frame inventories. Such inventories can in turn inform such knowledge-intensive applications as information retrieval, information extraction, and question answering.

Acknowledgements

This work has been supported in part by NSF ITR Grant IIS-0326553.

References

Eneko Agirre and German Rigau. A proposal for word sense disambiguation using conceptual distance. *1st International Conference on Recent Advances in NLP*.

Collin Baker, Jan Hajic, Martha Palmer, and Manfred Pinkal. 2004. Beyond syntax: Predicates, arguments, valency frames and linguistic annotations. Tutorial at *42nd Annual Meeting of the Association of Computational Linguistics*.

Christiane Fellbaum (Ed.) 1998. Introduction. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database.* MIT Press.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3): 245–288.

Rebecca Green, Bonnie J. Dorr, and Philip Resnik. 2004. Inducing frame semantic verb classes from WordNet and LDOCE. *42nd Annual Meeting of the Association of Computational Linguistics*.

Rebecca Green and Bonnie J. Dorr. 2004. Inducing a semantic frame lexicon from WordNet data. *Workshop on Text Meaning and Interpretation, 42nd Annual Meeting of the Association of Computational Linguistics*.

Erez Hartuv and Ron Shamir. 2000. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76:175–181.

Paul Kingsbury, Martha Palmer, and Mitch Marcus. 2002. Adding Semantic Annotation to the Penn Treebank. *Proceedings of the Human Language Technology Conference*.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 18(1):21–48.

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

Mechthild Stoer and Frank Wagner. 1997. A simple min-cut algorithm. *Journal of the ACM*, 44(4):585–591.

Nicola Stokes. 2004. Applications of lexical cohesion: Analysis in the topic detection and tracking domain. Ph.D. dissertation, National University of Ireland, Dublin.

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

A Sample Framesets

(a)
(stick_to stick_with follow) keep to
(comply follow abide_by) act in accordance with someone's rules, commands, or wishes
(b)
(sneer) smile contemptuously
(sneer) express through a scornful smile
(contemn despise scorn disdain) look down on with disdain
(c)
(muck) remove muck, clear away muck, as in a mine
(slime) cover or stain with slime
(clean make_clean) make clean by removing dirt, filth, or unwanted substances from
(dirty soil begrime grime colly bemire) make soiled, filthy, or dirty
(mire muck mud muck_up) soil with mud, muck, or mire
(d)
(federate federalize federalise) unite on a federal basis or band together as a league
(ally) become an ally or associate, as by a treaty or marriage
(confederate) form a confederation with; of nations
(divide split split_up separate dissever carve_up) separate into parts or portions
(unite unify) act in concert or unite in a common purpose or belief
(band_together confederate) form a group or unite
(e)
(fade melt) become less clearly visible or distinguishable; disappear gradually or seemingly
(get_down begin get start_out start set_about set_out commence) take the first step or steps in carrying out an action
(begin lead_off start commence) set in motion, cause to start
(end terminate) bring to an end or halt
(appear come_along) come into being or existence, or appear on the scene
(vanish disappear) cease to exist
(vanish disappear go_away) become invisible or unnoticeable
(begin start) have a beginning, in a temporal, spatial, or evaluative sense
(end stop finish terminate cease) have an end, in a temporal, spatial, or quantitative sense; either spatial or metaphorical

B Sample Frames

(a)
FRAME CONFORMITY (acting according to certain accepted standards):
- ATTRIBUTE (complaisance (a disposition or tendency to yield to the will of others)) [ ]
- COMMUNICATION (law (legal document setting forth rules governing a particular kind of activity)) [ ]
- PSYCH FEATURE (e.g., law (a rule or body of rules of conduct essential to or binding upon human society)) [ ]
- PERSON1/AGENT [ ]
- PERSON2/RECIPIENT OR PATIENT [ ]
- COMMUNICATION (advice (a proposal for an appropriate course of action)) [ ]
- ACT (e.g., accordance (the act of granting rights)) [ ]
(b)
FRAME CONTEMPT (open disrespect for a person or thing):
- COMMUNICATION (scorn (open disrespect for a person or thing)) [ ]
- PERSON1/AGENT [ ]
- PERSON2/RECIPIENT OR PATIENT [ ]
(c)
FRAME SOILING (the act of soiling something):
- ACTION (e.g., soiling (the act of soiling something)) [ ]
- STATE (e.g., soil (the state of being covered with unclean things)) [ ]
- CLEANER (the operator of dry cleaning establishment) [ ]
- CLEANER (someone whose occupation is cleaning) [ ]
(d)
FRAME CONFEDERATION (the act of forming an alliance or confederation):
- ACTION (e.g., division (the act or process of dividing)) [ ]
- SPLITTER (a taxonomist who classifies organisms into many groups on the basis of relatively minor characteristics) [ ]
- STATE (e.g., marriage (the state of being a married couple voluntarily joined for life (or until divorce))) [ ]
(e)
FRAME BEGINNING (the act of starting something):
- ACTION (e.g., beginning (the act of starting something)) [ ]
- COMMUNICATION (conclusion (the last section of a communication)) [ ]