Expanding Verb Coverage in Cyc With VerbNet

Clifton J. McFate
Northwestern University
Evanston, Il. USA.
c-mcfate@northwestern.edu

Abstract

A robust dictionary of semantic frames is an essential element of natural language understanding systems that use ontologies. However, creating lexical resources that accurately capture semantic representations en masse is a persistent problem. Where the sheer amount of content makes hand creation inefficient, computerized approaches often suffer from over generality and difficulty with sense disambiguation. This paper describes a semi-automatic method to create verb semantic frames in the Cyc ontology by converting the information contained in VerbNet into a Cyc usable format. This method captures the differences in meaning between types of verbs, and uses existing connections between WordNet, VerbNet, and Cyc to specify distinctions between individual verbs when available. This method provides 27,909 frames to OpenCyc which currently has none and can be used to extend ResearchCyc as well. We show that these frames lead to a 20% increase in sample sentences parsed over the Research Cyc verb lexicon.

1 Introduction

The Cyc[1] knowledge base represents general purpose knowledge across a vast array of domains. Low level event and individual facts are contained in larger definitional hierarchical representations and contextualized through microtheories (Matuszek *et al*, 2006). Higher order predicates built into Cyc's formal language, CycL, allow efficient inferencing about context and meta-language reasoning above and beyond first-order logic rules (Ramachandran *et al*, 2005).

Because of the expressiveness and size of the ontology, Cyc has been used in NL applications including word sense disambiguation and rule acquisition by reading (Curtis, Cabral, & Baxter, 2006; Curtis *et al*, 2009). Such applications use NL-to-Cycl parsers which use Cyc semantic frames to convert natural language into Cyc representations. These frames represent sentence content through a set of propositional logic assertions that first reify the sentence in terms of a real world event and then define the semantic relationships between the elements of the sentence, as described later. Because these parsers require semantic frames to represent sentence content, existing parsers are limited due to Cyc's limited coverage (Curtis *et al*, 2009). The goal is to increase this coverage by automatically translating the class frames in VerbNet into individual verb templates.

2 Previous Work

The Cyc knowledge base is continuously expanding and much work has been done on automatic fact acquisition as well as merging ontologies. However, the semantic frames remain mostly hand-made in ResearchCyc[2] and non-existent in the open-license OpenCyc[3]. Translating VerbNet frames into Cyc will expand the natural language capabilities of both.

There has been previous research on mapping existing Cyc templates to VerbNet, but thus far these approaches have not created new templates to address Cyc's lapses in coverage. One such attempt, King and Crouch's (2005) unified lexicon, compiled many lexical resources into a unified representation. While this research created a valuable resource, it did not extend the existing Cyc coverage. Of the 45, 704 entries in the UL only 3,544 have Cyc entries (King & Crouch, 2005).

Correspondences between a few VerbNet frames and ResearchCyc templates have also been mapped out through the VxC VerbNet Cyc

---

[1] http://www.opencyc.org/cyc

[2] http://research.cyc.com
[3] http://opencyc.org

Mapper (Trumbo 2006). These mappings became a standard that we later used to evaluate the quality of our created frames.

A notable exception to the hand-made paradigm is Curtis *et al*'s (2009) TextLearner which uses rules and existing semantic frames to handle novel sentence structures. Given an existing template that fits some of the syntactic constraints of the sentence, TextLearner will attempt to create a new frame by suggesting a predicate that fits the missing part. Often these are general underspecified predicates, but TextLearner is able to use common sense reasoning and existing facts to find better matches (Curtis *et al*, 2009).

While TextLearner improves its performance with time, it is not an attempt to create new frames on a large scale. Creating generalized frames based on verb classes will increase the depth of the Cyc Lexicon quickly. Furthermore, automatic processes like those in TextLearner could be used to make individual verb semantic frames more specific.

3   VerbNet

VerbNet is an extension of Levin's (1993) verb classes that uses the class structure to apply general syntactic frames to member verbs that have those syntactic uses and similar semantic meanings (Kipper *et al*, 2000). The current version has been expanded to include class distinctions not included in Levin's original proposal (Kipper *et al*, 2006).

VerbNet is an appealing lexical resource for this task because it represents semantic meaning as the union of both syntactic structure and semantic predicates. VerbNet uses Lexicalized Tree Adjoining Grammar to generate the syntactic frames. The syntactic roles in the frame are appended with general thematic roles that fill arguments of semantic predicates. Each event is broken down into a tripartite structure as described by Moens & Steedman (1988) and uses a time modifier for each predicate to indicate when specific predicates occur in the event. This allows for a dynamic representation of change over an event. (Kipper *et* al, 2000).

This approach is transferable to Cyc's semantic templates in which syntactic slots fill predicate arguments in the context of a specific syntactic frame. Both also have extensive connections to WordNet2.0, an electronic edition of Miller's (1985) WordNet (Fellbaum, 1998).

4   Method

The general method for creating semantic templates in Cyc requires creating Verb Class Frames and then using Cyc predicates and heuristic rules to create individual frames for each member verb.

4.1   OpenCyc

The existing semantic templates are accessible through the ResearchCyc KB. However, for the purposes of this study the OpenCyc KB was used. The OpenCyc KB is an open source version of ResearchCyc that contains much of the definitional information and higher order predicates, but has had much of the lower level specific facts and the entire word lexicon removed (Matuszek *et al*, 2006). However, the assertions generated by this method are fully usable in ResearchCyc. OpenCyc was used so as to minimize the effect of existing semantic frames on new frame creation. Since OpenCyc and VerbNet are open-licensed, our translation provides an open-license extension to OpenCyc to support its use in natural language research.

4.2   Knowledge Representation

The primary difficulty with integrating VerbNet frames into Cyc was overcoming differences in knowledge representation. Cyc semantic templates reify events as an instance of a collection of events. The arguments correspond to syntactic roles. The following is a semantic template for a ditransitive use of the word *give* from ResearchCyc.

```
(verbSemTrans Give-TheWord 0
(PPCompFrameFn
 DitransitivePPFrameType To-TheWord)
(and
(isa ACTION GivingSomething)

(objectGiven ACTION OBJECT)
(giver ACTION SUBJECT)
(givee ACTION OBLIQUE-OBJECT)))
```

However, VerbNet uses semantic predicates that describe relationships between two thematic roles. The following is a frame for the VerbNet class *Give* as presented in the Unified Verb Index[4].

```
NP V NP PP.recipient
example
```

---

[4] http://verbs.colorado.edu/verb-index/

```
"They lent a bicycle to me."
syntax
 Agent V Theme {to} Recipient
semantics
 -has_possession(start(E), Agent,
 Theme)
 -has_possession(end(E), Recipient,
 Theme)
 -transfer(during(E), Theme)
 -cause(Agent, E)
```

The predicate `has_possession` occurs twice, at the beginning and end of the event. In one case the `Agent` has possession and in the second the `Recipient` does. Both refer to the `Theme` which is being transferred.

In Cyc the `hasPossession` relationship to `Agent` and `Recipient` is represented with the predicates `giver` and `givee`. The subject and oblique-object of the sentence fill those arguments, and the actual change of possession is represented by the collection of events `GivingSomething`. The VerbNet `Theme` is the object in `objectGiven`. Thus an individual VerbNet semantic predicate often has a many-to-one mapping with Cyc predicates.

4.3 Predicates

To account for representation differences, a single Cyc predicate was mapped to a unique combination of Verbnet predicate and thematic role (ie. `Has_Possession Agent at start(E) => givee`). 56 of these mappings were done by hand. Though far from exhaustive, these hand mappings represent many frequently used predicates in VerbNet. The hand mapping was done by looking at the uses of the predicate across different classes.

Because the mappings were not exhaustive, a safety net automatically catches predicates that haven't been mapped. The VerbNet predicates `Cause` and `InReactionTo` corresponded to the Cyc predicates `performedBy, doneBy`, and `causes-Underspecified`. These predicates were selected whenever the VerbNet predicates occurred with a theme role that was the subject of the sentence. The more specific `performedBy` was selected in cases where the frame's temporal structure suggested a result. The predicate `doneBy` was selected in other cases. The `causes-Underspecified` predicate was used in frames whose time modifiers suggested that they were continuous states. The predicates `patientGeneric` and `patientGeneric-Direct` were used when a

predicate was not found for a required object or oblique object.

Some Cyc templates don't have predicates that reference the event. For example, the verb *touch* can be efficiently represented with the relation `(objectsInContact :SUBJECT :OBJECT)`. Situations like this were hand assigned.

4.4 Collections

In Cyc, concepts are represented by collections. Inheritance between collections is specified by the `genls` relationship, which can be viewed as subset. Most verb frames have an associated collection of events of which each use is an instance. The associated collection of the class frame templates was automatically selected using the common link that both resources share with WordNet (Fellbaum, 1998). To do this, the WordNet synsets of the member verbs for a class were matched with their Cyc-WordNet `synonymousExternalConcept` assertion. The Cyc representation became a denoted collection. The most general collection out of the list of viable collections was chosen as the general class frame collection. The number of `genls` links to a collection was used as a proxy for generality. In the case of a tie the first was chosen.

While the most general collection was used for the class semantic frame, at the level of individual verb frames the specific synset denoted collection was substituted for the more general one when applicable. Verbs with multiple meanings across classes were given a unique index number for each sense. However, within a given class each word only received one denotation. The general class level collection was used in cases where no Cyc-WordNet-VerbNet link existed. If no verb had a synset in Cyc, the general collection `Situation` was used.

4.5 Subcategorization Frames

Each syntactic frame is a subcategorization frame or a subset of one. In this case, the naming conventions were different between VerbNet and Cyc. Frames with prepositions kept Cyc's notation for prepositional phrases. However, since VerbNet had a much broader coverage the VerbNet subcat names were kept.

4.6 Assertions

The process above was used to create general class frames, for example,

```
(verbClassSemTrans give-13.1
(TransitiveNPFrame)
```

```
(and
(isa :ACTION
MakingSomethingAvailable)
(patient-GenericDirect :ACTION
:OBJECT)
(performedBy :ACTION :SUBJECT)
(fromPossessor :ACTION :SUBJECT)
(objectOfPossessionTransfer :ACTION
:OBJECT)))
```

These frames use more generic collections and apply to a VerbNet class rather than a specific verb.

Specific verb semantic templates were created by inferring that each member verb of a VerbNet class participated in every template in a class. Again, collections were taken from existing WordNet connections if possible. The output was assertions in the Cyc semantic template format:

```
(verbSemTrans Loan-TheWord 0
(PPCompFrameFn NP-PP (WordFn to))
(and
(isa :ACTION Lending)
(patient-GenericDirect :ACTION
:OBJECT)
(performedBy :ACTION :SUBJECT)
(fromPossessor :ACTION :SUBJECT)
(toPossessor :ACTION :OBLIQUE-
OBJECT)
(objectOfPossessionTransfer :ACTION
:OBJECT)))
```

This method for giving class templates to each verb in a class was written as a Horn clause for the FIRE reasoning engine. FIRE is a reasoning engine that incorporates both logical inference based on axioms and analogy-based reasoning over a Cyc-derived knowledge base (Forbus, Mostek, & Ferguson, 2002). FIRE could then be queried for implied verb templates which became the final list of verb templates.

4.7 Subclasses

VerbNet has an extensive classification system involving subclasses. Subclasses contain verbs that take all of the syntactic formats of the main class plus additional frames that verbs in the main class cannot.

Verbs in a subclass inherit frames from their superordinate classes. FIRE was used again to create the verb semantic templates.

Each subclass template's collection was selected using the same process as the main class. If no subclass member had a Cyc denotation, then the main class collection was used.

5   Results

The end result of this process was the creation of 27,909 verb semantic template assertions for 5,050 different verbs. This substantially increases the number of frames for ResearchCyc and creates frames for OpenCyc.

To test the accuracy of the results and their contribution to the knowledge base we ran two tests. The first was to compare our frames with the 139 hand-checked VxC matches by hand. Of the 139 frames from VxC, 81 were qualified as "good" matches, and 58 as "maybe" (Trumbo, 2006). Since these frames already existed in Cyc and were hand matched we used them as the current gold standard for what a VerbNet frame translated into Cyc should look like.

Matches between frames were evaluated along several criteria. First was whether the frame had as good a syntactic parse as the manual version. This was defined as having predicates that addressed all syntactic roles in the sentence or, if not enough, as many as the VxC match. Secondly we asked if the collection was similar to the manual version. Frames with collections that were too specific, unrelated, or just `Situation` were discarded. Because frame-specific predicates were not created on a large scale, a frame was not rejected for using general predicates.

It is important to note a difference in matching methodology between the VxC matches and our frames. First, the VxC mappings included frames in Cyc that only partially matched more syntactically robust VerbNet frames. Our frames were only included if they matched the intended VerbNet syntactic frame. Because of this some of our frames beat the VxC gold standard for syntactic completeness. The VxC frames also included multiple similar senses for an individual verb. Our verbs had one denotation per class or subclass. Thus in some cases our frames failed not from over generalizing but because they were only meant to represent one meaning per class. Since the strength of our approach lies in generating a near exhaustive list of syntactic frames and not multiple word senses, these kinds of failures are not necessarily representative of the success of the frames as a whole.

A total of 55 frames (39.5%) were correct with seventeen (30.9%) of the correct frames having a more complete syntactic parse than the manually mapped frame. 48 frames (34.5%) were rejected only for having too general or specific a collection; however ten (20.8%) of the collection

rejected frames had a more complete parse than their manual counterparts. Thus 103 frames (74.1%) were as syntactically correct or better than the existing Cyc frame mapped to that VerbNet frame. Nine (6.47%) frames failed syntactically, with four (44.4%) of the syntax failures also having the wrong collection. Thirteen frames (9.3%) were not matched.

Fifteen frames (10.8%) from the *Hold* class, were separated out for a formatting error that resulted in a duplicate, though not syntactically incorrect, predicate. The predicate repeated was `(objectsInContact :ACTION :OBJECT)`. 12 of 15 frames (80%) had accurate collections.

The second test compared the results of a natural language understanding system using either ResearchCyc alone or a version of ResearchCyc with our frames substituted for theirs. The test corpus was 50 randomly selected example sentences from the VerbNet frame examples. We used the EA NLU parser, which uses a bottom-up chart parser and compositional semantics to convert the semantic content of a sentence in CycL (Tomai & Forbus 2009). Possible frames are returned in choice sets. A parse was judged correct if it returned a verb frame for the central verb of the example sentence that either wholly or in combination with preposition frames addressed the syntactic constituents of the sentence with an acceptable collection and acceptable predicates. Again general predicates were acceptable.

ResearchCyc got sixteen out of 50 frames correct (32%). Eleven frames (22%) did not return a template but did return a denotation to a Cyc collection. Twelve verbs (24%) retuned nothing, while eleven (22%) returned frames that were either not the correct syntactic frame or were a different sense of the verb.

EA NLU running the VerbNet generated frames got 26 out of 50 (52%) frames correct. Twelve frames (24%) returned nothing. Eight frames, (16%) failed because of a too specific or too general collection. Four generated frames (8%) were either not the correct syntactic frame or were for a different sense of the verb. This was an overall 20% improvement in accuracy.

Five (10%) parses using the VerbNet generated correct frames that were labeled as noisy. Noisy frames had duplicate predicates or more general predicates in addition to the specific ones. The *Hold* frames separated out in the VxC test are an example of noisy frames. None of these frames were syntactically incorrect or contradictory. The redundant predicates arise because the predicate safety net had to be greedy. This was in the interest of capturing more complex frames that may have multiple relations for the same thematic role in a sentence.

This evaluation is based on parser recall and frame semantic accuracy only. As would be expected, adding more frames to the knowledge base did result in more parser retrievals and possible interpretations. The implications for this on word sense disambiguation is evaluated further in the discussion. To improve predicate specificity, the next phase of research with these frames will be to implement predicate strengthening methods that move down the hierarchy to find more specific predicates to replace the generalized ones. Thus in the future precision both in terms of frame retrieval and predicate specificity will be a vital metric for evaluating success.

6   Discussion

As has been demonstrated in this approach and in previous research like Curtis *et al*'s (2009) TextLearner, Cyc provides powerful reasoning capabilities that can be used to successfully infer more specific information from general existing facts. We hope that future research is able to use this feature to provide more specific individual frames. Because Cyc is consistently changing and growing, an approach that uses Cyc relationships will be able to improve as the knowledge base improves its coverage.

While many of the frames are general, they provide a solid foundation for further research. As they are now, the added 27,909 frames increase the language capabilities of OpenCyc which previously had none. For ResearchCyc the contribution is less clear-cut. The 27,909 VerbNet frames have approximately 7.93 times the coverage of the existing 3,517 ResearchCyc frames[5] and they improved ResearchCyc parser performance by 20%. However, with 35% of frames in the VxC comparison and 16% in the parse test failing because of collections, and 10.8% of the VxC comparison set and 10% of correct parses classified as noisy, these frames are not as precise as the existing frames. The goal of these frames is not necessarily to replace the existing frames, but rather to extend coverage and provide a platform for further development whether by hand or through automatic methods. Precision can be improved upon in future

---

[5] D. Lenat briefing, March 15, 2006

research and is facilitated by the expressiveness of Cyc. Predicate strengthening, using existing relationships to infer more specific predicates, is the next step in creating robust frames.

Additionally, there is a tradeoff between the number of frames covered and efficiency of disambiguation. More frame choices make it harder for parsers to choose the correct frame, but it will hopefully improve their handling of more complex sentence structures.

One possible solution to competition and over-generality is to add verbs incrementally by class. The class based approach makes it easy to separate verbs by types, such as verbs that relate to mechanical processes or emotion verbs. One could use classes of frames to strengthen specific areas of parsing while choosing not to take verbs from a class covering a domain that the parser already performs strongly in. This approach can reduce interference with existing domains that have been hand built and extended beyond the standard Cyc KB for individual research.

Furthermore, semi-automatic approaches like this generate information more quickly than one could do by hand. Thus an approach to computational verb semantic representation that is rooted in classes can take advantage of modern reasoning sources like Cyc to efficiently create semantic knowledge.

References

Crouch, Dick, and Tracy Holloway King. 2005. Unifying Lexical Resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbruecken, Germany

Curtis, John, David Baxter, Peter Wagner, John Cabral, Dave Schneider, and Michael Witbrock. 2009. Methods of Rule Acquisition in the TextLearner Systerm. In *Proceedings of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read*, pages 22-28, Palo Alto, CA. AAAI Press.

Curtis, John, John Cabral, and David Baxter. 2006. On the Application of the Cyc Ontology to Word Sense Disambiguation. In *Proceedings of the Nineteenth International FLAIRS Conference*, pages 652-657, Melbourne Beach, FL.

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

Forbus, Kenneth, Thomas Mostek , and Ron Ferguson. 2002. An Analogy Ontology for Integrating Analogical Processing and First-principle Reasoning. In *Proceedings of the Thirteenth Conference on Innovative Applications of Artificial Intelligence.* Menlo Park, CA. AAAI Press.

Kipper, Karin, Hoa Trang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*, Austin, TX.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with Novel Verb Classes. In *Fifth International Conference on Language Resources and Evaluation (LREC 2006).* Genoa, Italy.

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

Matuszek, Cynthia, John Cabral, Michael Witbrock, and John DeOliveira. 2006. An Introduction to the Syntax and Content of Cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, Stanford, CA.

Moens, Marc, and Mark Steedman. 1988. Temporal Ontology and Temporal Reference. *Computational Linguistics*. 14(2):15-28.

Miller, G. 1985. WORDNET: A Dictionary Browser. In *Proceedings of the First International Conference on Information in Data.*

Ramachandran, Deepak, Pace Reagan, and Keith Goolsbey. 2005. First-Orderized Research Cyc: Expressivity and Efficiency in a Common-Sense Ontology. In *Papers from the AAAI Workshop on Contexts and Ontologies: Theory, Practice and Applications.* Pittsburgh, PA.

Tomai, Emmet, and Kenneth Forbus. 2009. EA NLU: Practical Language Understanding for Cognitive Modeling. In *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference*, Sanibel Island, FL.

Trumbo, Derek. 2006. VxC: A VerbNet-Cyc Mapper. http://verbs.colorado.edu/verb-index/vxc/