Automatically Deriving Event Ontologies for a CommonSense Knowledge Base

James Allen[1,2], Will de Beaumont[1], Lucian Galescu[1], Jansen Orfan[2], Mary Swift[2] and Choh Man Teng[1]

[1] Institute for Human and Machine Cognition, Pensacola, FL
[2] Dept. Of Computer Science, University of Rochester
{jallen, wbeaumont, lgalescu, cmteng}@ihmc.us
{jorfan, swift}@cs.rochester.edu

Abstract

We describe work aimed at building commonsense knowledge by reading word definitions using deep understanding techniques. The end result is a knowledge base allowing complex concepts to be reasoned about using OWL-DL reasoners. We show that we can use this system to automatically create a mid-level ontology for WordNet verbs that has good agreement with human intuition with respect to both the hypernym and causality relations. We present a detailed error analysis that reveals areas of future work needed to enable high-performance learning of conceptual knowledge by reading.

1. Introduction

Most researchers agree that attaining deep language understanding will require systems that have large amounts of commonsense knowledge. Such knowledge will need to be expressed in terms that support semantic lexicons as used by parsing systems, with concept hierarchies and semantic roles, and provide knowledge required for disambiguation as well as deriving key entailments. While there have been many attempts to hand-build such knowledge, most notably within the Cyc project (Lenat, 1995), as well as ontology-building efforts such as SUMO (Niles & Pease, 2001), GUM (Bateman et al., 1995), DOLCE (Gangemi et al., 2002) and EuroWordNet (Vossen, 1998), these fall short of encoding the range and depth of needed knowledge. This motivates work in building a commonsense knowledge base automatically from reading online sources. Learning by reading offers the opportunity not only to amass a significant knowledge base for processing online sources, but also allows for learning on demand - i.e., looking up something in a dictionary or Wikipedia when needed.

Recently, there has been significant interest in acquiring knowledge using information extraction techniques (e.g., Etzioni et al, 2011; Carlson et al, 2010). Such work, however, remains close to the surface level of language - involving mostly uninterpreted words and phrases and surface relations between them (e.g., is-a-subject-of, is-an-object-of), or a limited number of pre-specified relations. In addition, information extraction tends to focus more on learning facts (e.g., *Rome is the capital of Italy*) than conceptual knowledge (e.g., *kill* means *cause to die*).

We have been exploring the feasibility of building extensive knowledge bases by reading definitional sources such as online dictionaries and encyclopedias such as Wikipedia. So far, we have focussed on what knowledge can be derived by reading the glosses in WordNet (Fellbaum, 1998). This is a good start for the project for several reasons. First, WordNet is the most used lexical resource in computational linguistics, and so a knowledge base indexed to WordNet would be most readily accessible for use in other projects. Second, a significant portion (i.e., about 50%) of the content words in WordNet glosses have been sense tagged by hand, thus giving us considerable help on tackling the word sense disambiguation problem. And third, WordNet has hand-built semantic structures, such as the hypernym and troponym hierarchies, as well as tagged relations such as *cause*, and *part-of*, which give us a hand-coded standard to compare against. While most previous work on extracting knowledge from WordNet has focused on exploiting these hand-built relations, we focus solely on what can be extracted by understanding the glosses, which consist of short definitions (e.g., kill: *cause to die*) and a few examples (e.g., *This man killed several people when he tried to rob a bank*), and use the hand-built relations for evaluation. The goal is a system that is not WordNet specific, but could be used on any source of definitional knowledge. This projects shares some of the same goals with the work of Nichols et al. (2005), who convert definitions from a machine readable dictionary of Japanese into underspecified semantic representations using Robust Minimal Recursion Semantics (Frank, 2004) and construct an ontology based on extracted hypernyms and synonyms.

While many complain about WordNet, it is an unparalleled lexical resource. Attempts to use WordNet as an ontology to support reasoning have mainly focussed on nouns, because the noun hypernym hierarchy provides a relatively good subclass hierarchy (e.g., Gangemi et al. 2003). The situation is not the same for verbs however. Verbs in WordNet have no organization into an ontology of event types in terms of major conceptual categories such as states, processes, accomplishments and achievements (cf. Vendler 1957). Instead, WordNet has a set of 15 semantic domains that serve as unique beginners for verbs, such as verbs of motion and verbs of communication. The verbs are then organized around a troponym hierarchy - capturing manner modifications (e.g., destroy is a killing done in a particular way). Fellbaum (1998) argues against a top-level verb distinction between events and states, or *be* and *do* as suggested in Pulman (1983), for several reasons. A goal of WordNet was to reflect human lexical organization, and there is a lack of psycholinguistic evidence that humans have strong associations between abstract concepts such as *do* and more specific verbs. This lack of a hierarchical mid-level[1] ontology for events creates a significant obstacle to unifying WordNet with ontologies that are built to encode commonsense knowledge and support reasoning.

In this paper, we report on work that attempts to address this problem and bring formal ontologies and lexical resources together in a way that captures the detailed knowledge implicit in the lexical resources. Specifically, we focus on building an ontology by reading word definitions -- and use WordNet glosses as our test case for evaluating the feasibility of doing so. It is important to remember here that our goal is to develop new techniques for building knowledge bases by reading definitions in general, and our work is not specific to WordNet, though we use WordNet for evaluation.

It is always difficult to evaluate the usefulness and correctness of ontologies. We resort to using several focussed evaluations of particular types of knowledge using human judgement. In some of these cases, we find that WordNet itself provides some information related to these aspects, so we can compare the coverage and accuracy of our automatically constructed ontology with the explicitly coded information in WordNet. For example, we can evaluate the coverage of our event hierarchy by comparing to the WordNet troponym hierachy, and we can compare the causal relationships we derive between events with the explicitly annotated *cause* relations in WordNet.

2. Encoding Knowledge in WordNet Glosses

There have been several prior attempts to process glosses in WordNet to produce axioms that capture entailments. For the most part, these representations are fairly shallow, and look more like an encoding of the syntactic information in a logical notation, with each word represented as a predicate. Furthermore, some of the encodings resist a formal interpretation. For instance, the representations in eXtended WordNet (Harabagiu et al. 2003) contain variables that are free, predicates that have variable arity, and lack a principled representation of logical operators, particularly disjunction. As such, it cannot support sound inference procedures. Furthermore the predicates are just words, not disambiguated senses. Clark et al. (2008) produce a representation where the predicates are senses, but share many of the other weaknesses of eXtended WordNet. Agerri & Peñas (2010) resolve a number of these issues and generate intermediate logical forms that have no free variables nor unconnected predicates in the definitions, but the formalism still resembles an encoding of syntax as opposed to a semantic representation. As an example, Figure 1 shows the representation generated for the definition of the adjective *bigheaded* as *overly conceited or arrogant.* It is not clear what the semantics of the encoding of disjunction (i.e., *conj_or(x3,x5)*) plays in the definition, as it appears that both modifiers *conceited* and *arrogant* appear in parallel *amod* relations to the variable *x1*. It is hard to imagine an inference mechanism that would handle the disjunction correctly given this representation.

[1] we distinguish between the upper ontology (identifying the fundamental conceptual distinctions underlying knowledge), the mid-level ontology (capturing general knowledge of events), and the domain ontology, capturing specific knowledge about particular domains.

something(x1) ^ amod(x1,x3) ^ amod(x1,x5) ^ overly(x2) ^ conceited(x3)
^ advmod(x3,x2) ^ conj_or(x3,x5) ^ arrogant(x5)

Figure 1: Agerri & Peñas (2010) representation of the gloss "overly conceited or arrogant"

Building a good ontology requires more than natural language processing--it requires sophisticated reasoning to identify subsumption relations implicit in the definitions. We pick our target formalism for the ontology to be description logic, specifically OWL, and use its associated reasoners to compute the subsumption relations. As an example, we encode the definition of *bigheaded* as

bigheaded ⊑ ∀_of.(person) ⊓ ((conceited ⊓ ∀_of -1.(degree-modifier and overly)) ⊔ arrogant)

i.e., *bigheaded* is a predicate that applies to people, and which is a subclass of the union of things that are conceited (with degree modifier overly) with things that are arrogant. Note that OWL allows types defined by relations and their inverses: ∀_of.(person) is the class of all objects that are in the domain of an of relation with only people (i.e., person) in the range, whereas ∀_of -1.(person) would be the class of all objects that are in the range of a relation with only person in the domain. While description logic is less expressive than first order logic, our experience has shown that it provides a good formalism for capturing much of the content in definitions and produces a representation that supports provably tractable inference about hierarchical relationships over complex types, making it suitable for encoding ontologies.

3. Parsing Glosses

We parse WordNet glosses with a slightly modified TRIPS parser (Allen et al., 2008). The TRIPS semantic lexicon provides information on semantic roles and selectional restrictions for about 5000 verbs, and the parser constructs a semantic representation of the language that is rich enough for reasoning. TRIPS has already shown promise in parsing WordNet glosses in order to build commonsense knowledge bases (Allen et al., 2011). The logical form is a graphical formalism that captures an unscoped modal logic (Manshadi et al. 2008). Figure 2 shows the logical form graph for the definition of *kill* as to *cause to die*. The graph consists of nodes that specify the word senses for each word (both its sense in the TRIPS ontology and the WordNet Sense) and quantification information, and relations between the nodes are labelled with semantic roles. The IMPRO nodes are derived from the gaps in the definition and become the arguments for the new concept, namely kill%2:35:00[2].

WordFinder

To attain broad lexical coverage beyond its hand-defined lexicon, the TRIPS parser uses input from a variety of external resources. WordFinder is a subsystem that accesses WordNet when an unknown word is encountered. The WordNet senses have hand-built mappings to semantic types in the TRIPS ontology, although sometimes at a fairly abstract level. WordFinder uses the combined information from WordNet and the TRIPS lexicon and ontology to dynamically build lexical entries with approximate semantic and syntactic structures for words not in the core lexicon.

WordFinder offers a powerful tool for increased lexical coverage. However, the information in entries constructed by WordFinder is frequently underspecified, so the parser must deal with significantly increased levels of ambiguity when dealing with dynamically constructed words. There are several

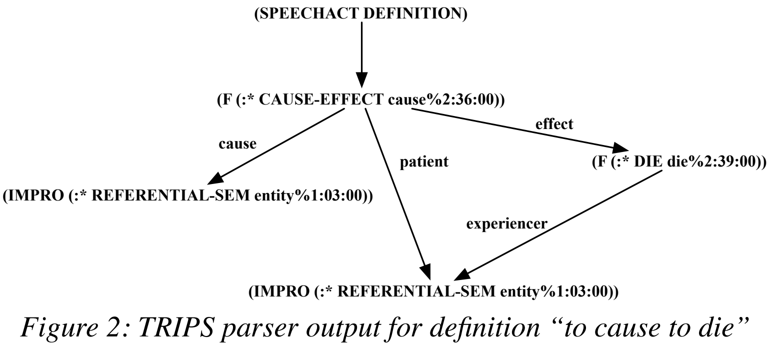
(SPEECHACT DEFINITION)
(F (:* CAUSE-EFFECT cause%2:36:00))
cause
effect
patient
(F (:* DIE die%2:39:00))
(IMPRO (:* REFERENTIAL-SEM entity%1:03:00))
experiencer
(IMPRO (:* REFERENTIAL-SEM entity%1:03:00))

Figure 2: TRIPS parser output for definition "to cause to die"

[2] We use the WordNet sense key notation throughout, which uses three values to identify a sense: kill%2:35:00 is a verb (2), is a verb of contact (35), and has a unique identifier (00) within this group.

settings that can be used to control how WordFinder is used during parsing. First, users can specify the number of senses returned from WordNet. WordNet may have multiple fine-grained senses for a given word, but depending on the application, selecting the most frequent senses listed in WordNet will suffice (cf. McCarthy et al. 2004).

Word Sense Disambiguation

As we mentioned earlier, one thing that makes WordNet glosses a good experimental dataset for our initial experiments is that many of the words in the glosses have been hand-tagged with their word senses (though see section 6 for an analysis of errors in the tagging). The remainder of the words, however, need to be tagged. We use a set of heuristic strategies to identify the WordNet senses for these words. First, for words that appear in the hand-built TRIPS lexicon, we simply use these TRIPS-WordNet mappings to identify the possible WordNet senses for each TRIPS sense, and then have the parser select the best interpretation in its usual manner, based on syntactic templates possible for each word, the selectional preferences, and finally frequency-based preferences among the senses. For words not in the TRIPS lexicon, we generate lexical entries for a small number of WordNet senses using WordFinder, drawing first from the Core WordNet senses (Boyd-Graber et al, 2006), and/or the most frequent senses (i.e., the first senses listed in WordNet).

4. Building Event Classes from Definitions

Because many glosses are complex, often highly elliptical and hard to parse, we depend on the ability of the TRIPS parser to produce semantically meaningful fragments when a full spanning parse cannot be found. In addition, we apply several strategies to create simplified definitions that are used as backup in case the full definition doesn't parse: These simplifications are

- if the definition starts with "verb or verb ....", truncate the first two words
- If the definition contains "or", "and", or comma, truncate the definition starting at that token

We parse the full definition and any simplifications produced, and then find the fragment or full interpretation that covers the greatest amount of the gloss while producing a definition that is semantically compatible with the target word (e.g., verbs must map to events, adjectives must map to predicates). Note that natural definitions, including those in WordNet, sometimes indicate necessary conditions while at other times indicate necessary and sufficient conditions, and do not reliably signal such cases. For the present, we treat all definitions as specifying only necessary conditions. Because of this, when we define a sense based on only part of its definition, it typically still produces useful knowledge.

We identify the likely arguments (i.e., semantic roles) of the concept using signals in the logical form such as the presence of gaps and the use of a few indefinite pro-forms such as *someone, somewhere, etc*. Note that most roles are **not** explicit in the definition. For example, the definition of *kill*, *cause to die,* does not explicitly express the subject or the object of the cause and the LF recovers this missing information, producing something like *<something> causes <something> to die*. We identify the semantic roles for these arguments by checking the TRIPS lexicon for the roles involved in the verb *cause*, or if there is no explicit entry in the lexicon, we use WordFinder to derive the likely roles by employing the WordNet to TRIPS ontology mapping. In this case, the roles for *kill%2:35:00* would be identified as *AGENT* and *PATIENT.*

To refine the roleset and compute selectional restrictions, we then try to parse the examples provided in WordNet, plus additional examples involving the current word sense being defined from the SEMCOR corpus[3]. These examples provide some evidence as to the range of syntactic templates and semantic roles that can occur with the verb, as well as providing examples of possible fillers. We compute a selectional preference for each role by attempting to find the most common subsumer of all the examples in either the WordNet hypernym

New Concept Name: kill%2:35:00
Roles: AGENT person%1:03:00
PATIENT organism%1:03:00
Definition: LF graph in Figure 2

Figure 3: The information derived for the concept corresponding to kill%2:35:00

[3] http://www.cse.unt.edu/~rada/downloads.html#semcor

hierarchy, or in the TRIPS ontology (and then mapping from this value back to the equivalent WordNet senses). At the end of this first phase of processing the definition, we have derived the information shown in Figure 3 for *kill%2:35:00*.

The next phase is to convert this information into OWL DL. In most cases we are performing a one-to-one mapping from the LF to OWL where concepts in the LF become OWL classes and roles are mapped to corresponding OWL object role restrictions. For example, we begin converting kill%2:35:00 with the selectional preferences by asserting that it is a subclass of the expression: $\forall\text{\_agent.person\%1:03:00} \sqcap \forall\text{\_patient.organism\%1:03:00}$ (i.e., all things that have agents that are person%1:03:00 and have patients that are organism%1:03:00). Note that the we can use the more informative universal restriction instead of an existential because we assume that verbs have at most one of each core role.

Next, we handle the conversion of the LF graph of the gloss shown in Figure 2. We begin at the head of the definition, the CAUSE-EFFECT node, by creating a new OWL class that uniquely represents that node, we will call C1, and assert, $\text{kill\%2:35:00} \sqsubseteq \text{C1}$. Next we define C1 simply as the subclass of the conjunction of its WordNet class, cause%2:36:00, and its semantic restrictions. To translate the :EFFECT role we first create a new class, D1, and then create the object role restriction $\forall\text{\_effect.D1}$. Doing this for each of C1's roles produces the axiom

$$\text{C1} \sqsubseteq \text{cause\%2:36:00} \sqcap \forall\text{\_effect.D1} \sqcap \forall\text{\_patient.R1} \sqcap \forall\text{\_cause.R2}.$$

We then recursively define any new classes; in this example, $\text{D1} \sqsubseteq \text{die\%2:39:00} \sqcap \forall\text{\_experiencer.R1}$, $\text{R1} \sqsubseteq \text{entity\%1:03:00}$, $\text{R2} \sqsubseteq \text{entity\%1:03:00}$.

We next must handle the multiple references to R1. The LF treats each object as a unique instance so when it is referred to more than once in an LF we know that each reference indicates the same instance. When we convert the LF to OWL the objects are no longer instances but are instead classes. In the above example, we no longer have the meaning that the patient and experiencer are the same individual - only that they belong to the same class, R1. In order to capture the intended meaning we introduce an OWL data property called varID which uniquely names the reference. varID acts as an indicator that when the classes are grounded those with the same varID are the same OWL instance. Using this methodology, we have the final set of assertions for the definition:

$$\text{kill\%2:35:00} \sqsubseteq \forall\text{\_agent.person\%1:03:00} \sqcap \forall\text{\_patient.organism\%1:03:00} \sqcap \text{C1}$$
$$\text{C1} \sqsubseteq \text{cause\%2:36:00} \sqcap \forall\text{\_effect.D1} \sqcap \forall\text{patient.(R1} \sqcap \text{varID=''r1'')} \sqcap \forall\text{\_cause.R2}$$
$$\text{D1} \sqsubseteq \text{die\% 2:39:00} \sqcap \forall\text{\_experiencer.(R1} \sqcap \text{varID=''r1'')}$$
$$\text{R1} \sqsubseteq \text{entity\%1:03:00}$$
$$\text{R2} \sqsubseteq \text{entity\%1:03:00}$$

Note that we are using a hierarchical roleset similar to the combining of VerbNet and LIRICS roles as described in Bonial et al (2011), with slight variations in the role names. Specifically, the Agent role is a specialization of the Cause role (i.e., the axiom $\text{agent} \sqsubseteq \text{cause}$ is in the OWL KB), thus we know that the the agent of kill%2:35:00 is the same as the cause role in the definition of C1.

Modifiers are indicated with a relation :MOD (see Figure 4) that indicates the presence of a backlink with semantic meaning but do not add any semantics itself. We remove these cycles and replace them with inverse object roles meant to represent the backlink. For the example, the concept defined in Figure 4 would be a subclass of $\text{die\%2:39:00} \sqcap \exists\text{\_OF}^{-1}\text{.quickly\%4:02:00}$. Notice that modifiers use the less restrictive existential rather than the universal since we do not restrict objects to have only one modifier. This is a very simple example. The same technique works for more complex cases like dealing with relative clauses.

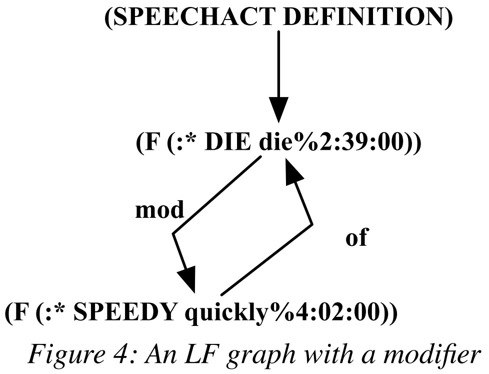

Figure 4: An LF graph with a modifier

	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
# new verb senses	559	255	169	150	99	75	66	41	34	29	15	15	10
# new senses	559	853	988	970	748	543	437	318	230	163	106	64	46

Table 1: The number of new senses introduced with each iteration

Logical operators such as conjunction, disjunction and negation are converted directly into the corresponding OWL operators, allowing the conversion of arbitrarily complex logical forms.

The translation process described above captures enough of the meaning in the LF to support the system described in the rest of the paper but it does not capture all the possible entailments one might be able to derive. In the future, we would like to encode core semantic roles in the gloss (not the ones found in selectional preferences) as the more appropriate exactly-one cardinality constraint coupled with an existential constraint. For instance, ∀_effect.D1 (if there is an effect then it is of type D1) becomes =1_effect.⊤ ⊓ ∃_effect.D1 (there is only one effect and it is of type D1). We are also exploring how to better handle negation in glosses. Consider the gloss for acquitted, “declared not guilty of a specific offense or crime; legally blameless”. What “not guilty” actually indicates is the opposite of guilty, i.e., innocent. While it would be correct to say that the _effect of the declare action is of the class ¬guilty, it isn't very useful. A lot of unrelated things could be ¬guilty: dog, blue, running, etc.

5. Building a Mid-Level Ontology for WordNet Verbs

As mentioned earlier, defining a mid-level ontology was not one of the goals of the WordNet designers. The hierarchical organization of verbs is the troponym hierarchy, which captures *manner* specialization (e.g., *beating* is a type of *striking* which is a type of *touching*). The sense *touch%2:35:00* is a top-level sense and has no more abstract characterization. There are 559 such synsets in WordNet that have no hypernyms, and these concepts range from concepts that would serve as useful primitives (like *touch, breathe)* to more specific senses such as three senses of the verb *keep up* (prevent from going to bed, keep informed, and maintain a required pace or level). The sense of *kill* we have used as an example is also one of the top-level verbs. In addition, over 200 of these verbs have no troponyms either, leaving these sense essentially unrelated hierarchically to any other verbs in WordNet.

The idea underlying this experiment is that we can build a mid-level ontology by reading the glosses of these words. The consequence of this is that each of the previous top-level verb synsets will now have a superclass concept, e.g., kill%2:35:00 will now have a superclass of cause%2:36:00 ⊓ ∃_effect.die%2:30:00 (i.e., “cause to die”) which of course is a specialization of the general class cause%2:36:00. Note that while many linguistic ontologies capture only subclass links between atomic types, we are generating much richer information that captures the definition in terms of a complex type. In this example, we not only have derived a hierarchical relation between kill%2:35:00 and cause%2:36:00, but also the causal relationship between kill%2:35:00 and die%2:30:00.

After this first iteration, we will have introduced a new set of word senses, both verbs and non verbs, that have not yet been defined. So we then iterate using the same procedure on this new set of words to define them. In principle, we continue this iteration as long as new undefined senses are introduced. In the evaluation described below, we stopped after twelve iterations and completed the remaining undefined terms by adding the hypernym chain for the concept. Table 1 shows the number of new senses that were introduced with each iteration. It takes another dozen iterations, each one adding just a few verbs in order to exhaust the generation of new undefined senses. One might think that this continual defining of verb senses would produce a full event hierarchy rooted at some “mother” verb-sense! This does not happen however, because of the presence of cycles in the definitions. Circular definitions “short-circuit” the identification of more abstract classes and tend to collapse sets of synsets together. We examined these circular classes by hand and found that most result from errors in the sense tagging provided in the Princeton WordNet Gloss Corpus. By correcting these tagging errors, we can avoid the unwanted circularities. Other cycles appear to cluster around core definitional primitives that simply are hard to define in any formal decompositional way, and we leave them as they are.

Class	*Definition*
Air%2:32:03	be broadcast%2:32:01
broadcast%2:32:01	**broadcast%2:32:00** over the airwave%1:10:00, as in radio or television%1:06:01
broadcast%2:32:00	cause to become widely known%3:00:00

Table 2: The definitions used to infer that 'airing something' causes it 'to become known'

We discuss our analysis of the cycles generated from processing the top-level WordNet verb classes in a later section. The evaluation examines systems with and without these word sense corrections.

Empirical Evaluation

While we have built a knowledge base containing significant amounts of conceptual information by reading the glosses, here we focus on evaluating just two aspects of this knowledge base. First is the hierarchical relations between the bare WordNet classes, which is a mid-level ontology for WordNet verbs. The second involves causal relationships that can be derived from the knowledge. Some of these are trivial (e.g., *kill%2:35:00* causes *die%2:39:00*), while others are revealed from inference. For instance, the subsumption algorithm will compute that the verb class *air%2:32:03* causes the event of something becoming *known%3:00:00*. There is much more information in this knowledge base than we are going to evaluate here. For instance, it contains knowledge about the changes of state and transitions that serve to define many verbs, and in Allen et al (2011) we demonstrate an ability to perform temporal inference using the knowledge base. But in this paper we focus solely on evaluating just the hierarchical and causal relations between bare WordNet classes in order to enable a direct comparison with WordNet.

We randomly selected 6N (N=8) pairs of verb concepts (A, B) from those that our system successfully processed (columns 0-11 in Table 1), such that at least N of them fell into each of the four categories "{WordNet, our OWL-DL knowledge base} says that A {is a kind of, causes} B", and such that 2N pairs were unrelated in either source. We then presented the pairs in different randomized orders to a set of human judges and asked them to identify whether there was a causal or hierarchical relation between the events, or whether they were unrelated. As judges, we used six researchers who had been involved with the project as well as five people who have no relation to the work. We computed the inter-rater agreement (IRA) using Cohen's kappa score (Cohen, 1960). Kappa was computed for each pair of judges, then averaged to provide a single index of IRA (Light, 1971). The resulting kappa indicated substantial agreement, $\varkappa = 0.63$ (Landis & Koch, 1977). In order to eliminate the cases where their was no consensus among the judges, we only consider the cases in which eight or more judges agreed, which was 83% of the samples, and used the majority decision as the gold standard. We can then evaluate the accuracy of the hand-coded relations in WordNet against two versions of our system: one processing the raw glosses in WordNet and the other with 79 corrected word sense tags out of over 5000 glosses processed.

The precision and recall results are shown in Table 3. The most important property we desire is that the knowledge produced is accurate, i.e., the precision score. This reflects the ability of the systems to produce accurate knowledge from processing glosses. If precision is high, we could always improve recall by processing more definitional sources. We see that the precision scores for the system generated relations are quite good, over 80% for the hypernym relations and a perfect 100% for the causal relations.

Regarding WordNet, we see that the hand-coded relations had a 100% precision, indicating that the structural information in WordNet is highly accurate. The recall numbers, however, show that a significant number of possible relations are missed, especially for causal relations. This suggests that it is worth exploring whether the information implicit in the glosses is redundant given the hand-coding, or whether they serve as an important additional source of knowledge. We can explore this by comparing the sets of relations produced by the system with the relations in WordNet. If they overlap significantly, then the hand-built WordNet relations are fairly complete. If they are disjoint, then the glosses contain an important additional source of these structural relations. The analysis is summarized in Table 4. We look at each relation proposed by WordNet or the system, and look at the overlap

Source	*Hypernym*			*Causal*		
	P	R	F1	P	R	F1
Processing Raw Glosses	80%	33%	47%	100%	36%	53%
Processing Corrected Glosses	83%	42%	56%	100%	55%	71%
Explicit WordNet relations	100%	83%	91%	100%	55%	71%

Table 3: Precision and Recall Scores Against Human Judgement

Relation	*WordNet*	*System*	*count*	*Human Judgement*	
				yes	*no*
Causation	Yes	Yes	1	1	0
	Yes	No	5	5	0
	No	Yes	5	5	0
	No	No	29	0	29
Hypernym	Yes	Yes	3	3	0
	Yes	No	7	7	0
	No	Yes	3	2	1
	No	No	27	0	27

Table 4: Comparing the Redundancy between WordNet & System-generated relations

and disjoint cases. The data show a surprising disjointness between what is explicitly coded in WordNet and the information derived from the glosses. Out of 11 cases of causal relations, there is only one overlap between WordNet and the system, and the remaining relations are equally divided, with five causal relations in WordNet that were not derivable by the system, and five causal relations the system derived that are not coded in WordNet. Thus there is significant causal knowledge derivable from the glosses that is not currently encoded in WordNet. With hypernyms, results are similarly disjoint, with only three out of thirteen cases both encoded in WordNet and derived by the system.

6. Error Analysis

Consider the cases where a hand-coded hypernym relation was not derived from the definitions. In general, the most common reasons for this include problems in parsing and an inability to reason from the provided definitions to the desired entailments. Interestingly, virtually all the errors in the evaluation set are problems the reasoning side. Some of these are because the definitions simply don't provide enough information, and in other cases the system lacked of an ability to resolve vagueness in the definitions. For instance, by failing to make a connection between "deprive of life" and "cause to die", the system misses that *annihilate%2:30:00* is a subclass of *kill%2:35:00*. In another case, it fails to note the relationship between *compose* and *create* due to the definition creating a disjunction that cannot be reasoned through. Specifically, *compose%2:36:01* is found to be a subclass of the class (OR create%2:36:00 construct%2:36:01). In other cases, the conclusion is not found because of sense tagging errors. For instance, the system cannot conclude that *corrupt%2:41:00* is a subclass of *alter%2:30:01* in either version of the system. The system running on uncorrected tags ended in a circular definition of *corrupt%2:41:00*. The system running with corrected tags infers that corrupting is making a mess of someone morally, and cannot relate this to causing a change in someone. As a final example, definitions sometimes involve phrasal verbs that are not defined in WordNet. For instance, *posit%2:32:02* is defined as "put%2:35:00 before" where the system knows nothing about a sense of *put before* as a verb of communication, and this phrasal verb is not defined in WordNet.

The one false positive in the evaluation was when the system derived that *excogitate%2:36:00*, defined by "come up with (an idea, plan, explanation, theory, or principle) after a mental effort", is a subclass of *execute%2:36:00*, defined as "put into effect". This conclusion results from a long chain of reasoning through definitions of *come up with*, to *bring forth*, to *bring*, to *take* and finally to *accomplish%2:36:00*, which is in the same synset as *execute%2:36:00*. It is hard to identify a specific flaw in this chain, but the human judges resoundingly judged this pair as being unrelated.

In general, exploring the results beyond this specific evaluation, the most common problem found was word sense tagging errors, mostly by the system on words that were not tagged in the glosses (and one hand-tagged in the WordNet files). Most of these were light verbs, specifically *have, give* and *put*, and generally the system tagged a more common concrete sense (e.g., *have* as possession) rather than the abstract causal sense (e.g., *have* as causing something). We believe such errors can be reduced by specializing the WSD algorithm to more specifically bias the senses useful in definitions. Other cases arose because the system identified the incorrect semantic roles in the definition, thereby losing the required entailments, and the system has significant problems in getting the right scoping for definitions containing disjunctions. We explore the sense tagging issues in more detail below.

Word Sense Corrections

As mentioned before, the initial, automatically generated ontology contained a number of senses with circular definitions that prevented deriving desired entailments. For example, we have in WordNet the following definition (showing only the relevant sense keys) for the synset *stick%2:35:00***:** (stick%2:35:00 to firmly).

In general, cycles indicate equivalence of the senses involved and logically collapse the synsets into one single class. We manually examined these cycles and determined that many of their definitions had been mis-tagged, and used the follow strategies to break many of the cycles.

- *Selecting an Alternative Sense:* We re-tagged the offending lemma with a different sense of the lemma. In the example of stick%2:35:00:: above, its definition should refer to a more basic sense stick%2:35:01:: (come or be in close contact with; stick or hold together and resist separation)
- *Replacing with a Hypernym:* There may not always be an alternative sense that seems appropriate. We replaced some of these circular senses with their hypernyms. For the circular definition *cast_away%2:40:00: (throw_away%2:40:00 or cast_away%2:40:00),* we replaced both words in the definition with their (common) hypernym: *cast_away%2:40:00: (get_rid_of%2:40:01::)*
- *Unpacking Phrases:* In WordNet phrasal verbs are often defined in entries separate from those of their head verbs. For example, *go_into%2:42:00* has its own definition *(be used or required for).* However, WordNet also includes an entry for the non-phrasal-verb sense of "go into" *go_into%2:38:00: (to come or go_into%2:38:00).* In this second example, "go into" literally means "go" + "into". We broke the phrase into these two components in the definition: *go_into%2:38:00: (to come or go%2:38:00 into)*
- *Simplifying Definitions:* Some definitions contain elaborate, detailed and slightly redundant information. For example: *pronounce%2:32:01: (speak, pronounce%2:32:01, or utter in a certain way)* Logically, with one of the disjuncts being identical to the sense being defined, the definition is vacuous. However, here "speak", "pronounce" and "utter" are closely related. We could break the cycle by deleting "pronounce" in the definition. Arguably this strategy could lose some information, but we only apply this simplification when the disjunct is nearly synonymous with some of the other elements in the definition.

There remain, however, some cycles that represent core concepts not easily reducible to other even more basic concepts. For example, the four-synset cycle containing

change%2:30:00 < undergo%2:39:04 < pass%2:38:00 < go%2:38:00 < change%2:30:00

are all related to the concept of change. We elected not to contrive a re-definition but rather leave these cycles in place. Such cycles are prime candidates for core concepts that would benefit from being hand axiomatized in an upper ontology.

7. Discussion

We have described initial steps in constructing common-sense knowledge bases by reading word definitions. The focus of this work is to derive conceptual knowledge, i.e., definitions of concepts associated with word senses, to facilitate deeper language understanding. This stands in contrast to much current work on learning by reading, which is focused on building surface level word/phrase relationships. For instance, Etzioni et al (2011) have an impressive system that scans the web and extracts surface patterns such as (Starbucks, has, a new logo). NELL (Carlson et al, 2010) derives similar knowledge by learning extraction patterns for a predefined set of relations. Neither of these systems attempt to disambiguate word senses or construct definitional knowledge. The evaluation is performed

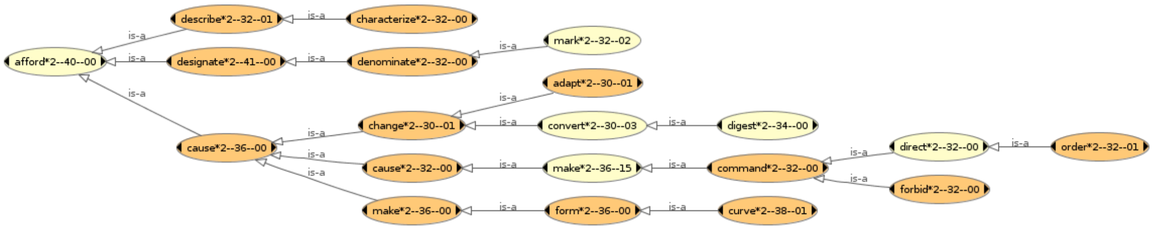

Figure 5: A fragment of the event hierarchy derived from the glosses

by human judges who, of course, used their ability to understand natural language in order to validate the data (e.g.., picking word senses that make sense).

As a demonstration of the promise of our techniques, we have shown that we can construct a mid-level ontology for WordNet verbs from the WordNet glosses, starting from the 559 verb senses in WordNet that have no hypernym. We evaluate the results using human judges comparing relations between word senses in WordNet, where each sense is carefully defined in the evaluation. We have shown that the knowledge we derive is not only quite accurate, but is substantially different from the information already in the explicitly defined WordNet relations (e.g., hypernym and cause relations). As such, our techniques have the potential to produce an expanded set of WordNet style relations that could be very useful for improving current techniques that use WordNet as a source of entailments.

Most prior work linking WordNet to ontologies has involved producing mappings from the synsets into an upper ontology, without developing the intermediate detail. For instance, SUMO has a comprehensive mapping from WordNet to its upper ontology, but 670 WordNet verb synsets are mapped to the single SUMO class *IntentionalProcess* (3 equivalences and 667 subsumptions), including senses as diverse as *postdate* (establish something as being later relative to something else), *average* (achieve or reach on average), *plug* (persist in working hard), *diet* (follow a regimen or a diet, as for health reasons), *curtain off* (separate by means of a curtain) and *capture* (succeed in representing or expressing something intangible). While these links connect WordNet into SUMO, they don't provide significant extra knowledge to enable entailments. Our work can provide links to an upper ontology with significant additional structure providing an opportunity for entailment. As an example, Figure 5 shows a small part of the derived ontology. This encodes such information like *forbidding* is a form of *commanding*, which involves *making someone do something*, which itself is a form of *causation*. With each of the concepts along this chain having a detailed definition in the style described in Section 4, we can use reasoning systems developed for OWL-DL to draw a rich set of entailments about the consequences of performing a *forbidding* act.

Much remains to be done to realize our dream of building rich knowledge bases by reading. There are short term issues and longer term issues. On the short term, the biggest improvement would result from improving word sense disambiguation, especially for the light verbs such as *have* and *go*. It is not a coincidence that these verbs generally are not tagged in the Princeton Gloss corpus. They are difficult to tag, and it is not clear that the senses offered in WordNet always provide the right set of choices. We are considering special processing of these abstract senses, possibly encoding them directly in a hand-built upper ontology. In the longer term, we need to expand our evaluation methods to verify that the knowledge derived beyond hypernym and causal relations is accurate and useful. This will presumably involve more complex entailment tests. Finally, in the long run, we do not believe that effective knowledge bases can be derived entirely from processing individual definitions without some inferentially-based "knowledge cleaning" where raw knowledge is combined from several sources, abstracted and revised in order to create more consistent and coherent knowledge.

8. Acknowledgements

This work was supported in part by the National Science Foundation under grant 1012205, by DARPA grant 1012205 FA8750-09-C-0179 in the Machine Reading program, and ONR grant N000141110417.

9. References

Agerri, R., Anselmo Peñas (2010) On the Automatic Generation of Intermediate Logic Forms for WordNet Glosses. CICLing 2010: 26-37.

Allen, J., W. de Beaumont, N. Blaylock, G. Ferguson, J. Orfan, M. Swift (2011) Acquiring Commonsense Knowledge for a Cognitive Agent. AAAI Advances in Cognitive Systems (ACS 2011), Arlington, VA.

Allen, J., M. Swift and W. de Beaumont. Deep Semantic Analysis for Text Processing. Symposium on Semantics in Systems for Text Processing (STEP 2008) Shared Task: Comparing Semantic Representations. Venice, Italy, September 22-24, 2008.

Bateman, J.A., B. Magnini, and G. Fabris (1995) The generalized upper model knowledge base: Organization and use. In N.J.I. Mars (Ed.). Towards very large knowledge bases: Knowledge building and knowledge sharing. Amsterdam: IOS Press.

Bonial, C., Brown, S.W., Corvey, W., Palmer, M., Petukhova, V., and Bunt, H. (2011). An Exploratory Comparison of Thematic Roles in VerbNet and LIRICS, Proceedings of the Sixth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-6).

Boyd-Graber, J., Fellbaum, C., Osherson, D., and Schapire, R. (2006). "Adding dense, weighted connections to WordNet." In: Proceedings of the Third Global WordNet Meeting, Jeju Island, Korea, January 2006

Carlson, A. Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In AAAI'10, 2010.

Clark, P., Fellbaum, C., Hobbs, J.R., Harrison, P., Murray, W.R., Thompson, J. (2008) Augmenting WordNet for Deep Understanding of Text. In: Bos, J., Delmonte, R. (eds.) Semantics in Text Processing. STEP 2008 Conference Proceedings. Research in Computational Semantics, vol. 1. College Publications.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.

Etzioni, O., A. Fader, J. Christiansen, S. Soderland, and Mausam. Open Information Extraction: The next generation, IJCAI, 2011.

Fellbaum, S. (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Frank, A. (2004) Constraint-based RMRS construction from shallow grammars. COLING-2004, Geneva.

Gangemi, A., N. Guarino, C. Masolo, A. Oltramari, and L. Schneider (2002). Sweetening ontologies with DOLCE. In A. Gómez-Pérez and V. Benjamins (Eds.), Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, Berlin, Heidelberg: Springer Berlin.

Gangemi, A., Navigli, R., Velardi, P. Axiomatizing WordNet Glosses in the OntoWordNet Project (2003) Workshop on Human Language Technology for the Semantic Web and Web Services, 2nd International Semantic Web Conference (ISWC2003). Sanibel Island, Florida.

Harabagiu, S.M., Miller, G.A., Moldovan, D.I. (2003): eXtended WordNet - A Morphologically and Semantically Enhanced Resource, http://xwn.hlt.utdallas.edu.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Lenat, D. B. (1995): Cyc: A Large-Scale Investment in Knowledge Infrastructure. The Communications of the ACM 38(11):33-38.

McCarthy, D., R. Koeling, J. Weeds, and J. Carroll, (2004) Finding predominant senses in untagged text. In Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain..

Mehdi H. Manshadi, James Allen, Mary Swift, "Towards a Universal Underspecified Semantic Representation", Proc. 13th Conference on Formal Grammar (FG 2008), Hamburg, Germany, August 9-10, 2008

Nichols, E., F. Bond, and D. Flickinger, Robust ontology acquisition from machine-readable dictionaries, IJCAI-2005.

Niles, I. and Pease, A. Towards a Standard Upper Ontology (2001) In Proc. 2nd International Conf. on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine.

Pulman, S. G. (1983) Word meaning and belief. London: Croom Helm.

Vendler, Z. Verbs and Times. The Philosophical Review, Vol. 66, No. 2. (Apr., 1957), pp. 143-160.

Vossen, P. (Ed.). (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.