Lexical Resource Integration across the Syntax-Semantics Interface

Rebecca Green†‡ and **Lisa Pearl†** and **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

Abstract

This paper examines extending a database of English verbs, grouped into syntactico-semantic classes, with WordNet senses. Probabilistic associations between θ-grids and WordNet verb frames, SEMCOR frequency data, and disambiguation based on an information-theoretic notion of semantic similarity are used. Mapping successes and failures are illustrated with *drop*.

1 Introduction

We are interested in mapping entries in a database of 4069 English verbs automatically to WordNet senses (Miller and Fellbaum, 1991), (Fellbaum, 1998) in order to integrate these lexical resources for multilingual applications such as machine translation and cross-language information retrieval. For example, the English verb *drop* has many potential translations in Spanish: *bajar*, *caerse*, *dejar*, *caer*, *derribar*, *disminuir*, *echar*, *hundir*, *soltar*, etc. Our database specifies a set of interpretations for the verb *drop*, differentiated by the context in which they appear in the source-language. Integration of these two lexical resources allows us to associate this interpretation with a set of WordNet senses; these, in turn, are used in choosing an appropriate verb in the target language.

Our work in lexical resource integration parallels the building of multilingual thesauri (Hudon, 2001), the mapping of dozens of medical vocabularies to MeSH (2000) within the Unified Medical Language System (UMLS, 2001), (Bodenreider and Bean, 2001), and work in ontology integration (Hovy, In press). As semantic resources (e.g., machine-readable dictionaries, thesauri, ontologies) begin to proliferate, we find that their underlying classificatory structures differ, making the establishment of equivalences across them anything but trivial. But as we are able to create such mappings, we both extend the power of individual resources and add to the larger research effort to generate standardized semantic resources, e.g., EAGLES.[1]

On the one hand, the verb database contains mostly syntactic information about its entries, with much of that information applying at the level of the classes used within the database. WordNet, on the other hand, is a significant source for information about semantic relationships, with much of that information applying at the "synset" level. Thus, by mapping entries in the database to their corresponding WordNet senses, the semantic potential of the verb database is extended significantly. At the same time, the fully mapped database becomes itself a data set in the larger effort to find commonalities across lexical resources.

2 Nature of the Resources

While it is commonly agreed in theory that words may have multiple senses, there is often little agreement in practice how many senses a given word has or whether word senses should be broadly or narrowly defined (Palmer, 2000). Detailed examination of the treatment of specific words in seemingly comparable dictionaries reveals that words are divided into senses in divergent ways (Fillmore and Atkins, 1992). Establishing equivalences across lexical resources under such circumstances is seldom a matter of generating one-to-one mappings. Indeed, mappings between lexical resources are not necessarily symmetrical; for example, when health experts mapped terms in various terminologies to the UMLS Metathesaurus and could not find an exact match, they opted for more general

[1] Information about the Expert Advisory Group on Language Engineering Standards (EAGLES) is available at http://www.ilc.pi.cnr.it/EAGLES/intro.html.

terms almost ten times as often as more specific ones (Bean, 2000).

In our lexical resource integration task, we have sought to identify sets of WordNet senses that best correspond to entries in the verb database and not vice versa. To understand the challenges involved, it is first necessary to compare the characteristics of the two resources.

2.1 Verb Database

Our database is a classification of 4069 English verbs, based initially on *English Verbs Classes and Alternations* (EVCA) (Levin, 1993) and extended through the splitting of some classes into subclasses and the addition of new classes. The resulting 491 classes (e.g., Roll Verbs, Group I: *drop, glide, roll, swing*) are referred to here as *Levin+ classes*. As verbs may be assigned to multiple Levin+ classes, the number of entries in the database is rather larger, viz., 9611.

Following the model of (Dorr and Olsen, 1997), each Levin+ class is associated with a *thematic grid* (henceforth abbreviated θ-grid), which summarizes a verb's syntactic behavior through specifying its predicate argument structure. For example, the Levin+ class 'Roll Verbs, Group I' is associated with the θ-grid `[theme goal]`, in which a theme and a goal are used (e.g., *The ball dropped to the ground*).

As (Levin, 1993) convincingly demonstrates, there is a correlation between a verb's syntactic behavior and its semantics. Thus, while the inclusion of a single verb in multiple Levin+ classes is grounded in syntactic behavior–specifically in its predicate argument structure (as captured in one or more corresponding θ-grids) and in permissible diathesis alternations–it may also be reasonably supposed that the multiple entries of a verb in the database represent different senses of the verb.

2.2 WordNet

WordNet 1.6 covers 10,319 verbs, organized into 12,127 synsets, representing 22,066 verb senses. Most of the verbs in our database (4056 of 4069) are also in WordNet;[2] these verbs have 12,561 senses in WordNet and belong to 8147 synsets. The ratio of verb senses to verbs is 3.10 for verbs in both WordNet and in the verb database; the ratio of verb senses to verbs for our database is 2.36. This indicates that WordNet uses more fine-grained word sense distinctions than the verb database. Moreover the basis on which the distinctions are made differ: syntactic behavior in the case of the verb database, semantic relationships in the case of WordNet.

In contrast to the syntactic emphasis of the verb database, WordNet gives mostly semantic information in its entries. For example, WordNet records semantic relations of several types between synsets. Using the semantically tagged Brown corpus files contained in the SEMCOR package, WordNet also indicates how frequently the various senses of a word are used, thus yielding the prior probability of a specific sense for any occurrence of a word. While information about the syntactic behavior of words has not been emphasized in WordNet, increasingly such information is being incorporated. Glosses indirectly indicate the predicate argument structure of verbs in a synset; example sentences and verb frames spell out the predicate argument structure more definitively. To some extent the verb frames–a set of 35 generic sentence frames, e.g., *Somebody ----s somebody something, Something ----s*–fill the same role as θ-grids. However, they are only partially comparable and thus cannot, on their own, support mapping verb database entries to WordNet senses.

It is worth noting that, although the two resources under consideration were constructed according to different principles, WordNet's relational organization captures some of the same information as decompositional theories of verb meaning, such as the one underlying EVCA (Fellbaum, 1998). Along these same lines, Dang et al. (1998) discuss a refinement of the EVCA class organization and its potential mapping to WordNet senses.

3 Data for Mapping between the Verb Database and WordNet

Since it is not possible to map directly between verb database entries and WordNet senses, we used 1791 entries that had been manually tagged with WordNet senses as training data to generate probabilistic associations between data from the two resources. For example, one of our measures captured the association between θ-grids and WordNet verb

[2] As we are mapping from entries in the verb classification to WordNet senses, the existence of verbs in WordNet but not in our database are of no significance.

frames, from the perspective of both individual θ-roles/verb frames and overall θ-grids/sets of verb frames. This will be referred to as a *syntactic similarity* measure. We also used a disambiguation algorithm (Resnik, 1999a)–based on an information-theoretic notion of semantic similarity (Resnik, 1999b)–which computes the confidence that specific WordNet senses hold, given the accompanying set of verbs in the same (Levin+) class. This will be referred to as a *semantic similarity* measure. We also used SEMCOR frequency data to establish the prior probability of specific WordNet senses.

Based on a handful of probabilistic associations between syntactic and semantic characteristics of the two resources, including the syntactic similarity measure set out above, as well as the information-theoretic semantic similarity measure, and SEMCOR frequency data, we investigated a number of voting schemes for mapping entries in the verb database to WordNet senses. The best results achieved 72% precision and 58% recall, versus a lower bound of 62% precision and 38% recall for most frequent WordNet sense, and an upper bound of 87% precision and 75% recall for human judgment. Further details of the mapping and its evaluation are available in (Green et al., 2001).

4 Case Study: *Drop*

In this section we consider mapping the verb database entries for *drop* to their corresponding WordNet senses; the examples are taken from the 'best results' voting scheme, with two aggregate voters, one based on the product of the half dozen measures indicated above, the other based on their weighted sum. The discussion will focus on the θ-grid/WordNet verb frame syntactic similarity measure, the Resnik semantic similarity measure, and SEMCOR frequency data as the most salient of those measures. The contribution made by the syntactic similarity measure to the mapping process reflects the degree to which the θ-grid data in the verb database and WordNet's verb frames capture the same syntactic behavior. The contribution made by the semantic similarity measure reflects the degree of compatibility between the semantics of the EVCA-based verb classes and WordNet's hierarchical structure.

There are 8 entries for *drop* in the verb database, outlined in Table 1; there are 19 senses of *drop* in WordNet, outlined in Table 2. We will examine 4 cases: (1) an appropriate WordNet sense correctly mapped; (2) an inappropriate WordNet sense correctly not mapped; (3) an appropriate WordNet sense incorrectly not mapped; and (4) an inappropriate WordNet sense incorrectly mapped.

The first case involves a WordNet sense (sense 3; "stock prices dropped") that our mapping process correctly indicates is an appropriate choice for a verb database entry (sense 3; "the prices dropped"). The sample sentences clearly indicate an exact match between the WordNet sense and the verb database entry. The WordNet sense is the third most frequently occurring sense of *drop* in SEMCOR, representing almost 12% of its uses. Thus prior probability does not promote this sense very strongly. However, both the syntactic and semantic similarity measures identified this as the most likely sense. The association between the verb database entry's θ-grid **[theme]** and the WordNet verb frame *Something* ____*s* is particularly strong; the fact that there is only one component in the θ-grid and only one verb frame for the WordNet sense helps strengthen that association. Likewise, the presence of verbs such as *appreciate, fluctuate, grow, mushroom* and *vary* in the same Levin+ class strongly point the semantic similarity measure to a WordNet sense in the change domain, where WordNet sense 3 occurs. The strength of the evidence with regard to both syntactic and semantic similarity easily overcome the weakness of the prior probability measure.

The second case involves a WordNet sense (sense 1; "don't drop the dishes") that our mapping process correctly indicates is an inappropriate choice for a verb database entry (sense 3 again; "the prices dropped"). (Surprisingly, both human coders rated WordNet sense 1 a good choice, despite the literal, transitive use of the WordNet sense versus the figurative, intransitive use of the verb database entry!) Over one-third of all occurrences of *drop* in SEMCOR represent WordNet sense 1; the mapping process will always consider this the most appropriate sense on the basis of prior probability alone. However, the semantic similarity measure for this sense rated this motion sense of *drop* at a zero level of confidence, which pretty

#	**Levin+ class**	**Example sentence**	**Required θ roles**	**Optional θ roles**
1	Drop	She dropped the book to the ground.	agent theme goal	
2	Putting down	I dropped the stone down to the ground.	agent theme	mod-loc (down) source goal
3	Calibratable changes of state	The prices dropped.	theme	
4	Meander (to/from)	The river drops from the lake to the sea.	theme source (from) goal (to)	
5	Meander (path)	The river drops through the valley.	theme goal	
6	Roll 1	The ball dropped.	theme	
7	Roll 2	The ball dropped into the room	theme	source goal
8	Roll down	The stone dropped down into the ground.	theme particle (down)	source goal

Table 1: Senses of *drop* in Verb Database

much scotches the possibility of its being assigned. The syntactic similarity measure looked favorably on this sense from the perspective of correlation between individual components of the θ-grid **[agent theme]** and the WordNet verb frame, since the verb frame *Somebody ____s something* has a fairly strong association with the presence of a theme, but the verb frame combination (also including *Somebody ____s somebody*) has only a weak association with the overall θ-grid.

Having looked at two successes, we turn now to two failures. The third case involves a WordNet sense (sense 1; "don't drop the dishes") that should have been assigned to a verb database entry (sense 1; "she dropped the book to the ground"), but was not. As noted above, this WordNet entry is the most frequently occurring sense of *drop* in SEMCOR and thus is favored by the prior probability measure. The training data included no instances of the θ-grid for this Levin+ class with the set of verb frames for this WordNet sense, although the strength of association between individual components of the θ-grid and individual WordNet verb frames was fairly strong. Uncharacteristically, the semantic similarity value for this WordNet sense is quite low. The reason for this turns out to be that *drop* is the only verb in this Levin+ class. Thus the semantic similarity measure has no evidence for distinguishing among WordNet senses and assigns them all an equal, but insignificant, confidence level. In this case, data sparsity stands in the way of correct sense assignment. It is worth noting, however, that the available evidence promotes the correct sense.

The final case involves a WordNet sense (sense 6; "drop a hint") assigned to a verb database entry (sense 1; "she dropped the book to the ground) that should not have been assigned. Since we are looking at the same verb database entry as in the previous example, it will be instructive to contrast the two WordNet senses. As WordNet senses are listed in order of SEMCOR frequency, sense 6 occurs rather less often than sense 1. As explained before, the semantic similarity measure is unable to distinguish between WordNet senses when the Levin+ class has only one member, as in this case. What drives the different assignment here is the absence of a verb frame from sense 6: Sense 1 allows both (*Somebody ____s something*) and (*Somebody ____s somebody*), while sense 6 allows only (*Somebody ____s something*). The

#	**WordNet gloss**	**Verb frames**	SEMCOR **count**
1	let fall to the ground; "don't drop the dishes"	Somebody ____s something Somebody ____s somebody	36
2	fall vertically; "the bombs are dropping on enemy targets"	Something ____s Somebody ____s	21
3	go down in value; "stock prices dropped"	Something ____s	12
4	fall or drop to a lower place or level; "he sank to his knees"	Something ____s Somebody ____s	7
5	terminate an association with; "drop him from the Republican ticket"	Somebody ____s somebody Something ____s somebody	6
6	utter casually; "drop a hint"	Somebody ____s something	6
7	stop pursuing or acting; "drop a lawsuit"	Somebody ____s something	5
8	leave or unload, esp. of passengers or cargo	Somebody ____s something Somebody ____s somebody Somebody ____s somebody PP Somebody ____s something PP	3
9	as of trees or people	Somebody ____s something Somebody ____s somebody Something ____s something	2
10	of games, in sports; "the Giants dropped all 11 of their first 13"	Somebody ____s something	2
11	pay out; "spend money"	Somebody ____s something Somebody ____s something on somebody	1
12	lower the pitch of (musical notes)	Somebody ____s something	1
13	hang freely; "the light dropped for the ceiling"	Something ____s Something is ____ing PP	0
14	stop associating with; "they dropped her after she had a child out of wedlock"	Somebody ____s somebody	0
15	get rid of; "he shed his image as a pushy boss"	Somebody ____s something Something ____s something	0
16	leave undone or leave out; "how could I miss that typo?"	Somebody ____s something Somebody ____s somebody Somebody ____s to INFINITIVE	0
17	change from one level to another; "she dropped into Army jargon"	Something is ____ing PP Somebody ____s PP	0
18	grow worse; "her condition deteriorated"	Something ____s Somebody ____s	0
19	give birth, used for animals; "the cow dropped her calf this morning"	Something ____s something	0

Table 2: Senses of *drop* in WordNet

association of (*Somebody ____s something*) with each of the components of the **[agent theme goal]** θ-grid is much stronger in the training data than is true for (*Somebody ____s somebody*); moreover, the single verb frame for sense 6 has a much stronger association with the whole θ-grid than does the verb frame pair for sense 1. Data sparseness is again a problem, as is the difference between the classification of syntactic patterns in the two resources.

5 Conclusion

Semantic data in WordNet–SEMCOR frequency data and the hierarchical structure of WordNet–combine with associations between θ-grid information and WordNet verb frames to extend a verb classification based on syntactico-semantic classes with WordNet senses. Data sparseness is a major factor in at least some mapping failures. At the same time, syntax-based measures contribute less to mapping successes than do the semantic similarity and word sense frequency measures. This suggests a larger degree of compatibility between the semantics of Levin+ verb classes and the WordNet relational structure than between the systems used in the two resources to reflect verbs' syntactic behavior.

Acknowledgments

The authors are supported, in part, by PFF/PECASE Award IRI-9629108, DOD Contract MDA904-96-C-1250, DARPA/ITO Contracts N66001-97-C-8540 and N66001-00-28910, and a National Science Foundation Graduate Research Fellowship.

References

Carol A. Bean. 2000. Mapping Down: Semantic and Structural Relationships in User-Designated Broader-Narrower Term Pairs. In C. Beghtol, L. C. Howarth, and N. J. Williamson, editors, *Dynamism and Stability in Knowledge Organization*, pages 301–305. Ergon Verlag, Wurzburg.

Olivier Bodenreider and Carol A. Bean. 2001. Relationships among Knowledge Structures: Vocabulary Integration within a Subject Domain. In C.A. Bean and R. Green, editors, *Relationships in the Organization of Knowledge*, pages 81–98. Kluwer, Dordrecht.

Hoa Trang Dang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig. 1998. Investigating Regular Sense Extensions Based on Intersective Levin Classes. In *ACL/COLING 98, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 293–299, Montreal, Canada, August 10-14.

Bonnie J. Dorr and Mari Broman Olsen. 1997. Deriving Verbal and Compositional Lexical Aspect for NLP Applications. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 151–158, Madrid, Spain, July 7-12.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Charles J. Fillmore and Beryl T. Atkins. 1992. Toward a Frame-Based Lexicon: The Semantics of RISK and its Neighbors. In A. Lehrer and E. F. Kittay, editors, *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, pages 75–102. Lawrence Erlbaum, Hillsdale, NJ.

Rebecca Green, Lisa Pearl, Bonnie J. Dorr, and Philip Resnik. 2001. Mapping Lexical Entries in a Verbs Database to WordNet Senses. Technical Report CS-TR-4231, UMIACS-TR-2001-19, LAMP-TR-068, University of Maryland.

Eduard Hovy. In press. Comparing Sets of Semantic Relations in Ontologies. In R. Green, C.A. Bean, and S. Myaeng, editors, *The Semantics of Relationships: An Interdisciplinary Perspective*. Book manuscript submitted for review.

Michele Hudon. 2001. Relationships in Multilingual Thesauri. In C.A. Bean and R. Green, editors, *Relationships in the Organization of Knowledge*, pages 67–80. Kluwer, Dordrecht.

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

MeSH. 2000. Medical Subject Headings. National Library of Medicine.

George A. Miller and Christiane Fellbaum. 1991. Semantic Networks of English. In Beth Levin and Steven Pinker, editors, *Lexical and Conceptual Semantics*, pages 197–229. Elsevier Science Publishers, B.V., Amsterdam, The Netherlands.

Martha Palmer. 2000. Consistent Criteria for Sense Distinctions. *Computers and the Humanities*, 34:217–222.

Philip Resnik. 1999a. Disambiguating noun groupings with respect to wordnet senses. In S. Armstrong, K. Church, P. Isabelle, E. Tzoukermann S. Manzi, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, pages 77–98. Kluwer Academic, Dordrecht.

Philip Resnik. 1999b. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, (11):95–130.

UMLS. 2001. Unified Medical Language System. National Library of Medicine.