Physical Causality of Action Verbs in Grounded Language Understanding

Qiaozi Gao[†] Malcolm Doering[‡*] Shaohua Yang[†] Joyce Y. Chai[†]
[†]Computer Science and Engineering, Michigan State University, East Lansing, MI, USA
[‡]Department of Systems Innovation, Osaka University, Toyonaka, Osaka, Japan
{gaoqiaoz, yangshao, jchai}@msu.edu
doering.malcolm.robert@irl.sys.es.osaka-u.ac.jp

Abstract

Linguistics studies have shown that action verbs often denote some *Change of State (CoS)* as the result of an action. However, the causality of action verbs and its potential connection with the physical world has not been systematically explored. To address this limitation, this paper presents a study on *physical causality* of action verbs and their implied changes in the physical world. We first conducted a crowdsourcing experiment and identified eighteen categories of physical causality for action verbs. For a subset of these categories, we then defined a set of detectors that detect the corresponding change from visual perception of the physical environment. We further incorporated physical causality modeling and state detection in grounded language understanding. Our empirical studies have demonstrated the effectiveness of causality modeling in grounding language to perception.

1 Introduction

Linguistics studies have shown that action verbs often denote some *change of state (CoS)* as the result of an action, where the *change of state* often involves an attribute of the direct object of the verb (Hovav and Levin, 2010). For example, the result of "slice a pizza" is that the state of the object (pizza) changes from one big piece to several smaller pieces. This change of state can be perceived from the physical world. In Artificial Intelligence (Russell and Norvig, 2010), decades of research on planning, for example, back to the early days of the STRIPS planner (Fikes and Nilsson, 1971), have defined action schemas to capture the change of state caused by a given action. Based on action schemas, planning algorithms can be applied to find a sequence of actions to achieve a goal state (Ghallab et al., 2004). The state of the physical world is a very important notion and changing the state becomes a driving force for agents' actions. Thus, motivated by linguistic literature on action verbs and AI literature on action representations, in our view, modeling change of physical state for action verbs, in other words, *physical causality*, can better connect language to the physical world.

Although this kind of physical causality has been described in linguistic studies (Hovav and Levin, 2010), a detailed account of potential causality that could be denoted by an action verb is lacking. For example, in VerbNet (Schuler, 2005) the semantic representation for various verbs may indicate that a change of state is involved, but it does not provide the specifics associated with the verb's meaning (e.g., to what attribute of its *patient* the changes might occur).

To address this limitation, we have conducted an empirical investigation on verb semantics from a new angle of how they may change the state of the physical world. As the first step in this investigation, we selected a set of action verbs from a cooking domain and conducted a crowd-sourcing study to examine the potential types of causality associated with these verbs. Motivated by linguistics studies on typology for gradable adjectives, which also have a notion of change along a scale (Dixon and Aikhenvald, 2006), we developed a set of eighteen main categories to characterize physical causality. We then defined a set of change-of-state detectors focusing on visual perception. We further applied two approaches, a knowledge-driven approach and a learning-based approach, to incorporate causality modeling in

[*] This work was conducted at Michigan State University where the author received his MS degree.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1814–1824,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

grounded language understanding. Our empirical results have demonstrated that both of these approaches achieve significantly better performance in grounding language to perception compared to previous approaches (Yang et al., 2016).

2 Related Work

The notion of *causality* or *causation* has been explored in psychology, linguistics, and computational linguistics from a wide range of perspectives. For example, different types of causal relations such as causing, enabling, and preventing (Goldvarg and Johnson-Laird, 2001; Wolff and Song, 2003) have been studied extensively as well as their linguistic expressions (Wolff, 2003; Song and Wolff, 2003; Neeleman et al., 2012) and automated extraction of causal relations from text (Blanco et al., 2008; Mulkar-Mehta et al., 2011; Radinsky et al., 2012; Riaz and Girju, 2014). Different from these previous works, this paper focuses on the physical causality of action verbs, in other words, change of state in the physical world caused by action verbs as described in (Hovav and Levin, 2010). This is motivated by recent advances in computer vision, robotics, and grounding language to perception and action.

A recent trend in computer vision has started looking into intermediate representations beyond lower-level visual features for action recognition, for example, by incorporating object affordances (Koppula et al., 2013) and causality between actions and objects (Fathi and Rehg, 2013). Fathi and Rehg (2013) have borken down detection of actions to detection of state changes from video frames. Yang and colleagues (2013; 2014) have developed an object segmentation and tracking method to detect state changes (or, in their terms, consequences of actions) for action recognition. More recently, Fire and Zhu (2015) have developed a framework to learn perceptual causal structures between actions and object statuses in videos.

In the robotics community, as robots' low-level control systems are often pre-programmed to handle (and thus execute) only primitive actions, a high-level language command will need to be translated to a sequence of primitive actions in order for the corresponding action to take place. To make such translation possible, previous works (She et al., 2014a; She et al., 2014b; Misra et al., 2015; She and Chai, 2016) explicitly model verbs with predicates describing the resulting states of actions. Their empirical evaluations have demonstrated how incorporating resulting states into verb representations can link language with underlying planning modules for robotic systems. These results have motivated a systematic investigation on modeling physical causality for action verbs.

Although recent years have seen an increasing amount of work on grounding language to perception (Yu and Siskind, 2013; Walter et al., 2013; Liu et al., 2014; Naim et al., 2015; Liu and Chai, 2015), no previous work has investigated the link between physical causality denoted by action verbs and the change of state visually perceived. Our work here intends to address this limitation and examine whether the causality denoted by action verbs can provide top-down information to guide visual processing and improve grounded language understanding.

3 Modeling Physical Causality for Action Verbs

3.1 Linguistics Background on Action Verbs

Verb semantics have been studied extensively in linguistics (Pustejovsky, 1991; Levin, 1993; Baker et al., 1998; Kingsbury and Palmer, 2002). Particularly, for action verbs (such as *run*, *throw*, *cook*), Hovav and Levin (Hovav and Levin, 2010) propose that they can be divided into two types: ***manner verbs*** that "specify as part of their meaning a manner of carrying out an action" (e.g., *nibble, rub, scribble, sweep, flutter, laugh, run, swim*), and ***result verbs*** that "specify the coming about of a result state" (e.g., *clean, cover, empty, fill, chop, cut, melt, open, enter*). Result verbs can be further classified into three categories: *Change of State* verbs, which denote a change of state for a property of the verb's object (e.g. "to warm"); *Inherently Directed Motion* verbs, which denote movement along a path in relation to a landmark object (e.g. "to arrive"); and *Incremental Theme* verbs, which denote the incremental change of volume or area of the object (e.g. "to eat") (Levin and Hovav, 2010). In this work, we mainly focus on result verbs. Unlike Hovav and Levin's definition of *Change of State* verbs, we use the term *change of state* in a more general way such that the location, volume, and area of an object are part of its state.

Previous linguistic studies have also shown that result verbs often specify movement along a scale (Hovav and Levin, 2010), i.e., they are verbs of scalar change. A scale is "a set of points on a particular dimension (e.g. height, temperature, cost)". In the case of verbs, the dimension is an attribute of the object of the verb. For example, "John cooled the coffee" means that the temperature attribute of the object *coffee* has decreased. Kennedy and McNally give a very detailed description of scale structure and its variations (Kennedy and McNally, 2005). Interestingly, gradable adjectives also have their semantics defined in terms of a scale structure. Dixon and Aikhenvald have defined a typology for adjectives which include categories such as Dimension, Color, Physical Property, Quantification, and Position (Dixon and Aikhenvald, 2006). The connection between gradable adjectives and result verbs through scale structure motivates us to use the Dixon typology as a basis to define our categorization of causality for verbs.

In summary, previous linguistic literature has provided abundant evidence and discussion on change of state for action verbs. It has also provided extensive knowledge on potential dimensions that can be used to categorize change of state as described in this paper.

3.2 A Crowd-sourcing Study

Motivated by the above linguistic insight, we have conducted a pilot study to examine the feasibility of causality modeling using a small set of verbs which appear in the TACoS corpus (Regneri et al., 2013). This corpus is a collection of natural language descriptions of actions that occur in a set of cooking videos. This is an ideal dataset to start with since it contains mainly descriptions of physical actions. Possibly because most actions in the cooking domain are goal-directed, a majority of the verbs in TACoS denote results of actions (changes of state) which can be observed in the world.

More specifically, we chose ten verbs (*clean, rinse, wipe, cut, chop, mix, stir, add, open, shake*)) based on the criteria that they occur relatively frequently in the corpus and take a variety of different objects as their *patient*. We paired each verb with three different objects in the role of *patient*. Nouns (e.g., *cutting board, dish, counter, knife, hand, cucumber, beans, leek, eggs, water, break, bowl, etc.*) were chosen based on the criteria that they represent objects dissimilar to each other, since we hypothesize that the change of state indicated by the verb will differ depending on the object's features.

Each verb-noun pair was presented to turkers via Amazon Mechanical Turk (AMT) and they were asked to describe (by text) the changes of state that occur to the object as a result of the verb. The descriptions were collected under two conditions: (1) without showing the corresponding video clips (so turkers would have to use their imagination of the physical situation) and (2) showing the corresponding video clips. For each condition and each verb-noun pair, we collected 30 turkers' responses, which resulted in a total of 1800 natural language responses describing change of state.

3.3 Categorization of Change of State

Based on Dixon and Aikhenvald's typology for adjectives (Dixon and Aikhenvald, 2006) and turkers' responses, we identified a categorization to characterize causality, as shown in Table 1. This categorization is also driven by the expectation that these attributes can be potentially recognized from the physical world by artificial agents. The first column specifies the type of state change and the second column specifies specific attributes related to the type. The third column specifies the particular value associated with the attribute, e.g., it could be a binary categorization on whether a change happens or not (i.e., *changes*), or a direction along a scale (i.e., *increase/decrease*), or a specific value (i.e., *specific* such as "five pieces"). In total, we have identified eighteen causality categories corresponding to eighteen attributes as shown in Table 1.

An important motivation of modeling physical causality is to provide guidance for visual processing. Our hypothesis is that once a language description is given together with its corresponding visual scene, potential causality of verbs or verb-noun pairs can trigger some visual detectors associated with the scene. This can potentially improve grounded language understanding (e.g., grounding nouns to objects in the scene). Next we give a detailed account on these visual detectors and their role in grounded language understanding.

4 Visual Detectors based on Physical Causality

The changes of state associated with the eighteen attributes can be detected from the physical world

Type	Attribute	Attribute Value
Dimension	Size, length, volume	Changes, increases, decreases, specific
	Shape	Changes, specific (cylindrical, flat, etc.)
Color/Texture	Color	Appear, disappear, changes, mix, separate, specific (green, red, etc.)
	Texture	Changes, specific (slippery, frothy, etc.)
Physical Property	Weight	Increase, decrease
	Flavor, smell	Changes, intensifies, specific
	Solidity	Liquefies, solidifies, specific
	Wetness	Becomes wet(ter), dry(er)
	Visibility	Appears, disappears
	Temperature	Increases, decreases
	Containment	Becomes filled, emptied, hollow
	Surface Integrity	A hole or opening appears
Quantification	Number of pieces	Increases, one becomes many, decreases, many become one
Position	Location	Changes, enter/exit container, specific
	Occlusion	Becomes covered, uncovered
	Attachment	Becomes detached
	Presence	No longer present, becomes present
	Orientation	Changes, specific

Table 1: Categorization of physical causality.

Attribute	Rule-based Detector	Refined Rule-based Detector
Attachment / NumberOfPieces	Multiple object tracks merge into one, or one object track breaks into multiple.	Multiple tracks merge into one.
		One track breaks into multiple.
Presence / Visibility	Object track appears or disappears.	Object track appears.
		Object track disappears.
Location	Object's final location is different from the initial location.	Location shifts upwards.
		Location shifts downwards.
		Location shifts rightwards.
		Location shifts leftwards.
Size	Object's x-axis length or y-axis length is different from the initial values.	Object's x-axis length increases.
		Object's x-axis length decreases.
		Object's y-axis length increases.
		Object's y-axis length decreases.

Table 2: Causality detectors applied to *patient* of a verb.

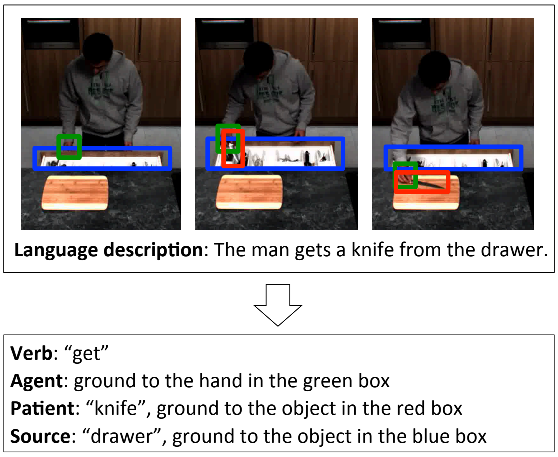
using various sensors. In this paper, we only focus on attributes that can be detected by visual perception. More specifically, we chose the subset: *Attachment*, *NumberOfPieces*, *Presence*, *Visibility*, *Location*, *Size*. They are chosen because: 1) according to the pilot study, they are highly correlated with our selected verbs; and 2) they are relatively easy to be detected from vision.

Corresponding to these causality attributes, we defined a set of rule-based detectors as shown in Table 2. These in fact are very simple detectors, which consist of four major detectors and a refined set that distinguishes directions of state change. These visual detectors are specifically applied to the potential objects that may serve as *patient* for a verb to identify whether certain changes of state occur to these objects in the visual scene.

5 Verb Causality in Grounded Language Understanding

In this section, we demonstrate how verb causality modeling and visual detectors can be used together for grounded language understanding. As shown in Figure 1, given a video clip V of human action and a parallel sentence S describing the action, our goal is to ground different semantic roles of the verb (e.g., *get*) to objects in the video. This is similar to the grounded semantic role labeling task (Yang et al., 2016). Here, we focus on a set of four semantic roles $\{$*agent*, *patient*, *source*, *destination*$\}$. We also assume that we have object and hand tracking results from video data. Each object in the video is represented by a track, which is a series of bounding boxes across video frames. Thus, given a video clip and a parallel sentence, the task is to ground semantic roles of the verb $\lambda_1, \lambda_2, \ldots, \lambda_k$ to object (or hand) tracks $\gamma_1, \gamma_2, \ldots, \gamma_n$, in the video.[1] We applied two approaches to this problem.

[1] For manipulation actions, the *agent* is almost always one of the human's hands (or both hands). So we constrain the grounding of the *agent* role to hand tracks, and constrain the grounding of the other roles to object tracks.



Language description: The man gets a knife from the drawer.

Verb: "get"
Agent: ground to the hand in the green box
Patient: "knife", ground to the object in the red box
Source: "drawer", ground to the object in the blue box

Figure 1: Grounding semantic roles of the verb *get* in the sentence: *the man gets a knife from the drawer*.

5.1 Knowledge-driven Approach

We intend to establish that the knowledge of physical causality for action verbs can be acquired directly from the crowd and such knowledge can be coupled with visual detectors for grounded language understanding.

Acquiring Knowledge. To acquire knowledge of verb causality, we collected a larger dataset of causality annotations based on sentences from the TACoS Multilevel corpus (Rohrbach et al., 2014), through crowd-sourcing on Amazon Mechanical Turk. Annotators were shown a sentence containing a verb-patient pair (e.g., "The person **chops** the **cucumber** into slices on the cutting board"). And they were asked to annotate the change of state that occurred to the *patient* as a result of the verb by choosing up to three options from the 18 causality attributes. Each sentence was annotated by three different annotators.

This dataset contains 4391 sentences, with 178 verbs, 260 nouns, and 1624 verb-noun pairs. After summarizing the annotations from three different annotators, each sentence is represented by a 18-dimension causality vector. In the vector, an element is 1 if at least two annotators labeled the corresponding causality attribute as true, 0 otherwise. For 83% of all the annotated sentences, at least one causality attribute was agreed on by at least two people.

From the causality annotation data, we can extract a *verb causality vector* $\mathbf{c}(\mathrm{v})$ for each verb by averaging all causality vectors of the sentences that contain this verb v.

Applying Knowledge. Since the collected causality knowledge was only for the *patient*, we first look at the grounding of *patient*. Given a sentence containing a verb v and its *patient*, we want to ground the *patient* to one of the object tracks in the video clip. Suppose we have the causality knowledge, i.e., $\mathbf{c}(\mathrm{v})$, for the verb. For each candidate track in the video, we can generate a causality detection vector $\mathbf{d}(\gamma_i)$, using the predefined causality detectors. A straightforward way is to ground the *patient* to the object track whose causality detection results has the best coherence with the causality knowledge of the verb. The coherence is measured by the cosine similarity between $\mathbf{c}(\mathrm{v})$ and $\mathbf{d}(\gamma_i)$.[2]

Since objects in other semantic roles often have relations with the *patient* during the action, once we have grounded the *patient*, we can use it as an anchor point to ground the other three semantic roles. To do this, we define two new detectors for grounding each role as shown in Table 3. These detectors are designed using some common sense knowledge, e.g., *source* is likely to be the initial location of the *patient*; *destination* is likely to be the final location of the *patient*; *agent* is likely to be the hand that touches the *patient*. With these new detectors, we simply ground a role to the object (or hand) track that has the largest number of positive detections from the corresponding detectors.

It is worth noting that although currently we only acquired knowledge for verbs that appear in the cooking domain, the same approach can be extended to verbs in other domains. The detectors associated with attributes are expected to remain the same. The significance of this knowledge-driven method is that, once you have the causality knowledge of a verb, it can be directly applied to any domain without additional training.

5.2 Learning-based Approach

Our second approach is based on learning from training data. A key requirement for this approach is the availability of annotated data where the arguments of a verb are already correctly grounded to the objects in the visual scene. Then we can learn the association between detected causality

[2] In the case that not every causality attribute has a corresponding detector, we need to first condense $\mathbf{c}(\mathrm{v})$ to the same dimensionality with $\mathbf{d}(\gamma_i)$.

Semantic Role	Rule-based Detector
Source	Patient track appears within its bounding box.
	Its track is overlapping with the patient track at the initial frame.
Destination	Patient track disappears within its bounding box.
	Its track is overlapping with the patient track at the final frame.
Agent	Its track is overlapping with the patient track when the patient track appears or disappears.
	Its track is overlapping with the patient track when the patient track starts moving or stops moving.

Table 3: Causality detectors for grounding *source*, *destination*, and *agent*.

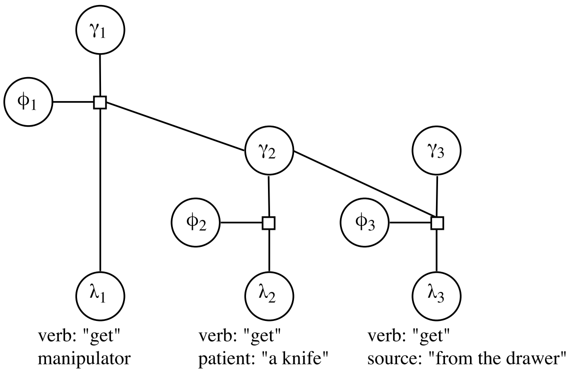


Figure 2: The CRF factor graph of the sentence: *the man gets a knife from the drawer*.

attributes and verbs. We use Conditional Random Field (CRF) to model the semantic role grounding problem. In this approach, causality detection results are used as features in the model.

An example CRF factor graph is shown in Figure 2. The structure of CRF graph is created based on the extracted semantic roles, which already abstracts away syntactic variations such as active/passive constructions. This CRF model is similar to the ones in (Tellex et al., 2011) and (Yang et al., 2016), where $\phi_1, \ldots, \phi_4$ are binary random variables, indicating whether the grounding is correct. In the learning stage, we use the following objective function:

$$\begin{aligned} & p(\Phi|\lambda_1, \ldots, \lambda_k, \gamma_1, \ldots, \gamma_k, \mathrm{v}) \\ = & \frac{1}{Z} \prod_i \Psi_i(\phi_i, \lambda_i, \gamma_1, \ldots, \gamma_k, \mathrm{v}) \end{aligned} \quad (1)$$

where Φ is the binary random vector $[\phi_1, \ldots, \phi_k]$, and v is the verb. Z is the normalization constant. Ψ_i is the potential function that takes the following log-linear form:

$$\Psi_i(\phi_i, \lambda_i, \Gamma, \mathrm{v}) = \exp\left(\sum_l w_l f_l(\phi_i, \lambda_i, \Gamma, \mathrm{v})\right) \quad (2)$$

where f_l is a feature function, w_l is feature weight to be learned, and $\Gamma = [\gamma_1, \ldots, \gamma_k]$ are the groundings. In our model, we use the following features:

1. Joint features between a track label of γ_i and a word occurrence in λ_i.
2. Joint features between each of the causality detection results and a verb v. Causality detection includes all the detectors in Table 2 and Table 3. Note that the causality detectors shown in Table 3 capture relations between groundings of different semantic roles.

During learning, gradient ascent with L2 regularization is used for parameter learning.

Compared to (Tellex et al., 2011) and (Yang et al., 2016), a key difference in our model is the incorporation of causality detectors. These previous works (Tellex et al., 2011; Yang et al., 2016) apply geometric features, for example, to capture relations, distance, and relative directions between grounding objects. These geometric features can be noisy. In our model, features based on causality detectors are motivated and informed by the underlying causality models for corresponding action verbs.

In the inference step, we want to find the most probable groundings. Given a video clip and its parallel sentence, we fix the Φ to be true, and search for groundings $\gamma_1, \ldots, \gamma_k$ that maximize the probability as in Equation 1. To reduce the search space we apply beam search to ground in the following order: *patient*, *source*, *destination*, *agent*.

5.3 Experiments and Results

We conducted our experiments using the dataset from (Yang et al., 2016). This dataset was developed from a subset of the TACoS corpus (Regneri et al., 2013). It contains a set of video clips paired with natural language descriptions related to two cooking tasks “cutting cucumber” and “cutting bread”. Each task has 5 videos showing how different people perform the same task, and each of these videos was split into pairs of video clips and corresponding sentences. For each video clip, objects are annotated with bounding boxes, tracks,

	All	***take***	***put***	***get***	***cut***	***open***	***wash***	***slice***	***rinse***	***place***	***peel***	***remove***
# Instances	279	58	15	47	29	6	28	13	29	29	10	15
With Ground-truth Track Labels												
Label Matching	67.7	70.7	46.7	72.3	69.0	16.7	85.7	69.2	82.8	37.9	90.0	60.0
Yang et al., 2016	84.6	93.2	91.7	93.6	77.8	80.0	93.5	86.7	90.0	66.7	80.0	38.9
VC-Knowledge	**89.6***	**94.8**	73.3	**100***	**93.1**	**83.3**	**100**	92.3	**96.6**	58.6	90.0	73.3*
VC-Learning	**90.3***	**94.8**	**86.7**	**100***	**93.1**	**83.3**	89.3	92.3	**96.6**	**75.9**	80.0	66.7*
Without Track Labels												
Label Matching	9.0	12.1	13.3	2.1	10.3	16.7	3.6	7.7	10.3	10.3	20.0	6.7
Yang et al., 2016	24.5	11.9	8.3	17.0	50.0	10.0	29.0	40.0	40.0	0	60.0	11.1
VC-Knowledge	**60.2***	**82.8***	**60.0***	**87.2***	**58.6**	**50.0**	**39.3**	**46.2**	**41.4**	**48.3***	10.0	**40.0**
VC-Learning	**71.7***	**91.4***	33.3	**87.2***	**72.4**	**83.3***	**46.4**	**84.6***	**51.7**	**65.5***	**80.0**	**60.0***

Table 4: Grounding accuracy on *patient* role

	Overall	**Agent**	**Patient**	**Source**	**Destination**
Number of Instances	644	279	279	51	35
With Ground-truth Track Labels					
Label Matching	66.3	68.5	67.7	41.2	74.3
Yang et al., 2016	84.2	86.4	84.6	72.6	81.6
VC-Knowledge	**86.8**	**89.3**	**89.6***	**60.8**	82.9
VC-Learning	**88.2***	**88.2**	**90.3***	**76.5**	88.6
Without Track Labels					
Label Matching	33.5	66.7	9.0	7.8	2.9
Yang et al., 2016	48.2	86.1	24.5	15.7	13.2
VC-Knowledge	**69.9***	**89.6**	**60.2***	**45.1***	**25.7**
VC-Learning	**75.0***	**87.1**	**71.7***	**41.2***	**54.3***

Table 5: Grounding accuracy on four semantic roles

and labels (e.g. "cucumber, cutting board" etc). For each sentence, the semantic roles of a verb are extracted using Propbank (Kingsbury and Palmer, 2002) definitions and each of them is annotated with the ground truth groundings in terms of the object tracks in the corresponding video clip. We selected the 11 most frequent verbs (*get, take, wash, cut, rinse, slice, place, peel, put, remove, open*) and the 4 most frequent explicit semantic roles (*agent, patient, source, destination*) in this evaluation. In total, this dataset includes 977 pairs of video clips and corresponding sentences, and 1096 verb-patient occurrences.

We compare our knowledge-driven approach (*VC-Knowledge*) and learning-based approach (*VC-Learning*) with the following two baselines.

Label Matching. This method simply grounds the semantic role to the track whose label matches the word phrase. If there are multiple matching tracks, it will randomly choose one of them. If there is no matching track, it will randomly select one from all the tracks.

Yang et al., 2016. This work studies grounded semantic role labeling. The evaluation data from this work is used in this paper. It is a natural baseline for comparison.

To evaluate the learning-based approaches such as *VC-Learning* and *(Yang, et al., 2016)*, 75% of video clips with corresponding sentences were randomly sampled as the training set. The remaining 25% were used as the test set. For approaches which do not need training such as *Label Matching* and *VC-Knowledge*, we used the same test set to report their results.

The results of the *patient* role grounding for each verb are shown in Table 4. The results of grounding all four semantic roles are shown in Table 5. The scores in bold are statistically significant ($p < 0.05$) compared to the *Label Matching* method. The scores with an asterisk ($*$) are statistically significant ($p < 0.05$) compared to *(Yang et al., 2016)*.

As it can be difficult to obtain labels for the track, especially when the vision system encounters novel objects, we further conducted several experiments assuming we do not know the labels for the object tracks. In this case, only geometric information of tracked objects is available. Table 4 and Table 5 also include these results.

From the grounding results, we can see that the causality modeling has shown to be very effective in grounding semantic roles. First of all, both the knowledge-driven approach and the learning-based approach outperform the two baselines. In

	All	***take***	***put***	***get***	***cut***	***open***	***wash***	***slice***	***rinse***	***place***	***peel***	***remove***
VC-Knowledge	89.6	94.8	73.3	100	93.1	83.3	100	92.3	96.6	58.6	90.0	73.3
P-VC-Knowledge	89.9	96.6	73.3	100	96.6	66.7	100	92.3	96.6	65.5	90.0	60.0

Table 6: Grounding accuracy on *patient* role using predicted causality knowledge.

particular, our knowledge-driven approach (*VC-Knowledge*) even outperforms the trained model (*Yang et al., 2016*). Our learning-based approach (*VC-Learning*) achieves the best overall performance. In the learning-based approach, causality detection results can be seen as a set of intermediate visual features. The reason that our learning-based approach significantly outperforms the similar model in (*Yang et al., 2016*) is that the causality categorization provides a good guideline for designing intermediate visual features. These causality detectors focus on the changes of state of objects, which are more robust than the geometric features used in (*Yang et al., 2016*).

In the setting of no object recognition labels, *VC-Knowledge* and *VC-Learning* also generate significantly better grounding accuracy than the two baselines. This once again demonstrates the advantage of using causality detection results as intermediate visual features. All these results illustrate the potential of causality modeling for grounded language understanding.

The results in Table 5 also indicate that grounding *source* or *destination* is more difficult than grounding *patient* or *agent*. One reason could be that *source* and *destination* do not exhibit obvious change of state as a result of action, so their groundings usually depend on the correct grounding of other roles such as *patient*.

Since automated tracking for this TACoS dataset is notably difficult due to the complexity of the scene and the lack of depth information, our current results are based on annotated tracks. But object tracking algorithms have made significant progress in recent years (Yang et al., 2013; Milan et al., 2014). We intend to apply our algorithms with automated tracking on real scenes in the future.

6 Causality Prediction for New Verbs

While various methods can be used to acquire causality knowledge for verbs, it may be the case that during language grounding, we do not know the causality knowledge for every verb. Furthermore, manual annotation/acquisition of causality knowledge for all verbs can be time-consuming. In this section, we demonstrate that the existing causality knowledge for some seed verbs can be used to predict causality for new verbs of which we have no knowledge.

We formulate the problem as follows. Suppose we have causality knowledge for a set of seed verbs as training data. Given a new verb, whose causality knowledge is not known, our goal is to predict the causality attributes associated with this new verb. Although the causality knowledge is unknown, it is easy to compute Distributional Semantic Models (DSM) for this verb. Then our goal is to find the causality vector $\mathbf{c}'$ that maximizes

$$\arg\max_{\mathbf{c}'} p(\mathbf{c}'|\mathbf{v}), \tag{3}$$

where $\mathbf{v}$ is the DSM vector for the verb v. The usage of DSM vectors is based on our hypothesis that the textual context of a verb can reveal its possible causality information. For example, the contextual words “pieces” and “halves” may indicate the CoS attribute “NumberOfPieces” for the verb “cut”.

We simplify the problem by assuming that the causality vector $\mathbf{c}'$ takes binary values, and also assuming the independence between different causality attributes. Thus, we can formulate this task as a group of binary classification problems: predicting whether a particular causality attribute is positive or negative given the DSM vector of a verb. We apply logistic regression to train a separate classifier for each attribute. Specifically, for the features of a verb, we use the Distributional Memory (*typeDM*) (Baroni and Lenci, 2010) vector. The class label indicates whether the corresponding attribute is associated with the verb.

In our experiment we chose six attributes to study: *Attachment*, *NumberOfPieces*, *Presence*, *Visibility*, *Location*, and *Size*. For each one of the eleven verbs in the grounding task, we predict its causality knowledge using classifiers trained on all other verbs (i.e., 177 verbs in training set). To evaluate the predicted causality vectors, we applied them in the knowledge-driven approach (*P-VC-Knowledge*). Grounding results were compared with the same method using the causality knowledge collected via crowd-sourcing. Table 6 shows the grounding accuracy on the *patient* role for each verb. For most verbs, using the predicted knowledge achieves very similar performance compared to using the collected knowledge. The overall grounding accuracy of using the predicted knowledge on all four semantic roles is only 0.3% lower than using the collected knowledge. This result demonstrates that physical causality of action verbs, as part of verb semantics, can be learned through Distributional Semantics.

7 Conclusion

This paper presents, to the best of our knowledge, the first attempt that explicitly models the physical causality of action verbs. We have applied causality modeling to the task of grounding semantic roles to the environment using two approaches: a knowledge-based approach and a learning-based approach.

Our empirical evaluations have shown encouraging results for both approaches. When annotated data is available (in which semantic roles of verbs are grounded to physical objects), the learning-based approach, which learns the associations between verbs and causality detectors, achieves the best overall performance. On the other hand, the knowledge-based approach also achieves competitive performance (even better than previous learned models), without any training. The most exciting aspect about the knowledge-based approach is that causality knowledge for verbs can be acquired from humans (e.g., through crowd-sourcing) and generalized to novel verbs about which we have not yet acquired causality knowledge.

In the future, we plan to build a resource for modeling physical causality for action verbs. As object recognition and tracking are undergoing significant advancements in the computer vision field, such a resource together with causality detectors can be immediately applied for any applications that require grounded language understanding.

Acknowledgments

This work was supported in part by the National Science Foundation (IIS-1208390 and IIS-1617682) and DARPA under the SIMPLEX program through UCLA (N66001-15-C-4035). The authors would like to thank anonymous reviewers for valuable comments and suggestions.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

Robert MW Dixon and Alexandra Y Aikhenvald. 2006. *Adjective Classes: A Cross-linguistic Typology*. Explorations in Language and Space C. Oxford University Press.

Alahoum Fathi and James M Rehg. 2013. Modeling actions through state changes. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2579–2586. IEEE.

Richard E. Fikes and Nils J. Nilsson. 1971. Strips: A new approach to the application of theorem proving to problem solving. In *Proceedings of the 2Nd International Joint Conference on Artificial Intelligence*, IJCAI'71, pages 608–620, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Amy Fire and Song-Chun Zhu. 2015. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):23.

Malik Ghallab, Dana Nau, and Paolo Traverso. 2004. *Automated planning: theory & practice*. Elsevier.

Eugenia Goldvarg and Philip N Johnson-Laird. 2001. Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive science*, 25(4):565–610.

Malka Rappaport Hovav and Beth Levin. 2010. Reflections on Manner / Result Complementarity. *Lexical Semantics, Syntax, and Event Structure*, pages 21–38.

Christopher Kennedy and Louise McNally. 2005. Scale structure and the semantic typology of gradable predicates. *Language*, 81(2)(0094263):345–381.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*.

Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. 2013. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970.

Beth Levin and Malka Rappaport Hovav. 2010. Lexicalized scales and verbs of scalar change. In *46th Annual Meeting of the Chicago Linguistics Society*.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Changsong Liu and Joyce Y. Chai. 2015. Learning to mediate perceptual differences in situated human-robot dialogue. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI15)*, pages 2288–2294, Austin, TX.

Changsong Liu, Lanbo She, Rui Fang, and Joyce Y. Chai. 2014. Probabilistic labeling for efficient referential grounding based on collaborative discourse. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 13–18, Baltimore, MD.

Anton Milan, Stefan Roth, and Kaspar Schindler. 2014. Continuous energy minimization for multitarget tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(1):58–72.

Dipendra Kumar Misra, Kejia Tao, Percy Liang, and Ashutosh Saxena. 2015. Environment-driven lexicon induction for high-level instructions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 992–1002, Beijing, China, July. Association for Computational Linguistics.

Rutu Mulkar-Mehta, Christopher Welty, Jerry R Hoobs, and Eduard Hovy. 2011. Using granularity concepts for discovering causal relations. In *Proceedings of the FLAIRS conference*.

Iftekhar Naim, Young C. Song, Qiguang Liu, Liang Huang, Henry Kautz, Jiebo Luo, and Daniel Gildea. 2015. Discriminative unsupervised alignment of natural language instructions with corresponding video segments. In *Proceedings of NAACL HLT 2015*, pages 164–174, Denver, Colorado, May–June. Association for Computational Linguistics.

Ad Neeleman, Hans Van de Koot, et al. 2012. The linguistic expression of causation. *The Theta System: Argument Structure at the Interface*, page 20.

J Pustejovsky. 1991. The syntax of event structure. *Cognition*, 41(1-3):47–81.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, pages 909–918. ACM.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36.

Mehwish Riaz and Roxana Girju. 2014. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDial)*, page 161.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition*, pages 184–195. Springer.

S. Russell and P. Norvig. 2010. *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Karin Kipper Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Lanbo She and Joyce Y. Chai. 2016. Incremental acquisition of verb hypothesis space towards physical world interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany.

Lanbo She, Yu Cheng, Joyce Chai, Yunyi Jia, Shaohua Yang, and Ning Xi. 2014a. Teaching robots new actions through natural language instructions. In *RO-MAN, 2014 IEEE*, Edinburgh, UK, August.

Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Chai, and Ning Xi. 2014b. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Proceedings of the SIGDIAL 2014 Conference*, Philadelphia, US, June.

Grace Song and Phillip Wolff. 2003. Linking perceptual properties to the linguistic expression of causation. *Language, culture and mind*, pages 237–250.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Matthew R Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. 2013. Learning semantic maps from natural language descriptions. In *Robotics: Science and Systems*.

Phillip Wolff and Grace Song. 2003. Models of causation and the semantics of causal verbs. *Cognitive Psychology*, 47(3):276–332.

Phillip Wolff. 2003. Direct causation in the linguistic coding and individuation of causal events. *Cognition*, 88(1):1–48.

Yezhou Yang, Cornelia Fermuller, and Yiannis Aloimonos. 2013. Detection of manipulation action consequences (mac). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2563–2570.

Yezhou Yang, Anupam Guha, C Fermuller, and Yiannis Aloimonos. 2014. A cognitive system for understanding human manipulation actions. *Advances in Cognitive Sysytems*, 3:67–86.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Y. Chai. 2016. Grounded semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, San Diego, CA.

Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 53–63.