High-Fidelity Lexical Axiom Construction from Verb Glosses

Gene Kim
University of Rochester
Department of Computer Science
gkim21@cs.rochester.edu

Lenhart Schubert
University of Rochester
Department of Computer Science
schubert@cs.rochester.edu

Abstract

This paper presents a rule-based approach to constructing lexical axioms from WordNet verb entries in an expressive semantic representation, Episodic Logic (EL). EL differs from other representations in being syntactically close to natural language and covering phenomena such as generalized quantification, modification, and intensionality while still allowing highly effective inference. The presented approach uses a novel preprocessing technique to improve parsing precision of coordinators and incorporates frames, hand-tagged word senses, and examples from WordNet to achieve highly consistent semantic interpretations. EL allows the full content of glosses to be incorporated into the formal lexical axioms, without sacrificing interpretive accuracy, or verb-to-verb inference accuracy on a standard test set.

Evaluation of semantic parser performance is based on *EL-match*, introduced here as a generalization of the *smatch* metric for semantic structure accuracy. On gloss parses, the approach achieves an *EL-match* F1 score of 0.83, and a whole-axiom F1 score of 0.45; verb entailment identification based on extracted axioms is competitive with the state-of-the-art.

1 Introduction

Words encapsulate a great deal of knowledge, and in conjunction with language syntax, allow human beings to construct sentences that convey novel ideas to one another. Any system intended for broad natural language understanding will need to be able to perform inferences on the words that are the building blocks of language. For this reason, there have been many attempts to transduce informal lexical knowledge from machine readable dictionaries into a formally structured form (Calzolari, 1984; Chodorow et al., 1985; Harabagiu et al., 1999; Moldovan and Rus, 2001; Hobbs, 2008; Allen et al., 2013).

Gloss – *slam2.v* : "strike violently"
Axiom – ((x slam2.v y) ** e)
→ ((x (violently.adv (strike.v y))) ** e)

Figure 1: Example of rule extraction from machine readable dictionaries for WordNet entry of *slam2.v*.

Consider an example of the types of knowledge these approaches seek to extract in Figure 1. WordNet defines *slam2.v*, i.e., sense 2 of the verb *slam*, as "strike violently". This gloss states an implication that if "x slams y" characterizes an event e, then "x strikes y violently" also characterizes event e. All language phenomena must be able to be represented and reasoned about for such axioms to be useful in a language understanding system. This is where previous approaches share a common shortcoming: the logical representations that the lexical knowledge is mapped into are insufficient for representing many common natural language devices or for performing inference.

The contributions of this paper are the following:

- We demonstrate limitations in previous approaches to extracting lexical knowledge from machine readable dictionaries, particularly in their choices of logical representation.

- We present an approach to extracting lexical axioms in EL, which is a logical representation that overcomes these limitations. Our approach includes novel preprocessing and

information synthesis strategies for making precise axioms.

- We present *EL-smatch*, a generalized *smatch* scoring metric for partial scoring of semantic parses with complex operators and predicates.

The remainder of the paper presents related work in Section 2, background in Section 3, then a description of our semantic parsing approach in Section 4. A description of *EL-smatch* is presented in Section 5, followed by experimental results in Section 6, and future work and conclusions in Section 7.

2 Related Work

There have been many approaches in the past to extracting lexical information from machine-readable dictionaries. Early approaches to this problem focused on surface-level techniques, including hypernym extraction (Calzolari, 1984; Chodorow et al., 1985), pattern matching (Alshawi, 1989; Vossen et al., 1989; Wilks et al., 1989), and co-occurrence data extraction (Wilks et al., 1989).

In an evaluation of such methods, Ide & Veronis (1993) identified key challenges that thwart progress on this problem—challenges that persist to this day. Among these are the fact that dictionary glosses are often abstract, sometimes miss important information (such as arguments), and may be inconsistent with one another. Evidently there is a need for sophisticated extraction techniques to acquire accurate and consistent knowledge from dictionaries.

Most modern approaches to this problem use WordNet (Miller, 1995) as the lexical resource because of the linguistic and semantic annotations that accompany the glosses. Some work encodes WordNet glosses into variants of first-order logic (FOL) (Harabagiu et al., 1999; Moldovan and Rus, 2001; Hobbs, 2008), such as Hobbs Logical Form (HLF) (Hobbs, 1985), while other work encodes them into OWL-DL (OWL Working Group, 2004; Allen et al., 2011; Allen et al., 2013; Orfan and Allen, 2015; Mostafazadeh and Allen, 2015). A particularly noteworthy line of work is that by Allen et al. (2013), which integrates information from a high-level ontology with glosses of semantically related clusters of words to construct inference-supporting micro-theories of concepts corresponding to these words. While these advances are significant, they are limited by the expressivity of the representations used, in comparison with the richness of natural language.

2.1 Limitations of Logical Representations Used by Previous Approaches

As discussed by Schubert (2015), the choice of semantic representation is an important component of the natural language understanding problem. Because of space constraints, we will discuss only a few of the relevant issues and point the reader to (Schubert, 2015) for a more in-depth analysis of the issues at hand. The logical representation used for robust language understanding must satisfy the following requirements:

- Express the semantic content of most, if not all, possible natural language constructions;
- Have associated methods of inference;
- Have a formal interpretation.

The semantic representations used by previous approaches fall short on at least one of the above requirements. FOL struggles to express predicate modification (especially nonintersective modification), nonstandard quantifiers such as *most* or *at least 50*, and modality. Approaches that rely on functionalizing predication and connectives as a means of allowing for arbitrary propositional attitudes ultimately fail because quantifiers cannot be functionalized; thus they cannot capture the meaning of sentences with a modally embedded quantifier such as the following (with *believes* taking scope over *every*):

Kim believes that every galaxy harbors life.

HLF (Hobbs, 1985) is another common choice of semantic representation. It strives to capture sentential meaning within a subset of FOL by treating all words as predicates, including negation, disjunction, quantifiers, and modifiers. But it is unable to distinguish between events and propositions and between predicate and sentence modifiers, and the formal interpretation of quantification in HLF can lead to contradiction (Schubert, 2015).

OWL-DL (OWL Working Group, 2004) was designed for knowledge engineering on specific domains and thus cannot handle many common

natural language phenomena, such as predicate and sentence reification, predicate modification, self-reference, and uncertainty. There have been many efforts to allow for such phenomena, with varying degrees of success. As just one example, consider the common practice in OWL-DL of treating predicate modification as predicate intersection. For example, "whisper loudly" is represented as whisper $\sqcap \forall_{of}$-1.(loudly). *whisper* is the set of individual whispering events and $\forall_{of}$-1.(loudly) is the set of individual events that are modified by the adverb *loudly*. But according to WordNet, to whisper is to speak softly, so under an intersective interpretation of the modifiers, a loud whisper is both soft and loud. Similarly, WordNet glosses the verb *spin* as *revolve quickly*, so that under an intersective interpretation, a slow spin is both quick and slow. Analogously for nouns, a large pond or large brochure would be both large and small (*brochure* is glossed as *a small book*, and *pond* as *a small lake*). Even more difficult issues, from an OWL-DL perspective, are generalized quantifiers, uncertainty, attitudes, and reification, such as exemplified in the sentence

When self-driving cars are properly adopted, vehicles that need humans to drive them will probably be banned, according to Tesla CEO Elon Musk.

For a fuller discussion of issues in representations based on FOL, HLF, OWL-DL, etc., again see (Schubert, 2015).

3 Background

This section describes background material underlying our semantic parsing approach. First, we describe WordNet (Miller, 1995), our input lexical resource. Then, we describe Episodic Logic (EL), our choice of semantic representation for lexical axioms.

3.1 WordNet

WordNet is a lexical knowledge base that contains glosses for words, enumerates the word senses of each word, groups synonyms into *synsets*, encodes generality/specificity relations as *hypernym/hyponyms*, and provides schematic sentence structures for each word in the form of simple *frames*. The semantic annotations accompanying the glosses help in building a robust parser by reducing the amount of inference necessary for building axioms and assisting in handling mistakes in the glosses. Also, a significant proportion of the words in WordNet glosses have been tagged with their word senses and part-of-speech in the Princeton Annotated Gloss Corpus.[1] This helps with the important but often neglected word sense disambiguation (WSD) aspect of the interpretation problem; certainly ambiguous or faulty WSD can lead to misunderstandings and faulty inferences (is *Mary had a little lamb* about ownership or dining?). We use WordNet 3.0, which at the time of writing is the most recent version that is fully available for the UNIX environment, and focus on the verbs in this paper.

3.2 Episodic Logic

EL (Schubert and Hwang, 2000) was designed to be close to natural language, with the intuition that a logical representation that retains much of the expressivity of natural language will be able to more fully represent the complex constructs in natural language. EL provides constructs that are not common in most FOL-based languages, such as predicate modifiers, generalized quantifiers, reification, and ways of associating episodes (events or situations) with arbitrarily complex sentences. Importantly, EL is backed by a comprehensive inference system, EPILOG, which has been shown to be competitive with FOL theorem provers despite its substantially greater expressivity (Morbini and Schubert, 2009).

EL uses infix notation for readability, with the "subject" argument preceding the predicate and any additional arguments following the predicate. For associating episodes with logical sentences, EL introduces two modal operators '**' and '*'. [Φ ** e] means that Φ *characterizes* (i.e. describes as a whole) episode e and [Φ * e] means that Φ is true in (i.e. describes a piece or aspect of) episode e.

We show that EL overcomes some of the limitations of previous work that have been discussed using an example. Below is the EL representation for the sentence *Kim believes that every galaxy harbors life*.

```
(Kim.name believe.v
  (That (∀x (x galaxy.n)
            (x harbor.v (K life.n)))))
```

`That` and `K` are sentence and predicate reifica-

[1] http://wordnet.princeton.edu/glosstag.shtml

tion operators, respectively and `(∀x Φ(x) Ψ(x))` is equivalent to `(∀x Φ(x) → Ψ(x))`.[2] For discussion of the semantic types of the operators alluded to in this section and the connection to Davidsonian event semantics and other variants of event/situation semantics, see the papers describing EL (Schubert and Hwang, 2000; Schubert, 2000).

4 Gloss Axiomatization

In this section, we describe our approach to semantic parsing and axiomatization of WordNet entries. Our approach consists of three major steps:

1. Argument structure inference (Section 4.1)
2. Semantic parsing of the gloss (Section 4.2)
3. Axiom construction (Section 4.3)

Figure 2 shows the entire process for the previously introduced example, *slam2.v*. The argument inference step refines the WordNet sentence frames using the provided examples. Specific pronouns associated with argument position are inserted as dummy arguments into the corresponding argument positions in the gloss, and the modified gloss is semantically parsed into EL. Axiom construction replaces the dummy arguments with variables and constructs a scoped axiom relating the entry word and the semantic parse of the gloss using the characterization operator '**'. In the simple example *slam2.v*, most of the subroutines used in each step have no effect. All transformations outside the scope of the BLLIP parser are performed with hand-written rules, which were fine-tuned using a development set of 550 verb synset entries.

4.1 Argument Structure Inference

We initially use the frames in WordNet to hypothesize the argument structures. For example, the frames for *quarrel1.v* are [Somebody quarrel1.v] and [Somebody quarrel1.v PP]. From this we hypothesize that *quarrel1.v* has a subject argument that is a person, no object argument, and may include a prepositional phrase adjunct.

Then we refine the frames by looking at the examples and gloss(es) available for the *synset*. The examples for *quarrel1.v*: *"We quarreled over the question as to who discovered America"* and *"These two fellows are always scrapping*[3] *over something"* suggest that the subject argument can be plural and the PP can be specialized to PP-OVER. We identify the arguments and semantic types of the examples through a semantic parse, which is obtained using the method described in Section 4.2. Then we either update existing frames or introduce additional frames based on the agreement among examples and the number of available examples. We similarly obtain semantic types for arguments from glosses. For example, *paint1.v* has the gloss *"make a painting"* and the frame [Somebody -s Something]. Based on the gloss, we infer that the semantic type for the object argument is *painting*. Gloss-based argument structure inference can be done during the gloss parsing step, to avoid redundant computation.

Finally, we merge redundant frames. For example, frames that differ only in that one has *somebody* in a certain argument position where the other has *something* are merged into one frame where we simply use *something* (as a category allowing for both things and persons). Also there are rules for merging predicate complement types (Adjective/Noun & PP → Adjective/Noun/PP) and adding dative alternations to ditransitive frames [Somebody -s Somebody Something] → [Somebody -s Something to Somebody].

4.2 Semantic Parsing of Glosses

Sentence-level semantic parsers for EL have been developed previously, which we can use for semantic parsing of the glosses (Schubert, 2002; Schubert and Tong, 2003; Gordon and Schubert, 2010; Schubert, 2014). For the parser to be effective, some preprocessing of the glosses is necessary because glosses often omit arguments, resulting in an incomplete sentence. There are also some serious shortcomings to general semantic parsers, particularly in handling coordinators *and/or*. In this section, we describe the complete semantic parsing process of glosses and the details of each step. Throughout our semantic parsing implementation, we use the tree-to-tree transduction tool (TTT) (Purtee and Schubert, 2012) for trans-

[2] However, EL's quantifier syntax also allows, e.g., `(most.det x Φ(x) Ψ(x))`, which is not reducible to FOL.

[3] *quarrel1.v* and *scrap2.v* are in the same synset, so they share example sentences and are interchangeable in this context.

WordNet entry
slam2.v
Tagged gloss: `(VB strike1) (RB violently1)`
Frames: `[Somebody slam2.v Something]`
`[Somebody slam2.v Somebody]`
Examples: ("slam the ball")

4.1 Argument Structure Inference
Refined Frames:
`[Somebody slam2.v Something]`

4.2 Semantic Parsing
Parse: `(Me.pro (violently1.adv (strike1.v It.pro)))`

4.3 Axiom Construction
Axiom:
```
(∀x1 (∀y1 (∀e [[x1 slam2.v y1] ** e]
               [[[x1 (violently1.adv (strike1.v y1))] ** e]
                and [x1 person1.n] [y1 thing12.n]
```

Figure 2: Example gloss axiomatization process for WordNet entry *slam2.v*. The numbering corresponds to the subsections where these stages are discussed in detail.

parent and modular tree transformations[4] and the BLLIP parser (Charniak, 2000) to get Treebank parses.

The general outline of the gloss processing steps is described below:

1. Create separate POS-tagged word sequences for distinct glosses:
   a. Label gloss g with POS tags using the Princeton Annotated Gloss Corpus, backing off to the synset type in the sense key.[5]
   b. Split multigloss trees along semicolons for individual POS tagged glosses $p_1, p_2, ..., p_n$.
2. Create an easy-to-parse sentence for each gloss:
   a. Factor out coordinators, leaving the first conjunct in the gloss. Save the coordinated phrases c_{p_i} for later insertion.
   b. Insert dummy arguments (*I*, *it*, *them*).
   c. Drop POS tags to create new gloss g_i'.
3. Syntactically parse each gloss sentence into initial LFs:
   a. Parse g_i' into tree t_i using the BLLIP parser.
   b. Refine POS tags in t_i using the Princeton Annotated Gloss Corpus.
   c. Run t_i through the sentence-level semantic parser to get logical form s_i.
4. Refine the initial LFs:
   a. Reinsert coordinated phrases c_{p_i} into s_i.
   b. Introduce word senses into the logical form.

We now describe the sentence-level semantic parser, coordinator factorization, argument insertion/inference, and word sense introduction in more detail.

4.2.1 Sentence Level Semantic Parser

The sentence-level semantic (EL) parser we use is modeled after the partial interpreter used by the KNEXT system (Van Durme et al., 2009; Gordon and Schubert, 2010). First, the parser applies corrective and disambiguating transformations to raw Treebank trees. For example, these correct certain systematic prepositional phrase (PP) attachment errors, distinguish copular *be* from other forms, assimilate verb particles into the verb, particularize SBAR constituents to relative clauses, adverbials, or clausal nominals, insert traces for dislocated constituents, etc. Second, the parser uses about 100 rules to compositionally interpret Treebank parses into initial interpretations. Finally, coreference resolution, quantifier, coordinator, and tense scoping, temporal deindexing, (non-embedded) Skolemization, equality reduction, conjunction splitting and other canonicalization operations are applied to refine the logical form.

4.2.2 Argument Insertion and Inference

WordNet glosses (and glosses in general) only include arguments when necessary to specify some semantic type for the argument. Figure 3 displays example glosses from WordNet that demonstrate this treatment of arguments. Both the subject and object arguments in the gloss of *slam2.v* are omitted, and the subject is omitted from the gloss of *paint1.v*, while the object in the gloss is included.

[4] We do not explicitly state where TTT is used in the algorithm since it is a general tree transformation tool, which is used throughout the algorithm whenever a tree transformation is necessary.

[5] Every word in the glosses of the Princeton Annotated Gloss Corpus is labeled with the POS tag or the sense key. The synset type distinguishes between nouns, verbs, adjectives, and adverbs.

Argument position	*English text*	*EL atom*
subject	I/my/myself	Me.pro
direct object	it	It.pro
indirect object	them	They.pro

Table 1: Mappings between dummy argument position, text, and EL atoms.

slam2.v – *subject* strike *object* violently
paint1.v – *subject* make a painting

Figure 3: Example glosses demonstrating the treatment of arguments in glosses. Underlined words are arguments and italicized arguments indicate where an argument should exist, but does not in the gloss.

We make arguments explicit and unify their treatment in order to improve Treebank and semantic parses and simplify the axiom construction step, described in Section 4.3. Figure 4 shows unified versions of the glosses that appear in Figure 3, *slam2.v* and *paint1.v*. In this unified treatment, all arguments are represented by argument position-specific dummy arguments. Table 1 lists the dummy arguments and their relation to the argument position and EL. Dummy arguments are inserted into the POS tagged gloss p_i based on the inferred argument structure from Section 4.1 and the insertions are achieved through pattern matching of the POS tags.

Finally, some glosses contain references to the subject using the terms *one*, *one's*, or *oneself* (e.g. *sprawl1.v* : *sit or lie with one's limbs spread out*). These are mapped to *I*, *my*, and *myself*, respectively to correctly corefer with the dummy subject argument *I*.

4.2.3 Coordinator Factorization

Treebank and semantic parsers are prone to errors for coordinated phrases, often mistaking them for appositives, or vice-versa. To minimize such errors, we developed a method of factorizing coordinated phrases. The conjuncts can usually be identified by syntactic and semantic relatedness. This

slam2.v – *I'll* strike *it* violently
paint1.v – *I'll* make *it*; (*it* : a painting)

Figure 4: Unified versions of WordNet glosses from Figure 3.

can be seen in the WordNet gloss for *edit1.v*: *prepare for publication or presentation by correcting, revising, or adapting*. We use linguistic phrase types as a proxy for syntactic and semantic relatedness. That is, we identify coordinated groups of verb phrases, noun phrases, adjectival phrases, and prepositional phrases. These phrase groups are pulled out of the sentence, and only the first phrase in the group is left in the sentence.

The phrase groups are identified using a set of rules that were fine-tuned with reference to the development set of verb synsets. The rules tend to handle common modifications, such as adjectives in noun phrases. For ambiguous cases, such as prepositional modification, factorization is not performed.

The phrase groups are passed through a modified sentence-level semantic parser (stopping short of the coordinator scoping step), and embedded back into the gloss logical form before the coordinator scoping step in the semantic parsing of the gloss. The place of insertion is identified by matching the first phrase in the phrase group with a phrase in the logical form.

4.2.4 Word Sense Introduction

Word sense introduction is assisted by the hand-tagged word senses in WordNet. All words that are not hand-tagged with a word sense are given the lowest numbered word sense with a frame matching the context of its use in the gloss. Generally, the lower numbered word senses in WordNet are the most relevant senses of the word.

4.3 Axiom Construction

Finally, we take the results from Sections 4.1 and 4.2 and construct the axiom. Dummy arguments in the parsed gloss are correlated with the arguments in the frame using the mapping in Table 1. We replace the arguments with variables, introduce logical formulas asserting the semantic types (from the argument structure in Section 4.1), and construct an axiom asserting that the truth of the entry word with the proper argument structure (without semantic types) implies the truth of the semantic parse of the gloss and semantic types of the arguments. Before axiom construction, the example from Figure 2, *slam2.v*, has the following refined frame and semantic parse of the gloss from Sections 4.1 and 4.2, respectively:

```
[Somebody slam2.v Something]
[Me.pro
```

```
(violently1.adv (strike1.v It.pro))]
```

After we replace the arguments and create formulas asserting the semantic types, we have:

```
[x1 slam2.v y1]
[x1 (violently1.adv (strike1.v y1))]
[x1 person1.n], [y1 thing12.n]
```

Finally, we construct an axiom of form $(\forall_x \Phi\ \Psi)$ (equivalent to $(\forall_x \Phi \rightarrow \Psi)$) and using the modal *characterization* operator **:

```
(∀x1,y1,e
  [[x1 slam2.v y1] ** e]
  [[[x1 (violently1.adv
          (strike1.v y1))] ** e]
   and [x1 person1.n] [y1 thing12.n]
```

We can easily generate converse axioms as well, such as that if a person strikes something violently, then it is probably the case that he or she slams it (in the *slam2.v* sense). EL allows us to express a degree of uncertainty in the formulation of the converse, and this is appropriate to the extent that lexical glosses cannot be expected to provide complete, "airtight" definitions, but rather just the most important semantic content. However, in this paper we limit ourselves to discussion of the "forward" version of gloss-derived axioms.

5 *EL-smatch*

In this section we introduce *EL-smatch*, a generalized formulation of *smatch* (Cai and Knight, 2013), the standard evaluation metric for AMR parsing (Banarescu et al., 2013). *Smatch* represents each logical form as a conjunction of triples of three types:

1. *instance(variable, type)*
2. *relation(variable, variable)*
3. *attribute(variable, value)*

Every node instance of the logical form is associated with a variable, and the nodes are described and related to each other using the above triples. Thus, *type* and *value* can both only be atomic constants. The *smatch* score is then defined as the *maximum f-score (of triples) obtainable via a one-to-one matching of variables between the two formulas* (Cai and Knight, 2013).

In order to capture complex types of EL, we introduce an additional triple:

instance(variable, variable).

The first variable argument is associated with the instance, and the second variable argument, with the type.

EL

```
(me.pro (very.adv happy.a))
```

***EL-smatch* Triple Representation**

```
instance(a, very.adv) ∧
instance(b, happy.a) ∧
instance(d, me.pro) ∧
ARG0(a, b) ∧
instance(c, a) ∧
ARG0(c, d)
```

***EL-smatch* Graph Representation**

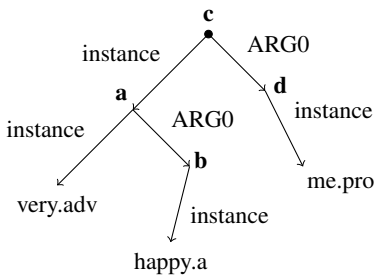

Figure 5: Example of syntactic mapping from EL to *EL-smatch* triple and graph representations for sentence "I am very happy".

With this addition to the representation, we can syntactically map EL formulas into a conjunction of triples by introducing a node variable for every component of the formula and then describing and relating the components using the triples. Since the representation used by *smatch* is the same as that of AMR, we can map the triple representation into a graph representation in the same manner as AMR formulas. Figure 5 shows an example of the use of the new instance triple in mapping the EL formula for "I am very happy" into these representations. However, this mapping does not relate the semantics of EL to AMR since the interpretation of the triples differ for AMR and EL formulas.

6 Experiments

We conducted two experiments to demonstrate the efficacy of our approach for semantic parsing and the usefulness of the resulting axioms for inference.[6]

[6]One reviewer suggested comparing our axioms with ontologies linked to WordNet, such as SUMO (Niles and Pease, 2001) and DOLCE (Gangemi et al., 2002), or with the hypernym hierarchy of WordNet. Such an experiment was performed by Allen et al. (2013), which showed that WordNet glosses contain information that is not found in the structural relations of WordNet. A similar experiment by us is unlikely to shed additional light on the topic.

Measure	Precision	Recall	F1
EL-smatch	0.85	0.82	0.83
Full Axiom	0.29	1.00	0.45

Table 2: Performance against gold standard parses of 50 synsets.

6.1 Semantic Parsing Evaluation

We constructed a gold standard set of axioms by selecting 50 random WordNet synsets that were not used during development. Gold standard axioms for these synsets were written by the first author, then refined in collaboration between the two authors.[7] The 50 synsets resulted in 52 axioms and 2,764 triples in the gold standard. The results in Table 2 show the system performance using both *EL-smatch* and full axiom metrics. In the full axiom metric, the precision measures the number of axioms that are completely correct, and the recall measures the number of axioms generated (which can vary due to merged glosses and multiple frames). The *EL-smatch* score of 0.83 shows that the axioms are generally good, even when not completely correct. Generating completely correct axioms is difficult because there are multiple non-trivial subproblems, such as prepositional attachment and word sense disambiguation. *EL-smatch* displays a more fine-grained measure of our system performance than the full axiom metric.

6.2 Inference Evaluation

To our knowledge, no earlier work evaluates inference in a manner that captures the details of semantically rich lexical axioms. Therefore, in order to compare our results to previous work, we evaluate a stripped-down version of our inference mechanism on a manually created verb entailment dataset (Weisman et al., 2012). This dataset contains 812 directed verb pairs, $v1 \rightarrow v2$, which are annotated 'yes' if the annotator could think of plausible contexts under which the entailment from $v1$ to $v2$ holds. For example, *identify* entails *recognize* in some contexts, does not entail *describe* is any context. Though the dataset is not rich, many previous systems (Mostafazadeh and Allen, 2015; Weisman et al., 2012; Chklovski and Pantel, 2004) have evaluated on this dataset, establishing it as a basic standard of comparison. In order to fit our axioms to this dataset, we remove semantic roles (verb arguments and adjuncts) from our axioms. Also, since the dataset has no word senses, the inferences begin with every synset that contains a sense of the starting word, and the final predicted entailments suppress sense distinctions. When generating inferences, we find verbs in the consequent of the axiom that are not modified by a negation or negating adverb (e.g., *nearly, almost,* etc.). Such inferences are chained up to three times, or until an abstract word is reached (e.g., *be, go,* etc.), which glosses do not sufficiently describe. This blacklist contains 24 abstract words.

Method	Precision	Recall	F1
Our Approach	0.43	0.53	0.48
TRIPS	0.50	0.45	0.47
Supervised	0.40	0.71	0.51
VerbOcean	0.33	0.15	0.20
Random	0.28	0.29	0.28

Table 3: Performance against gold standard parses of 50 synsets.

Table 3 shows the results on this dataset. *TRIPS* is an approach by Mostafazadeh & Allen (2015), which constructs axioms from WordNet using the TRIPS parser and represents its axioms in OWL-DL, *Supervised* is a supervised learning approach by Weisman et al. (2012), *VerbOcean* classifies entailments according to the strength relation of the VerbOcean knowledege-base (Chklovski and Pantel, 2004), and *Random* is a method that randomly classifies the pair with probability equal to the distribution in the testset (27.7%). The performance of our system is competitive with state-of-the-art systems *TRIPS* and *Supervised* on this task. Our system performance splits the performance of *TRIPS* and *Supervised* in all three measures.

The inference capabilities of our axioms exceed what is evaluated by this testset. Because of space constraints, an example of a more expressive inference using extracted axioms is included in supplementary material.

6.3 Error Analysis

In the semantic parsing evaluation, most of the parsing errors arose from a failure in the sentence parser or preprocessing directly preceding the sentence parser. That is, 17 out of the 52 axioms had errors arising from the sentence parser. These er-

[7] Due to time constraints, this evaluation was performed on a gold standard developed primarily by only one annotator. We hope to remedy this in future work including an analysis of interannotator agreement.

rors arose from either linguistic patterns that we did not encounter in our development set or in complex sentences (e.g. *take a walk for one's health or to aid digestion, as after a meal*). Many of these can be avoided in the future by increasing the development set. Fortunately, the semantic parser uses keywords to mark ambiguous attachments or phrases, so that in many cases, axioms that are not fully parsed can be identified and ignored, rather than using an incorrectly parsed axiom.

WSD and incorrect scoping of semantic types are also major sources of errors. The challenge of WSD was minimized by the subset of hand-tagged word senses in WordNet. We may be able to reduce such errors in the future by merging together redundant or overly specific word senses. Incorrect scoping of semantic types is particularly problematic when the semantic type is specified in the gloss itself, as the type constraint needed to move across scopes. Our system performed well on coordinator scoping. We correctly scoped 23 of the 27 instances of coordinators in the dataset. Coordinators are generally a great source of error in parsers and this result is evidence of the effectiveness of our coordinator handling mechanism. In all four instances, the disjunctions were extracted from the gloss correctly, but were not reintroduced into the axiom. As such, this error did not make these axioms incorrect, rather incomplete.

7 Future Work and Conclusions

There are many attractive directions for future work. The scope of this project can be broadened to include nouns, adjectives, and adverbs, as required for any system that actually tackles the natural language understanding problem. There are also many ways to refine and deepen the gloss interpretation process. The parses may be improved by looking through the *hypernym* graph and borrowing results from parses of parents (generalizations) of words. We can also incorporate techniques from Allen et al. (2011; 2013) and Mostafazadeh & Allen (2015) to integrate results from related sets of glosses. The high-level TRIPS ontology could be used to improve robustness in the face of inconsistencies in WordNet and interpretation errors. Also, more sophisticated WSD techniques, such as those from the SENSEVAL-3 task on WSD (Litkowski, 2004), could be used to improve semantic precision, and argument coherence could be improved using techniques from Mostafazadeh & Allen (Mostafazadeh and Allen, 2015). Another possible avenue is concurrent use of information from multiple dictionaries, such as Wiktionary, VerbNet, and WordNet, to construct more complete and reliable axioms, in particular with respect to argument structure and types.

We argued that the semantic representations used in previous approaches to extracting lexical axioms from dictionaries are insufficient for achieving a natural language understanding system. We presented an approach to extracting lexical axioms of verbs from WordNet into EL, an expressive semantic representation that overcomes the shortcomings of the representations used in the past. We also presented a generalized *smatch* scoring metric, *EL-smatch*, which we used to evaluate our system. The evaluation shows that our approach constructs precise verb axioms from WordNet. Furthermore, we demonstrate that the generated axioms perform competitively against the state of the art in a verb entailment task. We aim to apply these axioms to more comprehensive language understanding tasks and commonsense reasoning tests when we have sufficient coverage of the lexicon.

Acknowledgments

The work was supported by a Sproull Graduate Fellowship from the University of Rochester and NSF grant IIS-1543758. We are also grateful to the anonymous reviewers for their comments.

References

James Allen, William de Beaumont, Nate Blaylock, George Ferguson, Jansen Orfan, and Mary Swift. 2011. Acquiring commonsense knowledge for a cognitive agent. In *Proceedings of the AAAI Fall Symposium Series: Advances in Cognitive Systems (ACS 2011)*, Arlington, VA, USA.

James Allen, Will de Beaumont, Lucian Galescu, Jansen Orfan, Mary Swift, and Choh Man Teng. 2013. Automatically deriving event ontologies for a commonsense knowledge base. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 23–34, Potsdam, Germany, March. Association for Computational Linguistics.

Hiyan Alshawi. 1989. Analysing the dictionary definitions. In Bran Boguraev and Ted Briscoe, editors, *Computational Lexicography for Natural Language Processing*, pages 153–169. Longman Publishing Group, White Plains, NY, USA.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria, August. Association for Computational Linguistics.

Nicoletta Calzolari. 1984. Detecting patterns in a lexical data base. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 170–173, Stanford, California, USA, July. Association for Computational Linguistics.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 132–139, Stroudsburg, PA, USA. Association for Computational Linguistics.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 33–40, Barcelona, Spain, July. Association for Computational Linguistics.

Martin S. Chodorow, Roy J. Byrd, and George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting on Association for Computational Linguistics*, ACL '85, pages 299–304, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. 2002. Sweetening ontologies with DOLCE. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, EKAW '02, pages 166–181, London, UK. Springer-Verlag.

Jonathan Gordon and Lenhart Schubert. 2010. Quantificational sharpening of commonsense knowledge. In *Proceedings of the AAAI 2010 Fall Symposium on Commonsense Knowledge*.

Sanda Harabagiu, George Miller, and Dan Moldovan. 1999. WordNet 2 - A morphologically and semantically enhanced resource. In *SIGLEX99: Standardizing Lexical Resources*, pages 1–8, College Park, MD, USA, June. Association for Computational Linguistics.

Jerry R. Hobbs. 1985. Ontological promiscuity. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pages 60–69, Chicago, Illinois, USA, July. Association for Computational Linguistics.

Jerry R. Hobbs. 2008. Deep lexical semantics. In *Computational Linguistics and Intelligent Text Processing, 9th International Conference, CICLing Proceedings*, volume 4919 of *Lecture Notes in Computer Science*, pages 183–193, Haifa, Israel, February. Springer.

Nancy Ide and Jean Véronis. 1993. Knowledge extraction from machine-readable dictionaries: An evaluation. In *EAMT Workshop*, volume 898 of *Lecture Notes in Computer Science*, pages 19–34. Springer.

Ken Litkowski. 2004. Senseval-3 task: Word sense disambiguation of WordNet glosses. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 13–16, Barcelona, Spain, July. Association for Computational Linguistics.

George A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, November.

Dan Moldovan and Vasile Rus. 2001. Logic form transformation of WordNet and its applicability to question answering. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 402–409, Toulouse, France, July. Association for Computational Linguistics.

Fabrizio Morbini and Lenhart K. Schubert. 2009. Evaluation of EPILOG: a reasoner for Episodic Logic. In *Proceedings of the Ninth International Symposium on Logical Formalizations of Commonsense Reasoning*.

Nasrin Mostafazadeh and James F. Allen. 2015. Learning semantically rich event inference rules using definition of verbs. In *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing Proceedings, Part I*, volume 9041 of *Lecture Notes in Computer Science*, pages 402–416, Cairo, Egypt, April. Springer.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9, New York, NY, USA. ACM.

Jansen Orfan and James Allen. 2015. Learning new relations from concept ontologies derived from definitions. In *Proceedings of the AAAI 2015 Spring Symposium Series on Logical Formalizations of Commonsense Reasoning*.

W3C OWL Working Group. 2004. *OWL Web Ontology Language Guide*. W3C Recommendation. Available at `https://www.w3.org/TR/2004/REC-owl-guide-20040210`.

Adam Purtee and Lenhart Schubert. 2012. TTT: A tree transduction language for syntactic and semantic processing. In *Proceedings of the Workshop on Applications of Tree Automata Techniques in Natural Language Processing*, ATANLP '12, pages 21–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lenhart K. Schubert and Chung Hee Hwang. 2000. Episodic Logic meets Little Red Riding Hood: A comprehensive natural representation for language understanding. In Lucja M. Iwańska and Stuart C. Shapiro, editors, *Natural Language Processing and Knowledge Representation*, pages 111–174. MIT Press, Cambridge, MA, USA.

Lenhart Schubert and Matthew Tong. 2003. Extracting and evaluating general world knowledge from the Brown corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 7–13, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lenhart K. Schubert. 2000. The situations we talk about. In Jack Minker, editor, *Logic-based Artificial Intelligence*, pages 407–439. Kluwer Academic Publishers, Norwell, MA, USA.

Lenhart Schubert. 2002. Can we derive general world knowledge from texts? In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 94–97, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lenhart Schubert. 2014. From treebank parses to episodic logic and commonsense inference. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 55–60, Baltimore, MD, June. Association for Computational Linguistics.

Lenhart Schubert. 2015. Semantic representation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Benjamin Van Durme, Phillip Michalak, and Lenhart K. Schubert. 2009. Deriving generalized knowledge from corpora using WordNet abstraction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 808–816, Stroudsburg, PA, USA. Association for Computational Linguistics.

Piek Vossen, Willem Meijs, and M. den Broeder. 1989. Meaning and structure in dictionary definitions. In *Computational Lexicography for Natural Language Processing*, pages 171–192. Longman Publishing Group, White Plains, NY, USA.

Hila Weisman, Jonathan Berant, Idan Szpektor, and Ido Dagan. 2012. Learning verb inference rules from linguistically-motivated evidence. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 194–204, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yorick Wilks, Dan Fass, Cheng-ming Guo, James E. McDonald, Tony Plate, and Brian M. Slator. 1989. A tractable machine dictionary as a resource for computational semantics. In *Computational Lexicography for Natural Language Processing*, pages 193–228. Longman Publishing Group, White Plains, NY, USA.