The VerbCorner Project: Toward an Empirically-Based Semantic Decomposition of Verbs

Joshua K. Hartshorne
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139, USA
jkhartshorne@gmail.com

Claire Bonial, Martha Palmer
Department of Linguistics
University of Colorado at Boulder
Hellems 290, 295 UCB
Boulder, CO 80309, USA
{CBonial, MPalmer}@colorado.edu

Abstract

This research describes efforts to use crowdsourcing to improve the validity of the semantic predicates in VerbNet, a lexicon of about 6300 English verbs. The current semantic predicates can be thought of semantic primitives, into which the concepts denoted by a verb can be decomposed. For example, the verb *spray* (of the Spray class), involves the predicates MOTION, NOT, and LOCATION, where the event can be decomposed into an AGENT causing a THEME that was originally not in a particular location to now be in that location. Although VerbNet's predicates are theoretically well-motivated, systematic empirical data is scarce. This paper describes a recently-launched attempt to address this issue with a series of human judgment tasks, posed to subjects in the form of games.

1 Introduction

One key application of Natural Language Processing (NLP) is meaning extraction. Of particular importance is propositional meaning: To understand "Jessica sprayed paint on the wall," it is not enough to know who Jessica is, what paint is, and where the wall is, but that, by the end of the event, some quantity of paint that was not previously on the wall now is. One must extract not only meanings for individual words but also the relations between them.

One option is to learn these relations in a largely bottom-up, data-driven fashion (Chklovski and Pantel, 2004; Poon and Domingos, 2009). For instance, Poon and Domingos (2009) first extracts dependency trees, converts those into quasi-logical form, recursively induces lambda expressions from them, and uses clustering to derive progressively abstract knowledge.

An alternative is to take a human-inspired approach, mapping the linguistic input onto the kinds of representations that linguistic and psychological research suggests are the representations employed by humans. While the exact characterization of meaning (and by extension, thought) remains an area of active research in the cognitive sciences (Margolis and Laurence, 1999), decades of research in linguistics and psychology suggests that much of the meaning of a sentence – as well as its syntactic structure – can be accounted for by invoking a small number of highly abstract semantic features (usually represented as predicates), such as causation, agency, basic topological relations, and directed motion (Ambridge et al., 2013; Croft, 2012; Jackendoff, 1990; Levin and Rappaport Hovav, 2005; Pesetsky, 1995; Pinker, 1989). For instance, a given verb can appear in some syntactic frames (*Sally broke the vase. Sally broke the vase with the hammer. The vase broke.*) and not others (**Sally broke the vase to the floor. *Sally broke John the vase.*). When verbs are classified according to the syntactic frames they can appear in, most if not all the verbs in a class involve the same set of abstract semantic features.[1]

Interestingly, roughly these same features (causation, etc.) have been singled out by developmental psychologists as part of "core knowledge" – a set of early-learned or perhaps innate concepts upon which

[1] Whether all verbs in a class share the same abstract predicates or merely most is an area of active research (Levin and Rappaport Hovav, 2005).

the rest of cognition is built (Spelke and Kinzler, 2007). Thus these semantic features/predicates may be not only crucial to describing linguistic meaning but may be central organizing principles for a human's (reasonably successful) thinking about and conceptualization of the world. As such, they provide a potentially rewarding target for NLP.

2 VerbNet

2.1 Overview and Structure

Perhaps the most comprehensive implementation of this approach appears in VerbNet (Kipper et al., 2008; based on Levin, 1993). VerbNet classifies verbs based on the syntactic frames they can appear in, providing a semantic description of each frame for each class. An example entry is shown below:

Syntactic Frame NP V NP PP.DESTINATION
Example Jessica sprayed the wall.
Syntax AGENT V THEME {+LOC|+DEST_CONF} DESTINATION
Semantics MOTION(DURING(E), THEME)
NOT(PREP(START(E), THEME, DESTINATION))
PREP(END(E), THEME, DESTINATION)
CAUSE(AGENT, E)

The "Syntactic Frame" provides a flat syntactic parse. "Syntax" provides semantic role labels for each of the NPs and PPs, which are invoked in "Semantics". VerbNet decomposes the semantics of this sentence into four separate predicates: 1) the THEME (the paint) moves doing the event E; 2) at the start of the event E, the THEME (the paint) is not at the DESTINATION (on the wall), whereas 3) at the end of the event E, the THEME (the paint) is at the DESTINATION (on the wall), and; 4) the event is caused by the AGENT (Sally). Note that this captures only the core aspects of semantics shared by all verbs in the class; differences between verbs in the same class (e.g., *spray* vs. *splash*) are omitted.

Importantly, the semantics of the sentence is dependent on both the matrix verb (paint) and the syntactic frame. Famously, when inserted in the slightly different frame NP V NP.DESTINATION PP.THEME – "Sally sprayed the wall with paint" – "spray" entails that destination (the wall) is now fully painted, an entailment that does not follow in the example above (Pinker, 1989).

2.2 Uses and Limitations

VerbNet has been used in a variety of NLP applications, such as semantic role labeling (Swier and Stevenson, 2004), inferencing (Zaenen et al., 2008), verb classification (Joanis et al., 2008), and information extraction (Maynard, Funk, and Peters, 2009).

While such applications have been successful thus far, an important constraint on how well VerbNet-based NLP applications can be expected to perform is the accuracy of the semantics encoded in VerbNet. Here, several issues arise. Leaving aside miscategorized verbs and other inaccuracies, as noted above VerbNet assumes that all verbs in the same class share the same core predicates, which may or may not be empirically justified. Given the number of semantic predicates (146),[2] verb entries (6580), and unique verb lemmas (6284) it is not feasible for a single research team to check, particularly since after a certain number of verbs, intuitions become less clear. In any case, it may not be ideal to rely solely on the intuitions of invested researchers, whose intuitions about subtle judgments may be clouded by theoretical commitments (Gibson and Federenko, 2013); the only way to ensure this is not the case is through independent validation. Unfortunately, of the 280 verb classes in VerbNet, this has been done for only a few (cf Ambridge et al., 2013).

3 VerbCorner

The VerbCorner project was designed to address these issues by crowd-sourcing the semantic judgments online (gameswithwords.org/VerbCorner/). Several previous projects have successfully crowd-sourced linguistic annotations, such as Phrase Detectives, where volunteers have contributed 2.5 million judgments on anaphoric relations (Poesio et al., 2012). Below, we outline the VerbCorner project and describe one specific annotation task in detail.

3.1 Developing Semantic Annotation Tasks

Collecting accurate judgments on subtle questions from naive participants with limited metalinguistic

[2] Note that these vary in applicability from those specific to a small number of verbs (CHARACTERIZE, CONSPIRE) to those frequently invoked (BEGIN, EXIST).

skills is difficult. Rare is the non-linguist who can immediately answer the question, "Does the verb 'throw,' when used transitively, entail a change of location on the part of its THEME?" Thus, we began by developing tasks that isolate semantic features in a way accessible to untrained annotators.

We converted the metalinguistic judgments ("Does this verb entail this abstract predicate?") into real-world problems, which previous research suggests should be easier (Cosmides and Tooby, 1992). Each judgment tasks involved a fanciful backstory. For instance, in "Simon Says Freeze", a task designed to elicit judgments about movement, the Galactic Overlord (Simon) decrees "Galactic Stay Where You Are Day," during which nobody is allowed to move from their current location. Participants read descriptions of events and decide whether anyone violated the rule. In "Explode on Contact", designed to elicit judgments about physical contact, objects and people explode when they touch one another. The participant reads descriptions of events and decides whether anything has exploded.[3]

Each task was piloted until inter-coder reliability was acceptably high and the modal response nearly always corresponded with researcher intuitions. As such, these tasks cannot be used to establish whether researcher intuitions for the pilot stimuli are correct (this would be circular); however, there is no guarantee that agreement with the researcher will generalize to new items (the pilot stimuli cover a trivial proportion of all verbs in VerbNet).

3.2 Crowd-sourcing Semantic Judgments

The pilot experiments showed that it is possible to elicit reliable semantic judgments corresponding to VerbNet predicates from naive participants (see section 3.3). At the project website, volunteers choose one of the tasks from a list and begin tagging sentences. The sentences are sampled smartly, avoiding sentences already tagged by that volunteer and biased in favor of of the sentences with the fewest judgments so far. Rather than assessing annotator quality through gold standard trials with known answers (which wastes data – the answers to these trials are known), approximately 150 sentences were chosen to be "over-sampled." As the volunteer tags sentences, approximately one out of every five are from this over-sampled set until that volunteer has tagged all of them. This guarantees that any given volunteer will have tried some sentences targeted by many other volunteers, allowing inter-annotator agreement to be used to assess annotator quality.

Following the example of Zooniverse (zooniverse.org), a popular "Citizen Science" platform, volunteers are encouraged but required to register (requiring registration prior to seeing the tasks was found to be a significant barrier to entry). Registration allows collecting linguistic and educational background from the volunteer, and also makes it possible to track the same volunteer across sessions.

Multiple gamification elements were incorporated into VerbCorner in order to recruit and motivate volunteers. Each task has a leaderboard, where the volunteer can see his/her rank out of all volunteers in terms of number of contributions made. In addition, there is a general leaderboard, which sums across tasks. Volunteers can earn badges, displayed on their homepage, for answering certain numbers of questions in each task. Finally, at random intervals bonus points are awarded, with the explanation for the bonus points tailored to the task's backstory.

VerbCorner was launched on May 21, 2013. After six weeks, 555 volunteers had provided at least one annotation, for a total of 39,274 annotations, demonstrating the feasibility of collecting large numbers of annotations through this method.

3.3 Case Study: Equilibrium

"Equilibrium" was designed to elicit judgments about application of force, frequently argued to be a core semantic feature in the sense discussed above (Pinker, 1989). The backstory involves the "Zen Dimension," in which nobody is allowed to exert force on anything else. The participant reads descriptions of events (*Sally sprayed paint onto the wall*) and decides whether they would be allowable in the Zen Dimension – and, in particular, which participants in the event are illegally applying force.

In order to minimize unwanted effects of world

[3]Note that each task is designed to elicit judgments about entailments – things that must be true rather than are merely likely to be true. If John greeted Bill, they might have come into contact (e.g., by shaking hands), but perhaps they did not. Previous work suggests that it is entailments that matter, particularly for explaining the syntactic behavior of verbs (Levin and Rappaport Hovav, 2005)

knowledge, the verb's arguments are replaced with nonsense words or randomly chosen proper names (*Sally sprayed the dax onto the blicket*). In the context of the story, this is explained as necessary anonymization: You are a government official determining whether certain activities are allowable, and ensuring anonymity is an important safeguard against favoritism and corruption. An alternative wouod be to use multiple different content words, randomly chosen for each annotator. However, this greatly increases the number of annotators needed and quickly becomes infeasible.

3.3.1 Pilot Results

The task was piloted on 138 sentences, which comprised all possible syntactic frames for three verbs from each of five verb classes in VerbNet. After two rounds of piloting (between the first and second, wording in the backstory was adjusted for clarity based on pilot subject feedback and results), Kripp's alpha reached .76 for 8 annotators, which represents a reasonably high level of inter-annotator agreement. Importantly, the modal response matched the intuitions of the researchers in 137 of 138 cases.[4]

3.3.2 Preliminary VerbCorner Results

"Equillibrium" was one of the first tasks posted on VerbCorner, with data currently being collected on 12 of the 280 VerbNet classes, for a total of 5,171 sentences. As of writing, 414 users have submitted 14,294 judgments. Individual annotators annotated anywhere from 1 to 195 sentences (mean=8, median=4). While most sentences have relatively few judgments, each of the 194 over-sampled sentences has between 15 and 20 judgments.[5]

Comparing the modal response with the researchers' intuitions resulted in a match for 184 of 194 sentences. In general, where the modal response did not match researcher intuitions, the modal response was itself not popular, comprising an average of 53% of responses, compared with an average of 77% where the modal response matched researcher intuitions. Thus, these appear to be cases of disagreement, either because the correct intuition requires more work to obtain or because of differences across idiolects (at the moment, there is no obvious pattern as to which sentences caused difficulty, but the sample size is small). Thus, follow-up investigation of sentences with little inter-coder agreement may be warranted.

4 Conclusion and Future Work

Data-collection is ongoing. VerbNet identifies approximately 150 different semantic predicates. Annotating every verb in each of its syntactic frames for each semantic predicate would take many millions of judgments. However, most of the semantic predicates employed in VerbNet are very narrow in scope and only apply to a few classes. Thus, we have begun with broad predicates that are thought to apply to many verbs and are adding progressively narrower predicates as work progresses. At the current rate, we should complete annotation for the half-dozen most frequent semantic predicates in the space of a year.

Future work will explore using an individual annotator's history across trials to weight that user's contributions, something that VerbCorner was specifically designed to allow (see above). How to assess annotator quality without gold standard data is an active area of research (Passonneau and Carpenter, 2013; Rzhetsky, Shatkay and Wilbur, 2009; Whitehill et al., 2009). For instance, Whitehill and colleagues (2009) provide an algorithm for jointly estimating both annotator quality and annotation difficulty (including the latter is important because some annotators will have low agreement with others due to their poor luck in being assigned difficult-to-annotate sentences). This algorithm is shown to outperform using the modal response.

Note that this necessarily biases against annotators with few responses. In our case study above, excluding annotators who contributed small numbers of annotations led to progressively worse match to researcher intuition, suggesting that the loss in data

[4]The remaining case was "The crose smashed sondily." for which four pilot subjects thought involved the crose applying force – matching researcher intuition – and four thought did not involve any application of force, perhaps interpreting the sentence was a passive.

[5]These are the same 15 verbs used in the piloting. The number of sentences is larger in order to test a wider range of possible arguments. In particular, wherever appropriate, separate sentences were constructed using animate and inanimate arguments. Compare *Sally sprayed the dax onto Mary* and *Sally sprayed the dax onto the blicket*.

caused by excluding these annotations may not be worth the increased confidence in annotation quality. Future research will be needed to assess this trade-off.

The above work shows the feasibility of crowd-sourcing VerbNet semantic entailments, as has been shown for a handful of other linguistic judgments (Artignan, Hascoet and Lafourcade, 2009; Poesio et al., 2012; Venhuizen et al., 2013). There are many domains in which gold standard human judgments are scarce; crowd-sourcing has considerable potential at addressing this need.

References

B. Ambridge, J. M. Pine, C. F. Rowland, F. Chang, and A. Bidgood. 2013. The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. *Wiley Interdisciplinary Reviews: Cognitive Science*. 4:47-62.

G. Artignan, M. Hascoet, and M. Lafourcade. 2009. Mutliscale visual analysis of lexical networks. *Proceedings of the 13th International Conference on Information Visualisation*. Barcelona, Spain.

T. Chklovski and P. Pantel. 2004. VerbOcean: Mining the Web for fine-grained semantic relations. *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*. Barcelona, Spain.

L. Cosmides and J. Tooby. 1992. Cognitive adaptations for social exchange. in *The Adapted Mind*. (J. Barkow, L. Cosmides, and J. Tooby, Eds.) Oxford University Press, Oxford, UK.

W. Croft. 2012. *Verbs: Aspect and Argument Structure*. Oxford University Press, Oxford, UK.

D. R. Dowty. 1991. Thematic proto-roles and argument selection. *Language*. 67:547-619.

E. Gibson and E. Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*. 28(1-2):88–124.

R. Jackendoff. 1990. *Semantic Structures*. The MIT Press, Cambridge, MA.

E. Joanis, S. Stevenson, and D. James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*. 14(3):337-367.

K. Kipper, A. Korhonen, N. Ryant and M. Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation Journal*, 42:21–40

E. Margolis and S. Laurence 1999. *Concepts: Core Readings*. The MIT Press, Cambridge, MA.

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

B. Levin and M. Rappaport Hovav. 2005. *Argument Realization*. Cambridge University Press, Cambridge, UK.

D. Maynard, A. Funk, and W. Peters. 2009. Using lexico-syntactic ontology design patterns for ontology creation and population. *Proceedings of Workshop on Ontology Patterns (WOP 2009)*. Washington, DC

R. J. Passonneau and B. Carpenter 2013. The benefits of a model of annotation. 7th Linguistic Annotation Workshop and Interoperability with Discourse. Sofia, Bulgaria.

D. Pesetsky. 1995. *Zero Syntax: Experiencers and Cascades*. The MIT Press, Cambridge, MA.

S. Pinker. 1989. *Learnability and Cognition*. The MIT Press, Cambridge, MA.

M. Poesio, J. Camberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. 2012. The Phrase Detective Multilingual Corpus, Release 0.1. *Proceedings of the Collaborative Resource Development and Delivery Workshop*. Istanbul, Turkey

H. Poon and P. Domingos. 2009. Unsupervised semantic parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore.

A. Rzhetsky, H. Shatkay, and W. J. Wilbur. 2009. How to get the most out of your curation effort. *PLoS Computational Biology*, 5(5):1–13.

E. S. Spelke and K. D. Kinzler. 2007. Core knowledge. *Developmental Science*, 10(1):89–96.

R. Swier and S. Stevenson. 2004. Unsupervised semantic role labeling. *Proceedings of the Generative Lexicon Conference, GenLex-09*. Pisa, Italy.

N. Venhuizen, V. Basile, K. Evang, and J. Bos. 2013. Gamification for word sense labeling. *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*. Potsdam, Germany

J. Whitehill, P. Ruvolo, T. F. Wu, J. Bergsma. and J. Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems, 22*. Vancouver, Canada

A. Zaenen, C. Condoravdi, and D G. Bobrow. 2008. The encoding of lexical implications in VN. *Proceedings of LREC 2008*. Morocco