Learning Semantically Rich Event Inference Rules Using Definition of Verbs

Nasrin Mostafazadeh[1] and James F. Allen[1,2]

[1] Computer Science Department, University of Rochester, Rochester, New York
[2] Institute for Human and Machine Cognition, Pensacola, Florida
{nasrinm,james}@cs.rochester.edu

Abstract. Natural language understanding is a key requirement for many NLP tasks. Deep language understanding, which enables inference, requires systems that have large amounts of knowledge enabling them to connect natural language to the concepts of the world. We present a novel attempt to automatically acquire conceptual knowledge about events in the form of inference rules by reading verb definitions. We learn semantically rich inference rules which can be actively chained together in order to provide deeper understanding of conceptual events. We show that the acquired knowledge is precise and informative which can be potentially employed in different NLP tasks which require language understanding.

1 Introduction

Systems (Allen et al., 2011) performing NLP tasks such as Question Answering (QA), Recognizing Textual Entailment (RTE) and reading comprehension depend on extensive language understanding techniques to function. Deep language understanding enables an intelligent agent to construct a coherent representation of the scene intended to be conveyed through natural language utterances, connecting natural language to the concepts of the world. Developing a deep understanding system requires large amounts of conceptual and common-sense understanding of the world. As an example, consider a QA system which is given the question in Figure 1. One pre-requisite for answering this question is to semantically understand and interpret both query and the snippet. Figure 1 shows a generic semantic interpretation of the question and the snippet with grey labels. Throughout this paper we use the verbal semantic roles[1] as distinguished by TRIPS system (Allen et al., 2005).

After the semantic interpretation, the system understands that it should look for a *kill* event with Einstein as the *affected* person. However, the system does not see any explicit connection between the event in the question and the event presented in the snippet. Now let us provide the system with the following piece of knowledge in the form of an inference rule about the event *kill*:

$$(X_{agent}\ \textbf{kills}\ Y_{affected}) \xrightarrow{entails} (X_{agent}\ \textbf{causes}\ Y_{affected}\ to\ die) \qquad (1)$$

[1] http://trips.ihmc.us/parser/LFDocumentation.pdf

A. Gelbukh (Ed.): CICLing 2015, Part I, LNCS 9041, pp. 402–416, 2015.
DOI: 10.1007/978-3-319-18111-0_30

By having access to such an inference rule, the system will know that 'killing' entails 'cause to die', where explicitly the 'killer' causes the 'affected' to die. One can imagine many more complex pieces of knowledge presented in the form of inference rules, each of which can provide a new clue for a system which requires language understanding. It is obvious that a system should have various natural language processing capabilities in order to successfully answer questions, however, here we focus on the bottleneck of conceptual knowledge on events.

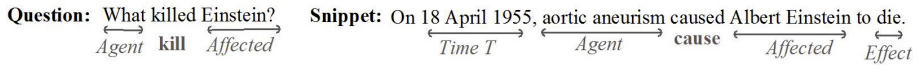

Fig. 1. Example question and its corresponding relevant information posed to a question answering system

As the earlier example shows, having conceptual knowledge about the events in the form of semantically rich inference rules – such as knowing what happens to the participants before and after it occurs or the consequences of the event – can play a major role in language understanding in different NLP applications. We believe that an effective conceptual knowledge should provide semantic reasoning capabilities, with semantic roles and sense disambiguation. In this paper, we introduce a novel attempt to automatically learn a semantically rich knowledge base for events, which provides high precision inference rules aligned by their semantic roles. We propose to learn the knowledge base by automatically processing large amounts of definitional knowledge about verbs, using their WordNet (Miller, 1995) word sense definitions (glosses). We accomplish this by deep semantic parsing of glosses, automatic extraction of inference rules, unsupervised alignment of semantic role labels, and chaining inference rules together until hitting a 'core' concept.

The phases of our approach are shown in Figure 2. We will provide details about each of these phases in Sections 2-4. The main outcome of our approach is the Inference Rules corpus which could be used for different language understanding tasks. In Section 5 we show that our semantic role alignment methodology is a promising way for acquiring precise and semantically rich inference rules. Moreover, we show that the inference rules acquired by our approach have higher precision than any other related work. Although we use WordNet here, our approach is applicable to any other definitional resources.

2 Deep Semantic Parsing of Definitions

As the first phase of our approach, we need to have deep semantic understanding of the verb definitions. Here we use the TRIPS broad-coverage semantic parser[2] which produces state-of-the-art logical form (LF) from natural language text

[2] `http://trips.ihmc.us/parser/cgi/parse`

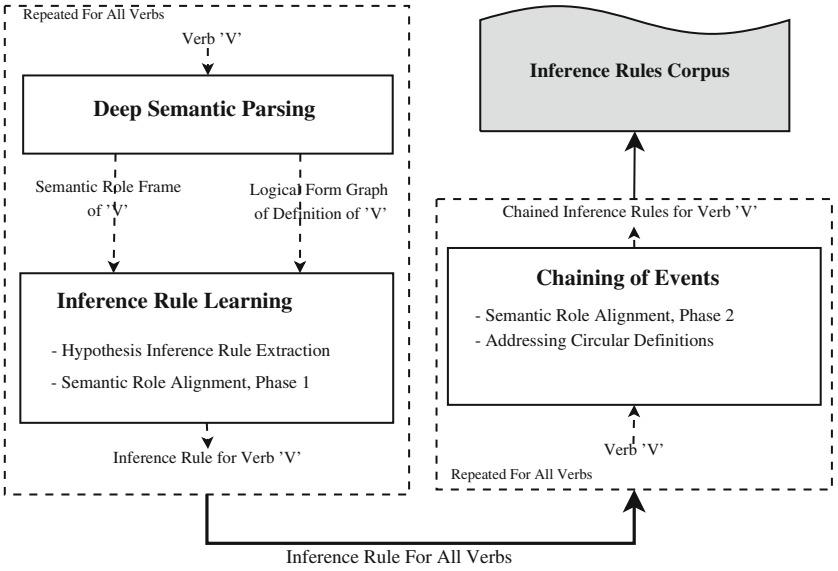

Fig. 2. The phases of our approach

(Allen et al., 2005). TRIPS provides an essential processing boost beyond other off-the-shelf applications, mainly sense disambiguated and semantically rich deep structures. The approaches presented in this paper can be applied to any other wide-coverage semantic parsers, such as Boxer system (Bos, 2008).

Many glosses are complex, often highly elliptical and hard to parse. For example, 'kill.v.1'[3] is defined as 'cause to die' which does not explicitly mention the subject or the object of the sentence. Another example is 'love.v.1' which has the gloss 'have a great affection or liking for', where the object of the sentence is missing. TRIPS recovers such missing information, producing a parse such as 'something causes something to die' (Allen et al., 2013) for the gloss of 'kill.v.1'. The output of this phase is the semantic role frame[4] (semframe) for each verb synset together with LF graph of its gloss. For instance, the semframe of the verb *kill.v.1* is given as $\{agent_{ont:person.n.1}, affected_{ont:organism.n.1}\}$. Figure 3 shows the simplified LF graph of the gloss of 'kill.v.1'.

3 Inference Rule Learning

In the second phase of our approach we aim to extract semantically rich inference rules for all verb synsets.

[3] We represent WordNet words sense disambiguated using their part of speech and sense number. So 'kill.v.1' is the first sense of the verb 'kill'.

[4] Semantic role frame is called to the set of semantic roles associated with a verb.

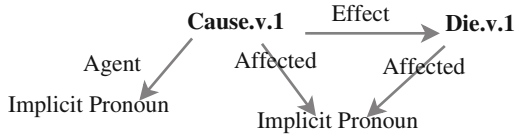

Fig. 3. Logical form produced by TRIPS for 'kill.v.1'

3.1 Hypothesis Inference Rule Extraction

Hypothesis inference rules are preliminary rules which are extracted using the two outputs of the deep semantic parsing phase. A hypothesis rule is an axiom with Left Hand Side (LHS) and Right Hand Side (RHS), each consisted of predicates where LHS logically entails RHS. There is always one predicate on the LHS, but there could be more than one predicates on the RHS (one of which is the root predicate, marked with '*'). LHS predicate comes from the semframe of the verb and the RHS predicates come from the LF graph of the verb's definition. We define a predicate to be either a verb or verb nominalization, as they inherently have the potential to occur at some time point as events. Here we stick to a very simple logical representation of axioms in the form of inference rules, which enables easy incorporation of our knowledge base in various systems. For instance, the following is the hypothesis rule that we deterministically extract for the verb 'kill.v.1':

$$(\mathbf{kill.v.1}\ X_{agent}\ Y_{affected}) \Rightarrow (\mathbf{cause.v.1}\ A_{agent}\ B_{affected}\ C_{effect})^* \\ \wedge(\mathbf{die.v.1}_C\ B_{affected}) \quad (2)$$

where each predicate is enclosed within parenthesis which has some arguments (semantic roles) realized with either variables or constants. As you can see, a predicate itself can be an argument of some other predicate, e.g., in the earlier example 'die.v.1' (reified with variable *C*) is the *effect* role of 'cause.v.1'. We call the set of hypothesis inference rules for all the WordNet verb sysnsets $corpus_{hypothesis}$. Now the question is whether a hypothesis rule is usable as an inference rule. The answer is that we often do not get a LF graph with all of the roles recognized correctly; and even if we do, more importantly we still do not know which role on the LHS corresponds to a role on RHS. This issue motivates 'semantic role alignment'.

3.2 Semantic Role Alignment, Phase 1

We want to know whether or not it is always the case that the *agent* role in the LHS of a rule maps to the *agent* role in the RHS and they should have the same realization. What happens to the *agent* role of LHS in case there is no *agent* role on the RHS? We call the problem of mapping the roles of the LHS to the roles of RHS 'Semantic Role Alignment' (SRA). The machine translation (MT) community has established an extensive literature on word alignment (Brown et al., 1993; Och and Ney, 2003), where translating 'she came' into French sentence 'elle est venue' requires an alignment between 'she' and 'elle', and between 'came' and

'est venue'. We believe that MT alignment approaches are suitable for the SRA task because of the following reasons:

- The semantic roles on the LHS and RHS tend to have semantic equivalence. So it is intrinsically the case that there is a (partial) mapping from roles on the RHS to the roles on the LHS.
- As opposed to the kind of inference rules learned based on distributional similarity (Harris, 1985) (to be discussed in Section 6), here the semantic content of LHS should not diverge substantially from RHS given the fact that RHS is basically defining LHS.
- MT alignment models are typically trained in an unsupervised manner, depending on sentence-aligned parallel corpora. For our task large volumes of training data are lacking, so an unsupervised training (to be explained in this section) is the most suitable approach.
- As it will be discussed in Section5, unsupervised aligners (which find hidden structures in data) can actually account for some frequent parsing errors in our system, which is very promising.

We model the SRA problem as a maximum bipartite matching problem: for each inference rule, we define n_{lhs} as the set of nodes such as l_i, each of which corresponds to a role in LHS; n_{rhs} is another set of nodes such as r_j, each of which corresponds to a role in RHS. Each pair of nodes (l_i, r_j) has an edge connecting them, which is weighted with the plausibility of the alignment of that pair. An alignment function a is defined as follows:

$$a : n_{lhs} \rightarrow n_{rhs} \cup \{null\}$$

which is a function mapping each role $\in n_{lhs}$ to a role $\in n_{rhs}$ or a null symbol, similar to IBM-style machine translation model (Brown et al., 1993). Here, mapping a LHS role to *null* means that the role should be 'inserted' in the RHS. Then the SRA problem is considered as a maximum weighted matching problem where the best alignment for the inference rule is the highest scoring a^*, under the constraint of 'one-to-one' matching, which is defined as follows:

$$a^* = \arg\max_a \{score(n_{lhs}, a, n_{rhs})\}$$

$$score(n_{lhs}, a, n_{rhs}) = \sum_{\substack{l_i \in n_{lhs} \\ r_j \in n_{rhs}}} score(l_i, a, r_j)$$

$$score(l_i, a, r_j) = log(Pr(l_i, a|r_j))$$

The training of the probability of aligning a role on LHS to a role on RHS, $Pr(l_i, a|r_j)$, is accomplished using the Expectation Maximization (EM) algorithm (Brown et al., 1993). In the E-step the expected counts for each role pair (l_i, r_j) are calculated and in M-step we normalize and maximize. We mainly estimate the so-called translation probability parameter $t(r|l)$ (Brown et al., 1993).

In order to prepare the data for performing the alignment explained above, we should firstly build an appropriate parallel corpus. Our idea is to build a corpus of LHS roles parallel with RHS roles from the set of hypothesis inference rules for all verbs in $corpus_{hypothesis}$. One issue to consider is that the rules with multi-predicate RHS cannot have a two-sided mapping. Among all 13,249 hypothesis inference rules that we generate, 649 one of them have two RHS predicates and only 10 of them have three RHS predicates. As the first step, we remove all the rules with multi-predicate RHS – about 0.4% of all the rules. With the remaining rules, we build a corpus of all LHS roles parallel with the RHS roles. We call this $corpus_{unary}$. Then we apply the alignment algorithm explained earlier to this corpus for learning the model parameters. Using the learned parameters, for each hypothesis inference rule we find the maximum weighted alignment. As an example, consider the verb *digest.v.3* which is defined as 'to tolerate something or somebody unpleasant'. The hypothesis inference rule produced for this verb is as follows:

$$(\textbf{digest.v.3}\ X_{pivot}\ Y_{affected}) \Rightarrow (\textbf{tolerate.v.4}\ A_{pivot}\ B_{theme})^* \tag{3}$$

Figure 4 shows the bipartite matching graph for this inference rule. The maximum weighted matching is shown by the dark edges. As a result of maximum weighted matching, the aligned inference rule for 'digest.v.3' is the following:

$$(\textbf{digest.v.3}\ X_{pivot}\ Y_{affected}) \Rightarrow (\textbf{tolerate.v.4}\ X_{pivot}\ Y_{affected})^* \tag{4}$$

We evaluate the outcome of this experiment, called '$phase1_{unary}$', in Section 5.

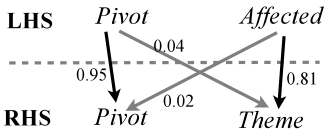

Fig. 4. The bipartite matching graph for alignment of inference rule 3

Our approach for SRA of the inference rules with multi-predicate RHS is linguistically motivated by the fact that the root predicate captures the core semantic meaning of the LHS. In short, our approach is as follows:

- Step 1: Discard the non-root RHS predicate and find the maximum weighted matching between LHS and the root RHS[5].
- Step 2: Make a set of nodes from the LHS roles which are matched to NULL in Step 1. Use this set as a new LHS, then find the maximum weighted matching to the non-root predicate.

[5] It is evident that a roles which is realized with the reification of another predicate, as with the *effect* role in (2), does not take part in the alignment problem.

This approach can generalized as a recursive SRA for rules which have more than two RHS predicates. The results of this experiment, named '$phase1_{bin}$', can be reviewed in Section 5. Applying this alignment approach the inference rule (2) results in the following aligned rule:

$$(\textbf{kill.v.1}\ X_{agent}\ Y_{affected}) \Rightarrow (\textbf{cause.v.1}\ X_{agent}\ Y_{affected}\ C_{effect})^* \\ \wedge(\textbf{die.v.1}_C\ Y_{affected}) \quad (5)$$

At the end of this phase, we will have an aligned and ready to use level-1 inference rule generated for each WordNet verb synset. We call this collection $corpus_{level-1-rules}$. We associate a score with each inference rule, which is its normalized weighted matching score.

4 Chaining of Events

Given the inference rule for each verb from the previous phase, we want to expand our understanding of each event by chaining verbs together. For example, consider a QA system that has encountered the sentence "the boy skinned his knee when he fell" and wants to know more about the concept of 'skinning' by looking up the verb 'skin.v.2' in our knowledge base. The ideal information that we would like to be able to get by forward chaining of the level-1 inference rules is as follows:

$\mathcal{X}$ **skin.v.2** $\mathcal{Y}$:
$\xrightarrow{means}$ $\mathcal{X}$ **bruise.v.1** the [skin] of $\mathcal{Y}$
$\xrightarrow{means}$ $\mathcal{X}$ **injure.v.1** [the underlying soft tissue] of the [skin] of $\mathcal{Y}$
$\xrightarrow{means}$ $\mathcal{X}$ **cause.v.1** [harm] to the [underlying soft tissue] of the [skin] of $\mathcal{Y}$

Obtaining the above chaining requires yet another phase of role alignment, going from each level to the next one and expanding each predicate on the RHS.

4.1 Semantic Role Alignment, Phase 2

Consider the inference rule (5) which we obtained in the previous section. For the expansion of 'die.v.1' on the RHS, we will use its inference rule which is as follows:

$$(\textbf{die.v.1}\ X_{agent}) \Rightarrow (\textbf{lose.v.1}\ X_{agent}\ bodily\_attributes_{theme}) \quad (6)$$

As you can see the semframe of 'die.v.1' in inference rule (6) does not match its semframe in inference rule (5). There are many cases similar to this one and the reason is that semantic parsing and sense disambiguation are not perfect and are error prone. Moreover, verbs can have different semframes in different contexts. Here we perform semantic role alignment phase 2, using a similar method to '$phase1_{unary}$'. This time we build a corpus of all LHS definitions parallel with any of their usages in the entire $corpus_{level-1-rules}$. We call this new corpus $corpus_{def-use}$. EM can find hidden error patterns here as well as actual semantic

alignment patterns. The results of this experiment named '$phase2_{EM}$' can be reviewed in Section 5.

After finding the maximum weighted matching using the trained parameters, we get a new inference rule proper for continuing the forward chaining on 'kill.v.1' which is as follows:

$$(\textbf{die.v.1}\ X_{affected}) \Rightarrow (\textbf{lose.v.1}\ X_{affected}\ bodily\_attributes_{theme}) \quad (7)$$

We have also obtained a probabilistic distribution on semframes for each synset given context, using the parsed glosses. We used this statistics together with EM alignment for favoring a specific semframe over another, resulting in a higher precision alignment in phase 2. The results of this experiment named '$phase2_{EM+}$' is reported in section 5.

The second phase of semantic role alignment results in high precision chaining and on average we get 10 new inference rules with high matching score after three levels of chaining – which increases the size of our inference rules corpus by an order of magnitude. For instance consider the verb 'kill.v.14', which is defined as 'to cause to cease operating'. This verb has three RHS predicates: cause, cease and operate and could have 2^3 different expansions just for the first level.

4.2 Addressing Circular Definitions

Usually there are some circular definitions for words in definitional resources including WordNet (Allen et al., 2011; Ide and Vronis, 1994). For example, the synset 'cause.v.1' is defined as 'cause.v.1 to happen.v.1' which is an immediate circulation. There have been some preliminary strategies (Allen et al., 2011) for breaking the definition cycles. Those findings show that some cycles can be resolved by selecting an alternative sense for the cyclical definition or simplifying the definitions. However, there are some key cycles which cannot be broken in this manner because there is essentially no specific simpler definition for some concepts, e.g., 'cause'. This is an essential problem with machine understanding, because machines have no direct experience with the world, which could have enabled them understand what a natural concept means.

This issue brings up an important psycholinguistic research, where it is believed that human lexicon is a complicated web of semantically related nodes instead of a one-to-one mapping of concepts to the words (Levary et al., 2012). According to earlier work, dictionaries have a set of highly interconnected nodes from which all other words can be defined (Picard et al., 2009). To our knowledge, there has not been any research on finding core concepts on WordNet verbs, using the graph theory experiments. Continuing the work on building dictionary graphs (Levary et al., 2012), we built a graph where directed links are drawn from a word to the words in its definition. In this graph, we found the strongly connected components using Tarjan's algorithm (Tarjan, 1972), which resulted in a set of 56 strongly connected components with size bigger than 1, which included 158 verb synsets. Other definitional paths of WordNet verbs converge to this set quickly, which we call the core verbs. Our idea is to stop forward chaining of definitions (avoiding circulation

trap) when we hit a core verb. The main core concepts that we have identified are as follows: cause, make, be, do, stop, start, begin, end, have, prevent, enable, disable.

After chaining of events and SRA phase 2, we obtain our final corpus of Inference Rules, containing new rules derived from chaining started from level-1 rules and going up to higher levels. We assign a score to each inference rule in different levels of the final corpus, which is the sum of normalized weighted matching scores divided to the number of levels.

5 Evaluation and Results

We have conducted two focused experiments for evaluating the two major contributions of our approach.

Semantic Role Alignment: We attempted to build a gold-standard corpus on semantic role alignment. For annotators, we used seven linguistics experts who had no relation to the work and three researchers who were involved. For each individual annotator, we randomly sampled 100 hypothesis inference rules from the $corpus_{hypothesis}$, and asked them to perform the role alignment for the given hypothesis rule[6]. The role alignment task was either of the following actions towards each RHS role:

- *Substitute*: substitute the role with one of the LHS roles (also decide about the realization value). This action corresponds to a role matching from LHS to RHS.
- *Delete:* Completely remove the role.

Moreover, they had the option of performing *Add* action, which involves adding a new role on RHS. This action corresponds to matching a LHS role with NULL.

The annotators were also asked to assign a confidence score (out of three) to the resultant aligned inference rule. This score takes into account the cases in which there is really no good alignment, and the annotator feels that his/her best possible alignment is not good at all[7]. We used this gold standard for computing precision scores for the SRA Phase 1 methods: $phase1_{base}$, $phase1_{unary}$, and $phase1_{bin}$. As it is critical to get the exact output, we used strict evaluation with no partial credit. We performed the same procedure on $corpus_{def-use}$, and built a gold-standard for evaluating the precision of SRA Phase methods: $phase2_{base}$, $phase2_{EM}$, and $phase2^{+}_{EM+}$). Both $phase1_{base}$ and $phase2_{base}$ are baselines which deterministically align LHS roles with the same RHS roles[8]. The results of these experiments reporting precision and average confidence score is presented in Table 1. In this table, Sc_{err} is the average annotator confidence score on incorrect alignments and Sc_{corr} is the average annotator confidence score on correct alignments of the corresponding method.

[6] We presented each rule together with some example usages of the synset, to give the annotators the context (Szpektor et al., 2007).

[7] This mostly happens for vague definitions or essential parsing errors.

[8] For 78% of all verb synsets we could find an exact name-based role match going from LHS to RHS.

Table 1. Semantic role alignment evaluation results

Method	$phase1_{base}$	$phase1_{unary}$	$phase1_{bin}$	$phase2_{base}$	$phase2_{EM}$	$phase2_{EM+}$
Precision	51%	**90%**	**87%**	10%	72%	79%
Sc_{corr}	3.0	2.7	2.5	3.0	2.28	2.32
Sc_{err}	2.5	2.3	1.2	2.7	1.3	1.4

The results show that the alignment using EM performs very well, providing promising framework for the task of semantic role alignment. The points that the system has missed are mostly for 'Delete' actions (91% of the time) of the annotators. System prefers not to delete any piece of information from the RHS, as it might be necessary for next chaining levels. However, there are some roles on the RHS which are artifacts of bad parse or inconsistent definitions, which annotator can pinpoint but the system cannot. Parsing artifacts are quite easy to be corrected by human, so the average confidence score on those errors is high, which has resulted in pretty high Sc_{err}.

The results obtained for $phase1_{bin}$ show that the alignment on binary rules (which initially seemed more complex) performs as good as the alignment on unary rules. Our observations show that this is because of the fact that many of binary rules are composed of a core predicate such as 'cause', 'stop', or 'do' which all have a recurring usage pattern, making the unsupervised alignment more successful. The baseline $phase1_{base}$ performs mediocre as a simple alignment method for phase1. Phase 2 alignment is always more complicated than phase 1. The baseline $phase2_{base}$ performs very poorly because verbs are mostly used (in context) with different semframe as compared with the semframe they are defined with (out of the context). The method $phase2_{EM}$ performs good, but is not enough for handling complicated alignments in def-use cases. The low Sc_{err} for phase 2 methods indicate the complexity of alignment task at this phase. $Phase2_{EM+}$ outperforms $phase2_{EM}$, which is mainly because it better predicts the cases of occasional bad parsing.

Inference Rules Corpus: To our knowledge, none of the earlier works on acquiring inference rules (details in Section 6) have inference rules with complex and semantically rich semantic roles and sense disambiguation as we do. Hence, in order to compare our inference rules to earlier works we simplify our inference rules dataset, removing all the sense tags and semantic roles. Here we use the most recent manually created verb inference rules dataset (Weisman et al., 2012), hereafter, test-set. This test-set is created by randomly sampling 50 common verbs in the Reuters corpus, and is then randomly paired with 20 most similar verbs according to the Lin similarity measure (Lin, 1998). This dataset includes 812 verb pairs, which are manually annotated by the authors as representing a valid entailment rule or not. They have used rule-based approach for annotation of entailment, where a rule $v1 \rightarrow v2$ is annotated 'yes' if the annotator could think of plausible contexts under which the rule holds (Szpektor et al., 2004). In this dataset 225 verb pairs are labeled as entailing and 587 verb pairs were labeled as non-entailing. Although this dataset is not very rich, it is a good testbed for comparing our inference rules against the state-of-the-art work on verb inference rules. Table 2 shows the results of the following methods:

- $Semantic-Rules_{simplified}$ is our simplified approach: given our final Inference Rules corpus (containing rules up to three levels of chaining or until hitting a core concept), simplify the rules by removing all the semantic roles, all the sense tags, and introduce a new rule for each of the predicates of a multi-predicate RHS. Given a pair (v_1, v_2) from the test-set, if the entailment $v_1 \rightarrow v_2$ exists in the simplified corpus, classify the pair as 'yes'.
- $Supervised_{linguistically-motivated}$ is the work on supervised learning of verb inference rules from linguistically-motivated evidence (Weisman et al., 2012).
- $VerbOcean_{KB}$ is the method that classifies a given pair as 'yes' if the pair appears in the strength relation in the VerbOcean knowledge-base (Chklovski and Pantel, 2004).
- *Random* is the method that randomly classifies a pair as 'yes' with a probability 27.7%, proportional to the number of 'yes' instances in the test-set against the number of 'no' instances.

Table 2. Evaluation results on hand annotated verb entailment pairs test-set

Method	Precision	Recall	F1-Score
$Semantic-Rules_{simplified}$	**50.0%**	45.1%	0.47
$Supervised_{linguistically-motivated}$	40.2%	71.0%	0.51
$VerbOcean_{KB}$	33.1%	14.8%	0.2
Random	27.9%	28.8%	0.28

As the results show, our simplified method outperforms the best method by 10% in precision. This reveals that the accuracy of our inference rules is high and our approach is capable of acquiring more precise verb inferences than the other methods. As expected, our coverage is lower than the *Supervised* method, which is due to the fact that we acquire our rules by reading verb definition and not by mining significantly large web-scale corpora, resulting in a smaller-scale dataset. However, our recall outperforms the $VerbOcean$ method and has also a competing F-1 score compared with the *Supervised* method. Of course for a successful usage of a knowledge base in an application, accuracy is crucial and coverage can be mitigated by using various kinds of precise knowledge bases. A large but noisy and unreliable knowledge base will be of little use in reasoning.

Analyzing the pairs that we have miss-classified as 'yes', there are many pairs which do not seem to be correctly annotated as 'no' in the test-set, such as (reveal, disclose) and (require, demand), where we argue that according to rule-based approach one can indeed think of a reasonable context under which *reveal* $\rightarrow$ *disclose* and *require* $\rightarrow$ *demand* hold. Another example is the pair (stop, prevent), which we classify as 'yes' in the context of the sixth sense of the verb 'stop', but is classified as 'no' in the test-set as the verbs in the test-set are not sense-disambiguated and do not have any context. Overall, our simplified approach proves to be competent with other works and also outperforms the state-of-the-art in precision, which is very promising.

6 Related Work

Early research has shown that definitions in online resources (such as dictionaries and lexicons) contain the type of knowledge that systems can benefit from for conceptual understanding of the world (Ide and Vronis, 1994). More specifically, WordNet's glosses have substantial world knowledge that could leverage semantic interpretation of text (Clark et al., 2008). Some earlier works (Moldovan and Rus, 2001; Clark et al., 2008) have tackled the problem of encoding WordNet glosses as axioms in first-order logic. These works use syntactically processed glosses for extracting the logical information (e.g., they map the NP in a subject position to an *agent* role), and successfully incorporate these axioms in QA and RTE tasks (Moldovan and Rus, 2001; Clark et al., 2008). However, their syntactic representation limits the functionality of semantic representation. Semantically rich logical representations (as opposed to syntactic ones) are proven to perform better on textual similarity and understanding tasks (Blanco and Moldovan, 2013).

The recent work on Multilingual eXtended WordNet (Erekhinskaya et al., 2014) attempts to semantically parse the glosses which is promising. The work on deriving event ontologies (Allen et al., 2013) using WordNet glosses best addresses the shortcomings of semantic interpretation in previous works. It tries to build complex concepts compositionally using OWL-DL description logic and enables reasoning to derive the best classification of knowledge. However, their work mainly derives ontological information, whereas our work extracts full axioms in the form of inference rules. Also, as shown in Section 3, we use a more simple approach for expressing our inference rules (axioms), which enables semantic role alignment (a novel task introduced in this paper), resulting in a more precise, accurate, and easily usable inference rules for other NLP tasks. The earlier works on predicate-argument alignment have been mainly focused on finding lexical similarity and overlaps between pairs of sentences (Wolfe et al., 2013) which is different from aligning the semantic roles of not necessarily similar predicates as we do.

The main relevant work is on automatic acquisition of inference rules. Inference rules, e.g., 'someone$_x$ commutes $\rightarrow$ someone$_x$ changes positions', are very useful for tasks such as QA and RTE. The predominant approach, DIRT (Lin and Pantel, 2001), is based on distributional similarity, where two templates (such as 'X murder Y' and 'X kill Y') are deemed semantically similar if their argument vectors are similar. This similarity measure results in weak (and often incorrect) entailments (Melamud et al., 2013), but results in huge datasets. Among the 12 million DIRT inference rules only about about 50% seem correct and reasonable (Melamud et al., 2013). One instance of an incorrect rule is 'X entered Y $\rightarrow$ X left Y', which captures temporal relation between two predicates, and is incorrect as an entailment. Some later works have attempted to make the inference rules more precise by using lexical expansions for argument vectors (Melamud et al., 2013). However, their approaches still tend to produce many incorrect or too general entailments, such as 'Y is hijacked in X $\rightarrow$ Y crashes in X', which is the result of reporting bias which means there have been many reported hijacking events which have resulted in crashes, but hijacking

does not entail crash necessarily. All of the earlier inference rule acquisition approaches mostly use predicates with two arguments, which can result in limited and less-accurate application of rules for textual understanding tasks; however, our approach covers many complex predicate structures with various number of arguments and inter-connected predicates.

VerbOcean (Chklovski and Pantel, 2004) is another related work, which identifies verb entailment through instantiation of some manually constructed patterns. This idea led to more precise rules, but weak coverage since verbs do not co-occur often with patterns. In Section 5 we show that our approach outperforms VerbOcean by about 17% in precision and 30% in recall. Another recent work on acquiring inference rules is the work on learning verb inference rules from linguistically-motivated evidence (Weisman et al., 2012). This work argues that although most of the works on learning inference rules are using distributional similarity, they utilize information from various textual scopes ranging from verb co-occurrence within a sentence to a document, as well as corpus statistics, which results in richer set of linguistically motivated features in their supervised classification framework. Although they outperform some earlier methods, their method is still limited to verb to verb entailment without typed entities and semantic roles, which could make their rules less effective in actual language understanding tasks. In Section 5 we show that our approach results in a more accurate verb inference rules, outperforming this work by about 10%. More importantly, our approach attempts to produce semantically rich inference rules, i.e, sense-disambiguated predicates with all of the necessary semantic roles, which is far beyond the simple inference rules produced by this work.

Furthermore, paraphrases can be viewed as bidirectional inference rules. The works on automatic derivation of paraphrase databases (Dolan et al., 2004; Quirk et al., 2004) share some of the shortcomings of the works on acquiring inference rules. Mostly the paraphrase sets with the highest precision contain too general/trivial paraphrase rules (Ganitkevitch et al., 2013) such as 'higher than 90% $\leftrightarrow$ higher than 90 per cent' or 'and its relationship $\leftrightarrow$ and its link'. Inherently, definitions provide non-trivial pieces of information, so our set of high precision inference rules hardly contains such rules. Unlike these works, we rely on reading definitions instead of web-scale free texts which gives us higher precision of non-trivial inference rules, however, results in a smaller set of rules. By incorporating and linking various definitional resources, one can increase the size of the inference rules yielded by our approach.

7 Conclusion

We presented a novel attempt to automatically build a conceptual knowledge about events in the form of inference rules, which can serve as a semantically rich knowledge base useful for various language understanding tasks. We accomplish this by deep semantic parsing of glosses, inference rule learning enhanced by semantic role alignment, and chaining of the events. The evaluation results show that our semantic role alignment technique is very promising and our inference

rules are precise and informative pieces of knowledge. We have shown that learning inference rules by reading definitional resources can result in high accuracy and inherently non-trivial pieces of knowledge. In order to expand the coverage of our knowledge base, we are planning to apply our approach to other dictionaries. Moreover, we are looking into improving our semantic role alignment techniques for chaining of events, which can potentially result in more accurate inference rules. Our future goal is to experiment employing our definitional knowledge in QA and Reading Comprehension Tests.

Acknowledgments. We would like to thank William de Beaumont and anonymous reviewers for their invaluable comments. This work is funded by The Office of Naval Research under grant number N000141110417 and Nuance Foundation.

References

Allen, J., Beaumont, W.D., Blaylock, N., Ferguson, L.G.G., Orfan, J., Swift, M., Teng, C.M.: Acquiring commonsense knowledge for a cognitive agent. In: Proceedings of the AAAI Fall Symposium Series: Advances in Cognitive Systems (2011)

Allen, J., Beaumont, W.D., Galescu, L., Orfan, J., Swift, M., Teng, C.M.: Automatically deriving event ontologies for a commonsense knowledge base. In: IWCS (2013)

Allen, J., Ferguson, G., Swift, M., Stent, A.: Two diverse systems built using generic components for spoken dialogue (recent progress on trips). In: Proceedings of the ACL Demo, ACLdemo 2005, pp. 85–88. ACL (2005)

Blanco, E., Moldovan, D.: A semantically enhanced approach to determine textual similarity. In: Proceedings of EMNLP, pp. 1235–1245. ACL, Seattle (2013), http://www.aclweb.org/anthology/D13-1123

Bos, J.: Wide-coverage semantic analysis with boxer. In: Bos, J., Delmonte, R. (eds.) Semantics in Text Processing, pp. 277–286. Research in Computational Semantics, College Publications (2008)

Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics 19, 263–311 (1993)

Chklovski, T., Pantel, P.: VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In: Proceedings of the EMNLP (2004), http://aclweb.org/anthology/W04-3205

Clark, P., Fellbaum, C., Hobbs, J.R., Harrison, P., Murray, W.R., Thompson, J.: Augmenting wordnet for deep understanding of text. In: Semantics in Text Processing (2008)

Dolan, B., Quirk, C., Brockett, C.: Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: Proceedings of the 20th COLING, COLING 2004. ACL, Stroudsburg (2004), http://dx.doi.org/10.3115/1220355.1220406

Erekhinskaya, T.N., Satpute, M., Moldovan, D.I.: Multilingual extended wordnet knowledge base: Semantic parsing and translation of glosses. In: LREC, pp. 2990–2994 (2014)

Ganitkevitch, J., Durme, B.V., Callison-Burch, C.: PPDB: The paraphrase database. In: Proceedings of NAACL-HLT, pp. 758–764. ACL, Atlanta (2013), http://cs.jhu.edu/~ccb/publications/ppdb.pdf

Harris, Z.: Distributional structure. In: Katz, J.J. (ed.) The Philosophy of Linguistics. Oxford University Press, New York (1985)

Ide, N., Véronis, J.: Knowledge extraction from machine-readable dictionaries: An evaluation. In: Steffens, P. (ed.) EAMT-WS 1993. LNCS, vol. 898, pp. 17–34. Springer, Heidelberg (1995)

Levary, D., Eckmann, J.P., Moses, E., Tlusty, T.: Loops and self-reference in the construction of dictionaries. Phys. Rev. X (2012)

Lin, D., Pantel, P.: Dirt: Discovery of inference rules from text. In: Proceedings of the Seventh ACM SIGKDD, pp. 323–328. ACM, New York (2001), http://doi.acm.org/10.1145/502512.502559

Melamud, O., Dagan, I., Goldberger, J., Szpektor., I.: Using lexical expansion to learn inference rules from sparse data. In. In: Proceedings of ACL 2013 (2013)

Miller., G.: Wordnet: A lexical database for english. Communications of the ACM (1995)

Moldovan, D.I., Clark, C., Harabagiu, S.M., Hodges, D.: Cogex: A semantically and contextually enriched logic prover for question answering. J. Applied Logic 5(1), 49–69 (2007), http://dx.doi.org/10.1016/j.jal.2005.12.005

Moldovan, D.I., Rus, V.: Explaining answers with extended wordnet. In: ACL (2001)

Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. CL 29(1), 19–51 (2003), http://dx.doi.org/10.1162/089120103321337421

Picard, O., Blondin-Masse, A., Harnad, S., Marcotte, O., Chicoisne, G., Gargouri, Y.: Hierarchies in dictionary definition space. In: NIPS Workshop on Analyzing Networks and Learning With Graphs (2009)

Quirk, C., Brockett, C., Dolan, W.: Monolingual machine translation for paraphrase generation (2004)

Szpektor, I., Shnarch, E., Dagan, I.: Instance-based evaluation of entailment rule acquisition. In: Carroll, J.A., van den Bosch, A., Zaenen, A. (eds.) Proceeding of ACL Conference. ACL (2007)

Szpektor, I., Tanev, H., Dagan, I., Coppola, B.: Scaling web-based acquisition of entailment relations. In: Lin, D., Wu, D. (eds.) Proceedings of EMNLP 2004, pp. 41–48. ACL (July 2004)

Tarjan, R.: Depth first search and linear graph algorithms. SIAM Journal on Computing (1972)

Weisman, H., Berant, J., Szpektor, I., Dagan, I.: Learning verb inference rules from linguistically-motivated evidence. In: Proceedings of EMNLP-CoNLL, pp. 194–204. ACL, Jeju Island (2012), http://www.aclweb.org/anthology/D12-1018

Wolfe, T., Durme, B.V., Dredze, M., Andrews, N., Beller, C., Callison-Burch, C., DeYoung, J., Snyder, J., Weese, J., Xu, T., Yao, X.: Parma: A predicate argument aligner. In: Proceedings of ACL Short (2013), http://www.cs.jhu.edu/~vandurme/papers/PARMA:ACL:2013.pdf