

MODELING PHYSICAL CAUSALITY OF ACTION VERBS FOR GROUNDED LANGUAGE
UNDERSTANDING

By

Qiaozi Gao

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science – Doctor of Philosophy

2019

ABSTRACT

MODELING PHYSICAL CAUSALITY OF ACTION VERBS FOR GROUNDED LANGUAGE UNDERSTANDING

By

Qiaozi Gao

Building systems that can understand and communicate through human natural language is one of the ultimate goals in AI. Decades of natural language processing research has been mainly focused on learning from large amounts of language corpora. However, human communication relies on a significant amount of un verbalized information, which is often referred as commonsense knowledge. This type of knowledge allows us to understand each other's intention, to connect language with concepts in the world, and to make inference based on what we hear or read. Commonsense knowledge is generally shared among cognitive capable individuals, thus it is rarely stated in human language. This makes it very difficult for artificial agents to acquire commonsense knowledge from language corpora. To address this problem, this dissertation investigates the acquisition of commonsense knowledge, especially knowledge related to basic actions upon the physical world and how that influences language processing and grounding.

Linguistics studies have shown that action verbs often denote some *change of state (CoS)* as the result of an action. 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. However, the causality of action verbs and its potential connection with the physical world has not been systematically explored. Artificial agents often do not have this kind of basic commonsense causality knowledge, which makes it difficult for these agents to work with humans and to reason, learn, and perform actions.

To address this problem, this dissertation models dimensions of physical causality associated with common action verbs. Based on such modeling, several approaches are developed to incorporate causality knowledge to language grounding, visual causality reasoning, and commonsense story comprehension.

Copyright by
QIAOZI GAO
2019

ACKNOWLEDGEMENTS

First of all, I would like to express my great appreciation to my advisor, Dr. Joyce Y. Chai, for her patient guidance and continuous support throughout my doctoral studies. Dr. Chai always has a very insightful and knowledgeable view about the field of study, and I always feel enlightened after talking to her. Without her invaluable expertise and enthusiastic encouragement, I could not have finished this dissertation. Her great passion and patience towards research has left an impact on me, which will always guide me throughout my career.

I would also like to express my deep gratitude to Dr. Arun Ross, Dr. Pang-Ning Tan and Dr. Daniel Morris, for being on my program committee and for providing a lot of guidance and help at every milestone of my Ph.D program.

I want to thank my fellow colleagues in the Language and Interaction Research (LAIR) group, especially Shaohua Yang, Lanbo She, Malcolm Doering, Sari Saba-Sadiya, Changsong Liu, Rui Fang, Guangyue Xu, Kenneth Stewart, James Peterkin II, Sarah Fillwock and James Finch, who gave me a lot of support during the past several years. I enjoyed the discussion and teamwork during our collaborations.

Finally, I want to express my warmest gratitude to my family. I thank my parents for their unwavering support and love. My parents made me who I am today and I know I could never repay them for all those lessons they have taught me. Lastly, I would like to thank my beautiful wife Xi Liu for always being supportive and encouraging, and always pushing me to be the best version of myself.

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	x
CHAPTER 1 INTRODUCTION	1
1.1 Modeling Physical Causality of Action Verbs	2
1.2 Physical Causality Modeling for Language Grounding Task	5
1.3 Visual Causality Reasoning	6
1.4 Commonsense Reasoning about Physical Actions	7
1.5 Contributions	8
1.6 Organization of this Dissertation	9
CHAPTER 2 RELATED WORK	11
2.1 Theoretical Linguistics on Verbs	11
2.2 Grounding Language to Perception	12
2.2.1 Grounding Words in Perception	12
2.2.1.1 Grounding to Discretized Perceptual Signal	12
2.2.1.2 Learning from Ambiguous Parallel Data	14
2.2.1.3 Grounding Verbs	16
2.2.1.4 Context-Dependent Word Meaning	17
2.2.2 Grounding Phrases and Sentences	17
2.2.2.1 Referent Grounding	17
2.2.2.2 Grounding Action Frames	18
2.2.2.3 Parsing and Perception	18
2.2.2.4 Jointly Modeling Parsing and Perception	20
2.2.2.5 Neural Network approaches	21
2.3 Natural Language Inference Tasks	22
2.3.1 Recognizing Textual Entailment (RTE)	22
2.3.2 Winograd Schema Challenge (WSC)	23
2.3.3 Causal Reasoning	24
2.3.3.1 Choice of Plausible Alternatives (COPA)	24
2.3.3.2 Story Cloze Test	25
2.4 Knowledge Resources	26
2.4.1 Hand-built Knowledge Resources	26
2.4.2 Automatically Extracted Knowledge	27
2.5 Related Work in Computer Vision and Robotics	28
2.5.1 Related Work in Computer Vision	28
2.5.2 Related Work in Robotics	29
CHAPTER 3 MODELING PHYSICAL CAUSALITY OF VERBS	30
3.1 Categorization of Physical Causality	30

3.1.1	Linguistics Background on Action Verbs	30
3.1.2	A Crowd-Sourcing Study	31
3.1.3	Categorization of Change of State	32
3.1.4	Evaluation: Verb Similarity Judgement and Thematic Fit Estimation	36
3.1.4.1	Verb Similarity Judgement	37
3.1.4.2	Thematic Fit Estimation	39
3.2	Modeling Causality Knowledge via Embedding Methods	40
3.2.1	Cause-Effect Data Collection	40
3.2.2	Causality Embedding Models	41
3.2.3	Evaluation: Causality Embedding in Causal QA	44
3.2.3.1	Ranking Algorithm	44
3.2.3.2	Dataset	45
3.2.3.3	Models for Comparison	45
3.2.3.4	Evaluation Results	45
CHAPTER 4 PHYSICAL CAUSALITY MODELING FOR LANGUAGE GROUND-		
	ING TASK	47
4.1	Introduction	47
4.2	Visual Detectors based on Physical Causality	48
4.3	Verb Causality in Language Grounding	49
4.3.1	Knowledge-driven Approach	49
4.3.1.1	Acquiring Knowledge	49
4.3.1.2	Applying Knowledge	50
4.3.2	Learning-based Approach	52
4.3.3	Experiments and Results	53
4.4	Causality Prediction for New Verbs	56
4.5	Conclusion	58
CHAPTER 5 VISUAL CAUSALITY REASONING		59
5.1	Introduction	59
5.2	Action-Effect Data Collection	60
5.2.1	Actions (verb-noun pairs)	60
5.2.2	Effects Described in Language	60
5.2.3	Effects Depicted in Images	61
5.3	Action-Effect Prediction	62
5.3.1	Extracting Effect Phrases from Language Data	63
5.3.2	Downloading Web Images	63
5.3.3	Models	65
5.3.4	Evaluation	66
5.3.4.1	Methods for Comparison	67
5.3.4.2	Evaluation Results	67
5.4	Generalizing Effect Knowledge to New Verb-Noun Pairs	69
5.4.1	Action-Effect Embedding Model	69
5.4.2	Evaluation	71
5.5	Discussion and Conclusion	73

CHAPTER 6	UNDERSTANDING PHYSICAL ACTIONS THROUGH NATURAL LANGUAGE STORIES	74
6.1	Introduction	74
6.2	Physical Commonsense Reasoning Tasks	75
6.2.1	Data Collection through Crowdsourcing	76
6.2.2	Underlying Commonsense Knowledge	78
6.2.3	Comparison with Existing Tasks	81
6.3	Methods	82
6.3.1	The Attentive-Reader Model	82
6.3.1.1	Leveraging Physical Causality Knowledge	84
6.3.1.2	Typed Physical Causality Knowledge	84
6.3.2	Models for Comparison	85
6.4	Experiments	86
6.4.1	Experimental Settings	86
6.4.2	Results and Analysis	87
6.4.3	Predicting Breakpoints in Negative Stories	88
6.5	Summary	89
CHAPTER 7	CONCLUSIONS AND FUTURE DIRECTIONS	91
BIBLIOGRAPHY	94

LIST OF TABLES

Table 3.1: Categorization of physical causality.	33
Table 3.2: Variability of causality labels over different object and scene conditions.	35
Table 3.3: Results of verb similarity judgement task using Distributional Memory (DM) model, and concatenation model (DM+CoS). (Pearson’s correlation ρ , all values are significant with $p < 0.001$.)	38
Table 3.4: Results of thematic fitness estimation using Distributional Memory (DM) model and concatenation model (DM+Causality). (Pearson’s correlation ρ , all values are significant with $p < 0.001$.)	40
Table 3.5: Example cause and effect text from our collected data.	41
Table 3.6: Example patterns that are used to extract state phrases (bold) from sample sentences.	43
Table 3.7: MAP results for verb causality question answering task.	46
Table 4.1: Causality detectors applied to <i>patient</i> of a verb.	48
Table 4.2: Causality detectors for grounding <i>source</i> , <i>destination</i> , and <i>agent</i>	51
Table 4.3: Grounding accuracy on <i>patient</i> role	55
Table 4.4: Grounding accuracy on four semantic roles	55
Table 4.5: Grounding accuracy on <i>patient</i> role using predicted causality knowledge.	57
Table 5.1: Example action and effect text from our collected data.	61
Table 5.2: Example patterns that are used to extract effect phrases (bold) from sample sentences.	63
Table 5.3: Results for the action-effect prediction task (given an action, rank all the candidate images).	68
Table 5.4: Results for the action-effect prediction task (given an image, rank all the actions).	68
Table 5.5: Results for the action-effect prediction task (given an action, rank all the candidate images).	71

Table 5.6: Results for the action-effect prediction task (given an image, rank all the actions).	71
Table 5.7: Example predicted effect phrases for new verb-noun pairs. Unseen verbs and nouns are shown in bold.	72
Table 6.1: Typed state attributes for physical causality knowledge.	85
Table 6.2: Prediction accuracy results on the physical commonsense reasoning tasks.	87
Table 6.3: Prediction accuracy results of training on one task and evaluating on the other task.	88

LIST OF FIGURES

Figure 1.1: An image showing apple slices. Question: “What actions could possibly cause this situation?”	2
Figure 3.1: Distributions of causality labels for verbs <i>clean</i> and <i>rinse</i>	34
Figure 3.2: Architecture of the verb causality embedding model.	42
Figure 4.1: Grounding semantic roles of the verb <i>get</i> in the sentence: <i>the man gets a knife from the drawer</i>	50
Figure 4.2: The CRF factor graph of the sentence: <i>the man gets a knife from the drawer</i> . . .	52
Figure 5.1: Positive images (top row) and negative images (bottom row) of the action <i>peel-orange</i>	62
Figure 5.2: Examples of image search results.	64
Figure 5.3: Architecture for the action-effect prediction model with bootstrapping.	66
Figure 5.4: Several example test images and their predicted actions and predicted effect descriptions. The actions in blue are ground-truth labels.	68
Figure 5.5: Architecture of the action-effect embedding model.	70
Figure 6.1: Example story data for the cloze task and the ordering task. Candidates in red are correct answers.	75
Figure 6.2: Interface used for annotating stories for the cloze task.	77
Figure 6.3: Interface used for annotating stories for the ordering task.	77
Figure 6.4: Network architecture for the Attentive-Reader. Note that this architecture only shows the computation structure for the anomaly scores corresponding to sentence 3 (score s_{31} and score s_{32}). The anomaly scores for other sentences are computed via similar processes.	83
Figure 6.5: Network architecture for the EntNet-based approach.	85

CHAPTER 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 [54]. 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 [126], decades of research on planning, for example, back to the early days of the STRIPS planner [33], 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 [39]. 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.

Physical causality is one important aspect of human commonsense knowledge. Suppose we are given a statement “the apple is in small pieces”, or given an image as shown in Figure 1.1, what actions could possibly cause the situation described in the text or illustrated by the image? We humans have no problem of inferring potential causes: an external action such as *cut* or *slice* most likely have happened to a whole apple. What allows us to make such inference is the common sense knowledge we have, especially in this case the very basic cause-effect knowledge about how actions (and thus action verbs) may affect the state of the world. Let’s suppose we give the same statement and the same image to an artificial agent, will the agent be able to infer the potential causes? The answer is most likely no.

Despite tremendous progress in knowledge representation, automated reasoning, and machine learning, artificial agents still lack the understanding of naive causal relations regarding the physical world. This is one of the bottlenecks in machine intelligence. If artificial agents ever become



Figure 1.1: An image showing apple slices. Question: “What actions could possibly cause this situation?”

capable of working with humans as partners, they will need to have this kind of physical action-effect understanding to help them reason, learn, and perform actions.

In this dissertation, to address these limitations mentioned above, a series of investigations on physical causality of action verbs are performed. First, crowd-sourcing experiments were designed and conducted to collect physical causality knowledge from human users. Based on the collected causality knowledge data, two different approaches were developed to model causality knowledge: a categorization-based approach and a language embedding-based approach. These modeling approaches transform human causality knowledge to machine understandable representations, which can enable commonsense reasoning. We then developed several approaches to incorporate physical causality knowledge to language grounding, visual causality reasoning, and commonsense story understanding, where such knowledge plays an important role.

1.1 Modeling Physical Causality of Action Verbs

Causation in the physical world has long been a central discussion to philosophers who study casual reasoning and explanation [28, 44], to mathematicians or computer scientists who apply computational approaches to model cause-effect prediction [107], and to domain experts (e.g., medical doctors) who attempt to understand the underlying cause-effect relations (e.g., disease and

symptoms) for their particular inquiries. Apart from this wide range of topics, this dissertation investigates a specific kind of causation, the very basic causal relations between a concrete action (expressed in the form of a verb-noun pair such as “cut-cucumber”) and the change of the physical state caused by this action. We believe that physical causality knowledge forms an essential component of verb semantics, and is crucial to better connecting natural language with the physical world.

Verb semantics have been studied extensively. Theoretical linguistics use a frame of semantic roles to capture semantics of verbs [73]. Semantic roles include *agent*, *patient*, *instrument*, *source*, *destination*, etc. Several knowledge base resources on verb semantics have been made available, such as VerbNet [127], FrameNet [7], and PropBank [66]. However these resources mainly focus on organizing verbs into classes, and representing verb semantics with action frames. They do not provide a detailed and formal account of potential causality denoted by verbs.

In the NLP community, there is an increasing amount of effort on capturing common knowledge or commonsense knowledge. Except for few [167] which acquires commonsense knowledge from annotated images, most of the previous effort applies information extraction techniques to extract facts from a large amount of web data. For example, DBpedia [71] and YAGO [144] knowledge bases contain millions of facts about the world such as people and places. However, these knowledge bases do not contain basic cause-effect knowledge related to concrete actions such as “drop a glass will cause the glass to break into pieces”; “grind coffee beans will cause coffee beans to become powder”. Lacking this kind of basic physical cause-effect knowledge hinders artificial agents from connecting natural language to the physical world, and thus inhibits the capability of reasoning, learning and performing actions.

Motivated by these observations, this dissertation investigates the acquisition and modeling of commonsense causality knowledge associated with concrete action verbs. The basic cause-effect knowledge is so fundamental for human beings and is shared by cognitive capable individuals. This kind of knowledge is often presupposed in our communication and not explicitly stated. Thus, it is difficult to extract cause-effect relations from existing textual data (e.g., web). To overcome

this problem, several crowd-sourcing tasks were designed to collect physical causality data. In these crowd-sourcing tasks, human subjects were asked to explicitly express their knowledge on action verbs, through natural language descriptions or through answering designed multiple choice questions.

After data collection, we propose two different approaches to model physical causality knowledge. One approach is categorization-based, where the changes of state are categorized by the physical attributes of objects, and the causality knowledge for an action verb is represented as its association vector with those attributes. Another approach utilizes neural network embedding models, where causality knowledge is modeled through similarities between language embedding vectors.

For the first approach, in order to examine the potential types of causality associated with action verbs, a pilot crowd-sourcing experiment was first conducted on a selected set of action verbs. Motivated by linguistics studies on typology for gradable adjectives, which also have a notion of change along a scale [23], we developed a set of eighteen main categories to characterize physical causality. Then, the evaluation results on verb similarity judgement task and thematic fit estimation task demonstrate that categorization-based causality modeling can be a good supplement of distributional semantics for verb meanings.

For the second approach, we first collected a dataset of natural language cause-effect descriptions for a set of most frequently used action verbs. A neural network structure was developed to learn a *cause* and *effect* embedding space from the collected language data to capture common-sense causality knowledge. The proposed embedding models were evaluated on causal question answering, for example, to answer questions such as “what action could cause the state of the world described in the text?” or “what state change could happen to the object as a result of this action?” The experimental results have shown the potential of this embedding approach in enabling causal reasoning of actions for artificial agents.

Further, we applied the collected physical causality knowledge together with different modeling approaches to several novel tasks, demonstrating that physical causality modeling has a good

potential for intelligent systems that can deeply understand human language and better connect language with the physical world.

1.2 Physical Causality Modeling for Language Grounding Task

Physical causality knowledge captures potential changes of physical states caused by action verbs. The change of state can be perceived from the physical world. Therefore, modeling physical causality knowledge can help the machine to better ground natural language components to concepts of the world, in other words, connecting words, phrases or sentences to objects, states and actions in the physical world.

We conduct a study to incorporate categorization-based physical causality modeling in a language grounding task [36]. In this task, a system is given parallel language and visual data as input, and the goal is to ground language components to objects from the visual data. Our hypothesis is that modeling physical causality can provide guidance for visual processing: once a parallel language and visual data about an action is given, the potential causality of the verb or the verb-noun pair can trigger some visual detectors that mainly focus on potential state changes caused by this action. Applying these visual detectors to the visual data can potentially improve the performance of grounded language understanding.

Based on the categorization of physical causality attributes, we designed a set of change-of-state detectors to detect the corresponding changes from visual perception of the physical environment. We further applied two approaches, a knowledge-driven approach and a learning-based approach, to incorporate causality modeling in grounded language understanding. The knowledge-driven approach incorporates the collected human physical causality knowledge with the change-of-state detectors to find the best groundings for semantic roles. The learning-based approach utilizes Conditional Random Field (CRF) to model the relations between physical objects and language components. Instead of using the collected human physical causality knowledge, it learns the association between causality attributes and verbs from training data. The empirical results have demonstrated that both of these approaches achieve significantly better performance in grounding

language to perception compared to previous approaches [162].

1.3 Visual Causality Reasoning

We humans share a vast amount of commonsense causality knowledge, and we use them in our daily lives without even noticing it. For example, given a verb (e.g., *grind*) and a noun (e.g., *coffee beans*), we can predict the effect on the state of the world caused by this action. Given a photo, for example, showing many small cucumber pieces, we can infer that some external action (e.g., cut) on a cucumber could cause such state. We can make such action-effect prediction because we have developed an understanding of this kind of basic action-effect relations at a very young age [6]. What about machines? Will artificial agents be able to make the same kind of predictions? The answer is not yet.

To address this problem, we introduce a new task on naive physical action-effect prediction [37]. This task includes both cause prediction: given an image which describes a state of the world, identify the most likely action (in the form of a verb-noun pair, from a set of candidates) that can result in that state; and effect prediction: given an action in the form of a verb-noun pair, identify images (from a set of candidates) that depicts the most likely effects on the state of the world caused by that action. Note that there could be different ways of formulating this problem, for example, both causes and effects are in the form of language or in the form of images/videos. Here we intentionally frame the action as a language expression (i.e., a verb-noun pair) and the effect as depicted in an image in order to make a connection between language and perception. This connection is important for physical agents that not only can perceive and act, but also can communicate with humans in language and act to the environment through planning. To our knowledge, there is no prior work in this nature that attempts to connect actions (in language) and effects (in images).

As a first step, we collected a dataset of 140 verb-noun pairs. Each verb-noun pair is annotated with possible effects described in language and depicted in images (where language descriptions and image descriptions are collected separately). We have developed an approach that applies

distant supervision to harness web data for bootstrapping action-effect prediction models. Our empirical results have shown that, using a simple bootstrapping strategy, our approach can combine the noisy web data with a small number of seed examples to improve action-effect prediction. In addition, for a new verb-noun pair, our approach can infer its effect descriptions and predict action-effect relations only based on 3 image examples. This opens up the possibility for humans to teach robots new tasks through language communication and small number of examples.

1.4 Commonsense Reasoning about Physical Actions

While it is trivial for humans to use natural language to communicate about actions and changes in the physical world, machines still struggle in developing similar skills. To investigate deeper understanding of human natural language, we create a new language benchmark, which can be used to evaluate machines' capability of understanding and reasoning about human physical actions. This benchmark contains short stories created by human annotators. Each story describes a short sequence of human physical actions in our daily lives. For example, a story could describe the action sequence of making a sandwich in the kitchen, or the actions of repairing a bike in the garage. Based on the collected stories, we present two tasks for evaluating machine reading systems: the **cloze task** (selecting the correct sentence to fill in the blank in a story) and the **ordering task** (selecting the correct order of sentences in a story).

Although the proposed tasks are easy for humans to solve, they are very challenging for machines. An analysis shows that understanding the stories and solving these tasks requires various types of commonsense knowledge, e.g., knowledge about action verbs, objects, and naive physics rules. Therefore, we believe this benchmark will be a valuable resource for evaluating machines' capability of acquiring and applying physical commonsense knowledge. Further, the setting of two sub-tasks can be naturally used to evaluate models generalization ability, via training on one task and evaluating on the other task. If a model can successfully learn the fundamental knowledge and the reasoning abilities via training on the data of one subtask, it can potentially perform well on the other subtask. By doing this, we encourage models that focus on learning underlying knowledge

instead of overfitting to shallow language patterns.

A neural network model was proposed for tackling the commonsense reasoning tasks. This model solves both the cloze task and the ordering task via explicitly examining the compatibility of each action with its context in those stories. Since the action-effect knowledge plays an essential role in understanding these commonsense stories, we further incorporated physical causality knowledge into the proposed model. Experiments were designed to compare the proposed model with several state-of-the-art models for machine comprehension tasks. The results demonstrate the effectiveness of the proposed model, and further show the improvement introduced by external physical causality knowledge. The results also suggest that this benchmark is challenging for current approaches, and better solving this task requires wider range of commonsense knowledge and richer semantic representation of actions and objects.

1.5 Contributions

In this dissertation, we focus on an investigation on verb semantics from a new angle of how they may change the state of the physical world. The contributions of this dissertation is listed as below:

1. A categorization of physical causality was developed, motivated by existing theoretical linguistic studies. This categorization provides a stepping stone for systematically exploring the physical causality knowledge.
2. Two human annotated physical causality knowledge datasets were created. One dataset was annotated with causality attributes defined in this dissertation. Another dataset was annotated with open-ended natural language.
3. Two novel approaches were presented to solve the semantic role grounding task via causality modeling. The empirical results have shown the potential of causality modeling on connecting language with the physical world.

4. A physical causality embedding structure was proposed. The embedded cause-effect knowledge will allow the agent to better infer underlying causes or predict potential effects given a situation. It can be applied to answer causal questions, which are an important type of questions for artificial agents, yet not well explored in either traditional QA or Visual Question Answering (VQA).
5. The bootstrapping approach for visual causal reasoning provides a cost-efficient way to connect causality embedding with a large number of images from the web. This approach is general and can be extended to other applications involving visual causal reasoning.
6. A benchmark dataset for physical commonsense reasoning task was created. This dataset evaluates a system’s capability of understanding and reasoning about state changes in the physical world.
7. A novel approach that leverages external knowledge for the physical commonsense reasoning task were proposed. Empirical results have shown the potential of physical causality knowledge on facilitating machines to better comprehend and reason about commonsense stories.

1.6 Organization of this Dissertation

The rest of this dissertation is organized as follows. In Chapter 2, we review works from different research fields that are closely related to our study of physical causality knowledge. Chapter 3 presents our modeling of physical causality knowledge, a categorization-based approach and a language embedding-based approach. Both approaches are also evaluated on several preliminary tasks. In Chapter 4, we utilize the categorization of causality attributes and causality knowledge to improve the language grounding task. In Chapter 5, we introduce the visual causality reasoning task, as well as a bootstrapping approach that harnesses large amount of web images to tackle this task. In Chapter 6, we introduce the benchmark dataset for understanding human physical actions through natural language stories, as well as several neural models trying to solve the proposed

tasks. Finally, in Chapter 7, we summarize this dissertation and discuss several promising future directions.

CHAPTER 2

RELATED WORK

The research work in this dissertation is motivated by studies from multiple research fields, including natural language processing, theoretical linguistics, psycholinguistic, computer vision, robotics, etc.

2.1 Theoretical Linguistics on Verbs

Verb semantics have been studied extensively in linguistics [110, 73, 7, 66]. Previous work [54] has divided action verbs into *manner verbs* and *result verbs*. Result verbs usually specify a result effect of an action, which often indicates objects's *Change of State* [74]. Hovav and Levin [54] also propose that result verbs often specify movement along a scale [54]. A scale usually denotes an attribute of an object, like size, temperature, cost. For example, "Mary shortened the skirt" indicates that the length of the object *skirt* has decreased. The analysis of gradable predicates in terms of scale structure motivates us to model verb causality using object attribute categories. A detailed description of scale structure can be found in Kennedy and McNally's work [60].

Several large-scale verb lexicon databases have been built, for example, VerbNet [127], FrameNet [7] and PropBank [66]. These resources have enabled significant strides in computational semantic processing such as semantic role labeling [105, 109, 20, 175] and its applications in information extraction [29] and question answering [135]. While instrumental for text processing, the current modeling of verb semantics only plays a limited role in moving language processing towards the physical world. Despite an increasing research effort on grounding language to the environment, connections that link verbs to perception and action in the physical world are still missing. Therefore, the study of verb causality knowledge in this dissertation could be a valuable supplement to these existing knowledge bases.

2.2 Grounding Language to Perception

We humans use natural language to communicate about things in the physical world: objects, actions, events, and their properties and relations. However, for the decades research of language processing, linguistic meaning is explained mainly by symbolic models. The circular definitions in symbolic explanation of linguistic meaning restricts language meaning to only symbols. For example, if a person had to learn his first language only from a dictionary, he would be passing endlessly from one meaningless symbol to another. To ground symbol meaning in something other than just more meaningless symbols, it is the task of *symbol grounding problem* [50]. Researchers have been trying to give machines the same abilities as human, to “bridge the symbolic realm of language with the physical realm of real-world referents” [124].

Recent years have seen an increasing amount of work on grounding language to perception [171, 150, 79, 98, 78]. The common goal of language grounding researches is to enable machines to automatically acquire beliefs about the physical world and to exchange their beliefs with humans through natural language communication.

Solving the symbol grounding problem and connecting language and the physical world is fundamental to many tasks, from identifying context-dependent shifts of word meanings, to enabling situated natural language communication between human and robots. Here we give a brief review on existing works on language grounding.

2.2.1 Grounding Words in Perception

One of the most fundamental tasks in grounded language learning is to associate words with perceptual input.

2.2.1.1 Grounding to Discretized Perceptual Signal

Words are discrete symbols and perceptions are usually represented by continuous sensory data. Therefore a common way of connecting them is to discretize the sensory feature space into categories

that are associated with linguistic words. Examples include models for *grounding color names* [38, 116, 56, 88] and *grounding spatial terms* [115, 140, 47]. In computer vision, this task of associating linguistic labels with perceptual categories is usually called *recognition*, e.g., *object recognition*, *action recognition*.

Grounding color terms is an actively studied topic in linguistic and cognitive science, since color is an important type of object properties in human visual system. And the studies of color name grounding could inspire new models of learning vague meanings for other continuous domains as quantity, space, and time.

In computational systems, color is usually represented by values in a color space, e.g., red-green-blue (RGB) or hue-saturation-value (HSV) color space. A cross-linguistic study of color naming [116] shows that the color prototypes for English are close to the clusters in other different languages, e.g., white, black, red, green, yellow, and blue. This result suggests that the human perceptual system tends to have strong bias on the meaning of basic color terms. The task of grounding color terms is usually done by associating color term with a prototypical point [2] or a convex region in an underlying color space [38, 56]. Since the association between words and perception is not definitive, McMahan and Stone [88] propose a Bayesian model of color naming that takes into account the uncertainty in categorization boundaries and distributions over vocabulary.

Grounding spatial terms is another actively studied topic. Regier and Carlson [115] propose the attention vector-sum (AVS) model to predict the acceptability judgement of linguistic spatial terms given two objects in a two-dimensional space. They use vector sum representation to model the human concerning attention. Skubic et al. [140] use a histogram of force to model spatial relations between 2D objects. In Guadarrama et al. [47] and Golland et al. [43], spatial relations between 3D objects are learned through logistic regression.

Object recognition tasks in computer vision can also be seen as grounding tasks. In object recognition tasks, we assign name tags to objects in the image. Generally, this is also a process of grounding language (object labels) to perceptual signals (object images). Thanks for large scale

image recognition datasets (e.g., ImageNet [22]), computer algorithms are closing their performance gap between human on object recognition task, or even overtaking human performance [51]. However, this does not mean computers have the same level of abilities as human in grounding language to perceptions. Because there are clear limitations in those studies: algorithms are only trained to recognize a fixed set of discrete categories (usually up to thousands of classes). Their training data are provided either with explicit labels for image classification task, or with localization annotations (e.g. bounding boxes) for object localization task. Fully annotated image datasets usually cost a lot to create and they have very limited categories. The number of categories provided by vision systems is still far from human-level vision.

That said, the task of object recognition still provides us valuable knowledge and tools for more complex language grounding tasks, like the representation of images (e.g., bag-of-visual-words, pre-trained CNN features).

Attribute-based recognition has recently gained in popularity in the computer vision community. Instead of assigning name tags for objects in an image, attribute-based recognition describes object properties using learned attributes [30]. Since attributes can be shared by different objects, the learned attributes can be used to recognize novel objects with a few or zero training example [1, 57]. The attribute words can be seen as an intermediate representation that bridges the visual space and the label space, therefore they provide useful information about relations between class labels. However, the process of attribute learning also requires more human annotation efforts when collecting attributes annotation.

2.2.1.2 Learning from Ambiguous Parallel Data

Above works need to learn from fully annotated language and perception data, which are expensive to collect and usually contain very limited words. Now we look at the more general task of learning from parallel language and perception data that have some sort of ambiguity. For example, when learning from parallel image and sentence data, we need to deal with the association relation between words and image locations; in the task of situated language perception, apart from the

ambiguity in aligning language and perceptual input, we also need to deal with speech recognition errors.

Learning the joint distribution of words and image features. There are large numbers of data sets that consist of parallel image and language data. These data usually do not contain alignment between words and image regions. In order to learn word grounding from this kind of image dataset, Barnard et al. [9] present an approach that links image segments and word semantics through clustering. The clusters captures the joint distribution of words and image segments. Barnard et al. [8] study a couple models of learning the joint distribution of image regions and words, including a multi-modal extension of Latent Dirichlet Allocation. Yu and Ballard [169] use a generative graphical model to model the correspondence of word and objects. It first generates a latent variable, then visual objects are generated based on the latent variable, and finally, words are generated conditioned on visual objects.

Models of infant word learning. In the community of NLP and cognitive science, plenty researches are focused on understanding the mechanism of how human acquire language. For example, the CELL system [125] acquires word meaning from speech and image input, mimicking the language learning process of infants. In this system, word meanings are represented by prototype feature vectors along with radii around them. A prototype can be seen as an ideal point for that category in the feature space. If an input perceptual feature vector is within the radius to a prototype, it will be treated as a member of this perceptual category.

Weakly supervised object recognition/localization. As mentioned earlier, the number of categories provided by vision systems is still far from human-level visual perception. Apart from large number of object categories, real world objects also have different states. For example, a potato can be in the state of “peeled”, “in pieces”, “cooked”, etc. Recognizing object state is a more difficult task. Due to the high cost of human annotation, it is very difficult to establish an open-domain image dataset that covers a large set of objects with state annotations. Therefore people seek ways of utilizing parallel language and vision data in weakly supervised settings. With the recent success of the Deformable Parts Model (DPM) detector [106], weakly-supervised object

localization techniques [102] have risen back to popularity. These works use weak supervision from web search images to train concept detectors (localization). Utilizing automatic web-search images helps to remove the obstacle of high-cost for collecting fully annotated image dataset. In Chapter 5, we borrow the similar idea of using web search images with distance supervision to facilitate the learning of action-effect prediction.

2.2.1.3 Grounding Verbs

Based on the representation of verb meanings, verb grounding researches can be categorized into several different types, 1) representation using world state, 2) representation using action control structures, 3) representation using motion profile.

Representation using world state. World state includes object properties (e.g., color, shape) and object relations (e.g., spatial relation). In Siskind’s work [137, 138, 139], state changes in force-dynamic relations between participant objects are visually recognized, and their temporal schemas are used to infer actions (verbs). Yang et al. [164] use a visual semantic graph to represent the consequence of manipulation actions. A number of work [133, 131, 94] explicitly model verbs with predicates describing the resulting states of actions.

Representation using action control structures. In robotic studies, in order for robot to carry out an action, the verb meaning need to be grounded to the control of action. Bailey et al. [5] uses x-schema to represent action verbs, which captures verb semantics using action control structures. Misra et al. [93] propose a data driven approach to ground natural language commands to sequences of robot basic actions.

Representation using motion profiles. Motion profiles are widely used for recognizing actions through computer vision [129, 151]. Approaches in this category usually perform well in recognizing actions from human gesture and motions. However, they usually highly rely on the training data (e.g., actors, lighting conditions, camera angle), and can hardly be generalized to grounding action verbs to robot actions.

Although lots of work have been done in grounding language to perception, no previous work

has investigated the link between physical causality denoted by action verbs and the change of state visually perceived. Chapter 4 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. In this dissertation, we are particularly interested in using world state to represent verb meaning. World state changes capture the causality information of action verbs, thus they can be beneficial for action reasoning and planning. Also as shown by experimental results [139, 164], using state change can be a more robust way to model verb meanings than using motion profiles.

2.2.1.4 Context-Dependent Word Meaning

One thing to notice is that, the grounding of words could be influenced by language context, e.g., the RGB values of “red wine” and “red hair” are likely to be different. [38] models the shift of word meaning given contexts. In fact, word meanings are determined by both linguistic convention and visual perception. Experiments in [18] show that human understanding of language depends on the listener’s evaluation of how to achieve the goal based on current situation. McMahan and Stone [88] claim that it is not accurate to use definitive mappings between words and the world. Their work models speaker judgment and speaker choice in the grounding of color words.

To summarize, a typical system that learns grounded word meanings usually takes two steps: 1) “parsing” the perception into ontological types and relations that could be explained by human language semantics; 2) learning the association between word and those perception categories. Sometime these two steps are done jointly.

2.2.2 Grounding Phrases and Sentences

2.2.2.1 Referent Grounding

Reference Grounding is the task of resolving referring expressions to a referent, the entity in the physical world to which they are intended to refer. In [122], perspective-taking mechanism is used

to find the referent based on a set of language descriptors. When the given information is not sufficient to make prediction, the agent can automatically raise a question for more distinguishing information.

In [85] and [62], the objects are distinct and represented via symbolically specified properties, such as color and shape. They use pre-defined property classifiers, like ‘green’, ‘triangle’, to identify object properties. In [63], word meaning is represented by a function from object perception features to a score of how well the word and that object fit each other. Phrases are represented by compositions of individual words. For example, a simple noun phrase is composed by averaging word meaning functions, a relational phrases (e.g., the book to the left of the mug) is composed by multiplying different word meaning functions.

2.2.2.2 Grounding Action Frames

Several works from computer vision community focus on the extraction of action frames from images [48, 168]. Yatskar et al. [168] creates an image dataset *imSitu*, where images are annotated with activities and semantic roles from FrameNet. The goal is to detect the activity and localize the objects of interactions from image input.

Yang et al. [162] extend traditional semantic role labeling (SRL) to grounded SRL where arguments of verbs are grounded to participants of actions in the physical world. Using a graphical probabilistic model to jointly learn the correspondence between language and vision, their approach grounds both explicit semantic roles and implicit semantic roles. In Chapter 4 of this dissertation, we model the physical causality of action verbs from crowd-sourced data, and demonstrates that physical causality modeling helps with the grounded semantic role labeling task.

2.2.2.3 Parsing and Perception

Many researchers treat grounded language acquisition as two subproblems: parsing and perception [70]. The first step is semantic parsing, which tries to map language to formal meaning representation. One of the most commonly used meaning representation is first-order logic. The

second step is mapping the meaning representation to perception categories. Researches in this direction often focus on the semantic parsing step, assuming that we have easy access to a logical representation of the world.

Zelle and Mooney [174] propose a natural language interface for database queries. Based on the CHILL parser acquisition system [172, 173], this interface transform the natural language queries to a logic form which can be used to query the database. Here the logic form bridges the natural language and database slots. The parser acquisition is regarded as a problem of learning search-control rules in a logic program through inductive logic programming. Apart from the logic form representation, there are other forms of semantic meaning representations. In Lu et al.'s work [83], they propose using hybrid trees to represent semantic meanings, where each tree node includes both natural language words and the corresponding meaning elements.

Borschinger et al. [14] propose that the grounded language learning problem can be solved by addressing the unsupervised Probabilistic Context Free Grammar (PCFG) induction problem. In their approach, the semantic information is encoded as part of the text string. However, this approach includes every possible meaning representation constituent as a nonterminal in the PCFG. It will have difficulties when dealing with complex sentences with a large number of potential meanings, since the number of possible subgraphs grows exponentially. Later, Kim and Mooney [64] propose to address the combinatorial explosion problem by introducing the Lexeme Hierarchy Graph (LHG), where a hierarchy of semantic lexemes is build for each ambiguous landmarks plan. In another work, Lin et al. [76] study the task of retrieving videos using complex natural language queries. In their proposed approach, a sentence is first parsed into semantic graph, and objects and their motions are detected from the video, then language is matched to visual concepts using a generalized bipartite matching algorithm.

Combinatory Categorical Grammar (CCG) is a popular tool for semantic parsing. It can model both the syntax and the semantics (expressions in λ -calculus) of a sentence. Matuszek et al. [86] propose a framework that uses CCG parsing to grounds natural language sentences to perceptions. A sentence is first parsed into logical forms using probabilistic CCG. Then an explicit model is

used to align logical constants and perception attribute classifiers. However this work only study a very limited number of objects and several attributes, like color and shape.

2.2.2.4 Jointly Modeling Parsing and Perception

There are also some grounding works that do not need explicitly semantic parsing as a first step. Yu and Siskind [171] propose an unsupervised approach to ground language descriptions to video clips of human interacting with multiple objects. They use an HMM-based model to jointly learn the object tracking and word meaning grounding. In their approach, each language element (verb, noun, adjective, adverb, preposition, etc) is modeled by an HMM, and the visual perception is represented by object detection results. Later, this model was extended to handle different application tasks [170]: 1) Language generation from vision, 2) Video/image retrieval using language query, 3) Language-Guided Activity Recognition.

In [147, 148], a probabilistic graphical model was created to map natural language commands to physical world groundings, like objects, paths and locations. Each command sentence is first decomposed into Spatial Description Clauses (SDCs) [68] with types of event, object, path, and place. The system then infers groundings in the world corresponding to each SDC using a Conditional Random Field (CRF) model.

Artzi and Zettlemoyer [4] present a joint model of meaning and context for interpreting and executing instructional sentences, using a grounded CCG semantic parsing approach. The joint modeling improves grounding performance by providing situated environment cues, like the set of visible objects. Matuszek et al. [87] build a joint model of linguistic meaning and action execution in a grounded CCG semantic parsing framework. In this work, a parser is learned to map language instructions to robot control language (RCL), which can later be executed by the robot in a simulation environment.

Above work which jointly address parsing and perception has some drawbacks, including: 1) the learning phase of these models requires large amounts of manual annotation, and 2) the semantic representations are limited by pre-defined predicates [68, 147, 87]. Krishnamurthy and Kollar [70]

partially solve these limitations by introducing a Logical Semantics with Perceptron(LSP) model, which jointly models the perception information and language meanings. Their task is given an environment with multiple objects, to map natural language statements to the referents in the environment. The authors introduce a weakly supervised method to train the LSP.

2.2.2.5 Neural Network approaches

Recent researches deploy deep-learning frameworks to model both image and word sequence. End-to-end deep neural network models have shown good performance in the tasks of visual description generation [26] and visual question answering tasks [3]. One drawback of end-to-end deep learning framework is the lacking of transparency. If the system fails on one example, it is hard for human to understand the reason. Therefore, researchers have been trying to develop neural models with explicit intermediate representations, e.g., the use of attention models.

Karpathy et al. [59] present an learning approach that grounds dependency-tree relations to regions in the image using a ranking technique. Rohrbach et al. [120] use attention model to ground phrases to image regions, their model works both with or without grounding supervision. Recently people start to utilize caption generation framework on grounding task. Hu et al. [55] propose Spatial Context Recurrent ConvNet (SCRC) to transfer visual-linguistic knowledge from image captioning tasks to facilitate the grounding of language query to bounding boxes in the image.

A common approach in processing parallel image and language data is to learn an embedding model that maps text and images into a shared latent space. In this shared space, vector representations for text and images can be compared directly. Therefore it is very convenient to retrieve related images given text, or retrieve related text given an image. For example, Wang et al. [152] propose an approach called Deep Structure-Preserving Embedding, which formulates the image-sentence retrieval as a ranking problem.

2.3 Natural Language Inference Tasks

Recent years have seen a trend that new language processing benchmarks shift from only targeting linguistic context to ones requiring world knowledge and reasoning process to solve. For instance, the Winograd Schema Challenge [72] requires commonsense knowledge to resolve pronouns. Some of the machine comprehension benchmarks [153, 113, 95] require comprehending and reasoning based on supporting documents, where world knowledge can be critical in deeply understanding the documents.

Natural Language Inference (NLI) is one of the most widely studied task that focuses on understanding and reasoning about natural language. NLI is the task of reasoning about the truth given some linguistic statement and premise. Researchers have created many challenging tasks that require language understanding and inference. Solving these challenges usually requires large amount of knowledge. Next we will review some popular language inference challenges and frameworks.

2.3.1 Recognizing Textual Entailment (RTE)

The Recognizing Textual Entailment (RTE) task aims to evaluate machines' capability of determining whether the meaning of one text is entailed by another text [21]. Given two sentences A and B , we say A entails B if a human reading A would infer that B is holding true. For example:

- Text: The purchase of Houston-based LexCorp by BMI for \$2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.
 - Hyp 1: BMI acquired an American company.
 - Hyp 2: BMI bought employee-owned LexCorp for \$3.4Bn.
 - Hyp 3: BMI is an employee-owned concern.

- Text: On 18 April 1955, Aortic aneurism killed Albert Einstein. This is when blood vessels gather in the aorta stretching out this part of the heart..
 - Hyp 1: A health issue caused Einstein to die.
 - Hyp 2: The Bell Inequalities were not presented while Einstein was alive.
 - Hyp 3: Einstein was executed by Nazi Germany.

In RTE tasks, the text is always an inherent part of the inference process for predicting the answer. In other words, solving a valid RTE question always requires information from both sentences.

2.3.2 Winograd Schema Challenge (WSC)

The Winograd Schema Challenge (WSC) is proposed by Hector Levesque [72]. One of the motivation of this challenge is to provide an alternative to the Turing Test, by enforcing human-like reasoning. Unlike the Turing Test, Winograd Schema does not involve a conversation between human and machine. Instead, the machine needs to answer binary questions. For example:

- The trophy would not fit in the brown suitcase because it was too big (small). What was too big (small)?
 - Answer 0: the trophy
 - Answer 1: the suitcase
- The town councilors refused to give the demonstrators a permit because they feared (advocated) violence. Who feared (advocated) violence?
 - Answer 0: the town councilors
 - Answer 1: the demonstrators

Note there is a word (“big”, “feared”) in each of the original sentence. If we switch this word with its alternative (“small”, “advocated”), the correct answer becomes the opposite. In the first

example, when “big” is used in the sentence, the answer is “the trophy”; when “small” is used in the sentence, the correct answer is “the suitcase”.

The questions in Winograd Schema Challenge are carefully designed to test a system’s world knowledge and human-like reasoning capability. Levesque introduces guidelines of creating the questions: 1) The questions are easy for human to answer; 2) The questions cannot be solved by coreference resolution techniques like selectional restrictions; 3) Mining statistical measures from large text corpora will not suffice to solve them. These guidelines provide valuable insights to the process of creating natural language understanding benchmarks. Recently, people are becoming more aware of the problem that machine learning models are often memorizing shallow statistical cues instead of truly understanding natural language [58, 100]. In Levesque’s design, a swap of the special word with its alternative leads to opposite ground-truth answer. This is closely related to the idea of building adversarial examples to eliminate the power of memorizing shallow statistical cues [100].

2.3.3 Causal Reasoning

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 [42, 156] have been studied extensively as well as their linguistic expressions [155, 141, 99] and automated extraction of causal relations from text [12, 97, 111, 118].

2.3.3.1 Choice of Plausible Alternatives (COPA)

The Choice of Plausible Alternatives (COPA) task [119] is created to evaluate a system’s capability of dealing with commonsense causal reasoning. It contains a thousand questions and each question contains a premise and two alternatives. The task is to select one of the alternatives that is most likely to be the cause or effect of the premise. Here are some examples:

- Premise: The man broke his toe. What was the CAUSE of this?

- Alternative 1: He got a hole in his sock.
- Alternative 2: He dropped a hammer on his foot.
- Premise: I knocked on my neighbor’s door. What happened as a RESULT?
 - Alternative 1: My neighbor invited me in.
 - Alternative 2: My neighbor left his house.

The COPA is a relatively challenging task in Natural Language Understanding. Since solving these questions requires more commonsense knowledge than the training data can provide. And causal relations in text are usually difficult to obtain. Common ways of acquiring causality knowledge adopted by existing approaches includes automatic extraction of causal relations, and pre-training on large-scale external datasets. Luo et al. [84] proposed an text data-based approach for the COPA task. They extracted causal-effect terms from a large web corpus and their approach achieves 70.2% accuracy. Li et al. [75] train a neural network model using data from two external datasets. They feed the model with training examples from other tasks, via transforming data from other tasks into COPA-style plausibility questions.

2.3.3.2 Story Cloze Test

The Story Cloze Test [95] consists of five-sentence stories with two alternate endings, requiring a system to decide which ending is more plausible. Below is an example story:

Context: Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.

- Right Ending: Karen became good friends with her roommate.
- Wrong Ending: Karen hated her roommate.

This benchmark contains lots of everyday events, which is related to a wide variety of commonsense knowledge. Most of the stories focus on human’s emotions, intentions and attitudes, i.e., naive psychology. In this dissertation, we are more interested in physical commonsense knowledge.

There are also many research works modeling cause-effect relations [40, 19, 24, 163], particularly for question answering (e.g., addressing *why* questions). Most of these works address high-level causal relations between events, for example, “the collapse of the housing bubble” causes the effect of “stock prices to fall” [130]. They do not concern the lower-level cause-effect relations associated with concrete actions. Different from these previous works, this dissertation has a specific focus on the physical causality of action verbs, in other words, change of state in the physical world caused by action verbs as described in [54].

2.4 Knowledge Resources

One bottleneck towards natural language inference is that machines lack world knowledge. Thus, there is an increasing amount of effort on developing knowledge representations and building knowledge resources. In this section, we discuss some of the major knowledge resources. Based on the approaches creating the knowledge resources, we categorize them into *hand-built knowledge resources* and *automatically extracted knowledge resources*.

2.4.1 Hand-built Knowledge Resources

WordNet [92] is a large lexical database for English. Words are grouped into synsets and then organized into a network structure. Each of the synsets is linked with other synsets by some conceptual relations. For nouns and verbs, they are arranged into hierarchies, with hypernymy and hyponymy representing links going up or going down the hierarchies. WordNet contains 117,000 synsets and provides different types of knowledge regarding the relations between them. However, WordNet mainly focus on conceptual-semantic and lexical relations, it does not contain commonsense knowledge about how the world changes.

ConceptNet [81] is a network that records large amounts of commonsense knowledge. Here

the term commonsense refers to “the millions of basic facts and understandings possessed by most people.” ConceptNet is built based on the human annotated database Open Mind Common Sense (OMCS) [136]. Extraction rules are designed to automatically extract ConceptNet’s binary relations from the OMCS sentences. Later, ConceptNet has grown to include knowledge from other human-built resources. In ConceptNet 5.5 [142], it contains over 21 million links between over 8 million nodes. Although ConceptNet has grown rapidly in size and coverage, it is still far from obtaining human-level commonsense knowledge. For example, causal relations are still sparse in ConceptNet, making it difficult to making inference in a human-like way.

2.4.2 Automatically Extracted Knowledge

Human annotation is usually very expensive to obtain. Therefore a large amount of studies have been done to automatically extract knowledge from existing large collections of data sets. Except for few that acquires knowledge from images [167], most of the previous effort apply information extraction techniques to extract facts from a large amount of text data [27, 112]. A commonly adopted way is to discover relations between named entities and automatically extract facts about those entities from the raw textual data [27, 112]. DBPedia [71], Freebase [13], and YAGO [144] extract structured information from document repositories on Wikipedia. Wikipedia is an ideal resource for knowledge extraction since it is maintained by a large community and it contains semi-structured documents that have great semantic heterogeneity.

These automatically minded knowledge base cover millions of facts about the world such as people and places, saving a lot of time and expenses compared with human annotation. However they emphasize more on relations and properties related to named entities (e.g., places, people, and organizations). They do not contain an important type of commonsense knowledge, which people usually do not mention explicitly, but is still critical for our communication. Physical causality knowledge is among this type.

2.5 Related Work in Computer Vision and Robotics

The idea of modeling object physical state change has also been studied in the computer vision community and the robotics community.

2.5.1 Related Work in Computer Vision

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 [69] and causality between actions and objects [31]. Fathi and Rehg [31] have broken down detection of actions to detection of state changes from video frames. Yang and colleagues [164, 165] 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 [34] have developed a framework to learn perceptual causal structures between actions and object statuses in videos. However these previous works only focus on the visual presentation of motion effects. In Chapter 5, we aim to make a connection between visual presentation and human language descriptions.

Recent years have seen an increasing amount of work integrating language and vision, for example, visual question answering [3, 35, 82]. Different approaches have been developed such as Multimodal Compact Bilinear Pooling (MCB) [35], Dynamic Memory Network [158], and the use of external knowledge bases [157]. Most of these work mainly focus on the Yes/No questions and *what* type questions related to object recognition. While many approaches require a large amount of training data, more recent works have developed zero/few shot learning for language and vision [96, 160, 159, 161, 149]. Different from these previous works, in Chapter 5 of this dissertation, we introduce a new task that connects language with vision for physical action-effect prediction, focusing on the causal relation between actions and state changes depicted by both language and visual data.

2.5.2 Related Work in Robotics

In the robotics community, an important task is to enable robots to follow human natural language instructions. Previous works [134, 94, 131, 132] explicitly model verb semantics as desired goal states and thus linking natural language commands with underlying planning systems for action planning and execution. In these works, action schemas are defined to capture the change of state caused by a given action. Based on action schemas and the goal state, planning algorithms can be applied to find a sequence of actions to achieve the goal [39]. Therefore, the state of the physical world is a very important notion and changing the state becomes a driving force for robot's actions.

However, these studies were carried out either in a simulated world or in a carefully curated simple environment within the limitation of the robot's manipulation system. And they only focus on a very limited set of domain specific actions which often only involve the change of locations. In Chapter 5 of this dissertation, we study a set of open-domain physical actions and a variety of effects perceived from the environment (i.e., from images).

CHAPTER 3

MODELING PHYSICAL CAUSALITY OF VERBS

In order to enable a robot or a computer to acquire and utilize the physical causality knowledge of verbs, we need to collect this kind of knowledge and transfer it into machine-understandable representations. Usually the most natural way for human to pass knowledge is through language. However, physical causality knowledge is usually not explicitly stated in human language, since we assume everyone possesses this kind of common sense knowledge. This makes it difficult to extract physical causality knowledge from existing language datasets, like large-scale text corpora. Therefore, in this study, crowd-sourcing tasks were designed to collect physical causality data. In these crowd-sourcing tasks, human subjects were asked to explicitly express their knowledge on action verbs, through natural language or through answering designed multiple choice questions.

After data collection, we investigate two different approaches to model physical causality knowledge. In one approach, the changes of state are categorized into classes, and the causality knowledge for an action verb is represented as its associations with those changes of state classes. In another approach, language descriptions of actions and their effects are embedded using neural network into a common vector space. The causality knowledge is modeled through similarities between embedding vectors.

3.1 Categorization of Physical Causality ¹

3.1.1 Linguistics Background on Action Verbs

Verb semantics have been studied extensively in linguistics [110, 73, 7, 66]. In this dissertation, we only focus on concrete action verbs (such as *run*, *throw*, *cook*), which denote physical actions in the world, instead of denoting states or abstract actions that can not be visually perceived. Hovav and Levin [54] propose that action verbs can be divided into two types: *manner verbs* that “specify

¹This is a joint work with Malcolm Doering. Part of this section (Section 3.1.1, 3.1.2 and 3.1.3) is also included in Doering’s Master of Science dissertation [25].

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, *Inherently Directed Motion* verbs and *Incremental Theme* verbs [74]. *Change of State* verbs denote a change in the property of object (e.g. “to melt”). *Inherently Directed Motion* verbs indicate a movement in regard to a landmark object (e.g. “to arrive”). *Incremental Theme* verbs stand for the incremental change of object, like mass, volume or area change (e.g. “to eat”). This dissertation has a main focus on result verbs. Unlike Levin and Hovav’s definition of *Change of State* verbs, here the term *change of state* is used 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 [54]. A scale usually denotes an attribute of an object, like size, temperature, cost. For example, “Mary shortened the skirt” indicates that the length of the object *skirt* has decreased. A detailed description of scale structure can be found in Kennedy and McNally’s work [60].

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 [23]. 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 work.

3.1.2 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 [117]. This corpus is a joint data of text and videos, where the videos capture different human subjects doing

cooking activities, and the text sentences describe the actions of the human subjects. The TACoS dataset contains mainly descriptions of physical actions, and a majority of the verbs belong to result verbs, which denote some changes of state that can be observed in the world. Therefore the TACoS dataset is very suitable for our study.

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 want to investigate the verb causality knowledge under different contexts.

Each verb-noun pair was presented to human annotators 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 annotators 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 annotators' responses, which resulted in a total of 1800 natural language responses describing changes of state.

3.1.3 Categorization of Change of State

Based on Dixon and Aikhenvald's typology for adjectives [23] and human annotators' responses, we identified a categorization to characterize causality, as shown in Table 3.1. 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 some possible values 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

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 3.1: Categorization of physical causality.

corresponding to eighteen attributes as shown in Table 3.1.

Through analyzing the crowd-sourcing data, we have made several interesting and important observations:

1. A verb can be associated with multiple changes of state. Our data show that each human description contains as many as three different changes of state. 43% descriptions contained only a single change of state, and 36% descriptions contained no change of state. 19% described two CoS and 2% described three CoS. Since our causality categories are mainly designed to capture low-level states, they do not include higher level attributes like *cleanliness*. For some of those descriptions counted as no change of state, they actually describe changes of high level state attributes.

Figure 3.1 shows the distributions of causality labels applied to two verbs, *clean* and *rinse*. Intuitively, these two verbs have similar meanings. As shown in the figure, their distribution of causality labels are also similar. They both have high weights on *PresenceOfObject* and *Wetness*.

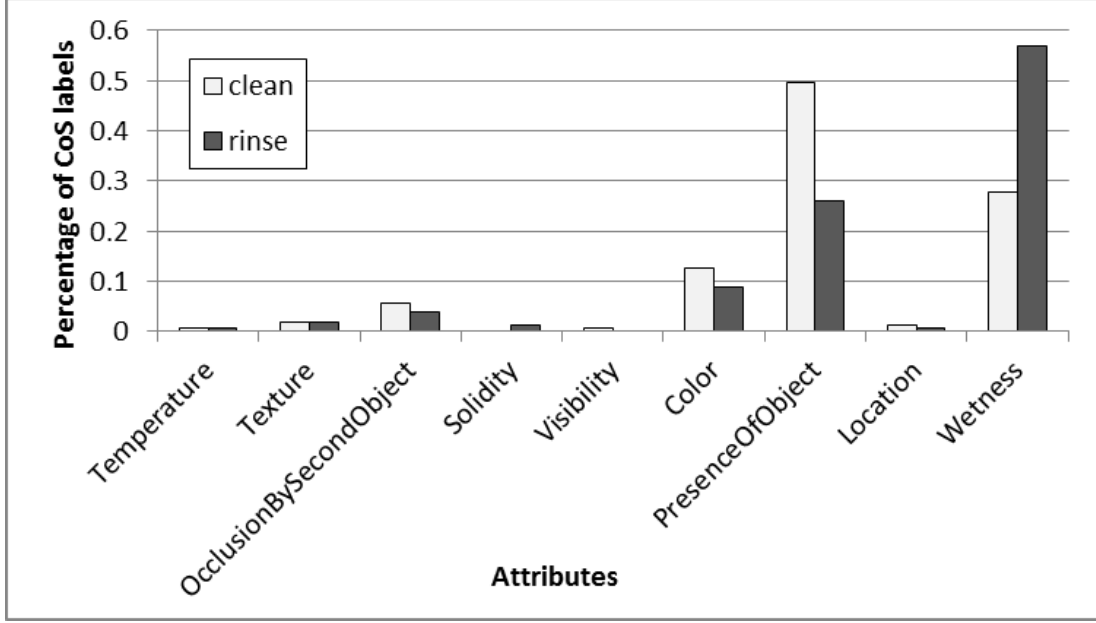


Figure 3.1: Distributions of causality labels for verbs *clean* and *rinse*.

However, the differences between these two verbs are also captured by the distributions. Human tends to describe the effect of *clean* with more *PresenceOfObject* label, and to describe the effect of *rinse* with more *Wetness* label. Partly because the result verb *clean* is more related to the final state that “dirt is no longer present”, while the manner verb *rinse* is more related to the use of water.

2. The causality for a verb is context dependent. Human’s description of verb causality not only depends on the nouns filling a particular semantic role for the verb, but also on the physical scenes where the verb-noun pairs appear in. Based on the collected data, we developed a metrics called *variability* using Jensen-Shannon divergence (JSD) to compare the distributions of causality labels associated to a verb under different conditions (e.g., taking different nouns or whether a video clip was shown or not).

The JSD of two distributions P and Q is defined as below.

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M), \quad (3.1)$$

where $M = (P + Q)/2$, and D is the well-known Kullback-Leibler divergence. JSD is a symmetric measure (i.e., $JSD(P||Q) = JSD(Q||P)$). The smaller JSD means two distributions are more similar.

Verb	+/-Scene	3 Objects
clean	0.03	0.04
rinse	0.01	0.05
wipe	0.02	0.14
cut	0.01	0.02
chop	0.02	0.03
mix	0.05	0.13
stir	0.09	0.21
add	0.12	0.22
open	0.09	0.32
shake	0.18	0.42

Table 3.2: Variability of causality labels over different object and scene conditions.

Variability describes how the causality label distributions of a verb vary with different conditions. The conditions include filling the patient role with different objects, or whether a video clip was shown to the human annotators. The variability is defined as below.

$$variability = \frac{\sum_{(i,j), i \neq j} JSD(d_i, d_j)}{num\ pairs} \quad (3.2)$$

where (i, j) indicates a pair of two distributions. The variability is an average of JSD for all pairs of causality label distributions. Each pair of distributions correspond to a pair of different values for the context variable. For example, the variability over the object conditions of a verb *chop* is calculated by averaging the JSD of three unique pairs of causality label distributions, i.e., averaging over $JSD(chop\ cucumber, chop\ bean)$, $JSD(chop\ cucumber, chop\ leek)$, and $JSD(chop\ bean, chop\ leek)$.

The variabilities over objects and scenes for each verb are shown in Table 3.2. The second column of the table shows that with or without video clips can influence human’s judgement of action effects. And some verbs (e.g., shake) are more sensitive to the changes of visual scenes. The third column of the table shows that for some verbs, their causality information is also closely related to the objects. These observations indicates that the causality of a verb depends on its context.

3. Causality models can be used to reflect similarities between verbs. Based on the data, we further applied Jensen-Shannon divergence (JSD) to calculate the divergence of causality label

distributions between different verbs. Our results indicate that similarity between verbs based on causality distributions is consistent with similarity based on verb semantics, for example, for two similar verbs $JSD(cut, chop) = 0.01$ and $JSD(mix, stir) = 0.03$, for two dissimilar verbs, $JSD(cut, shake) = 0.59$ and $JSD(rinse, chop) = 0.68$. This shows that the causality labels for verbs, while adding another dimension to verb semantics, still preserve the original meaning of these verbs.

In summary, the results from our empirical studies, although preliminary due to a small dataset, have shown it is possible to systematically model causality knowledge for a set of common verbs through crowd-sourcing studies. These results have motivated us to conduct more in-depth investigations on modeling and utilizing causality knowledge.

3.1.4 Evaluation: Verb Similarity Judgement and Thematic Fit Estimation

In this section, we demonstrate verb causality categorizations can potentially improve semantic modeling of verbs based on distributional semantics.

We collected a larger dataset of verb causality annotations based on sentences from the TACoS Multilevel corpus [121], through crowd-sourcing on Amazon Mechanical Turk. Annotators were shown a sentence containing a verb-object 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.

The dataset contains 4391 sentences, where there are 178 verbs, 260 nouns, and 1624 verb-object pairs. Note that multiple sentences could have a same verb-object pair. Each verb-object pair always contain a single verb, but could have two or more object nouns. 41.6% of the verb-object pairs contain two or more object nouns, e.g., “move-egg, bowl”. The causality annotation result for each sentence is represented as a 18-dimension binary valued vector, each dimension is 1 if at least two annotators labeled the corresponding causality attribute as true, 0 otherwise. In 80% of the vectors, only one dimension is 1, showing that on most sentences, at least two annotators agreed on

one causality attribute. In 3% of the vectors, more than one dimensions are 1, meaning at least two annotators agreed on more than one attributes. 17% of the vectors are zero vectors, meaning there is no agreed attribute between three annotators. All the experiments in this section are conducted on this dataset.

If an annotator believes that none of the 18 attributes is applicable to the verb, he/she has other choices of selecting “Current change of state frame is not applicable” (CoS-NA), or “No change of state” (No-CoS). In the overall annotation results, less than 1 percent of instances are labeled with CoS-NA or No-CoS, illustrating that the coverage of the proposed causality label categorization is quite thorough.

3.1.4.1 Verb Similarity Judgement

Distributional Semantic Models (DSM) [10, 17] use contextual distributions to represent word meaning. However, only using contextual information does not provide a complete picture of word meaning. In this section, we augment verb representation with causality information, and evaluate the performance of augmented models with human annotated verb similarity dataset. Since causality information captures possible change of the state of the physical world denoted by the verb, it could be a good supplement to the contextual information of DSM in terms of verb semantics.

For each verb, we use a vector of 18 dimension to represent its causality information. The vector is obtained by averaging all causality vectors of the sentences that contain this verb. For the contextual information, we adopt the Distributional Memory (*typeDM*) [10], from which we can get a vector representation for each verb. DM was constructed from three large-scale corpora, ukWaC, WackyPedia and BNC. To assemble the DM vector F_t and causality vector F_s , we use the linear weighted combination function from [17]:

$$F = \alpha \times F_t \oplus (1 - \alpha) \times F_s \quad (3.3)$$

where \oplus is the vector concatenation operator. The parameter α can be determined from a develop-

ment dataset.

As there is no existing word similarity dataset that has a good coverage on the verbs we study, we need to develop new benchmarks. Following previous work on similarity measurement [16, 10], we developed two benchmarks. Each of the benchmark contains 378 pairs of frequent verbs in the TACoS dataset. Each pair has an averaged similarity score obtained by crowd-sourcing on the Amazon Mechanical Turk. Ten different annotators were asked to rate each verb pair with a scale between 1 to 5. For example, the pair *cut-slice* receives a high average rating 4.2, *clean-pull* receives a low rating 1.2, in one of the benchmarks. The only difference between the two benchmarks is that, during the collection of one, annotators were informed that these verbs describe cooking activities in the kitchen, while no such information is provided during the other one. In this way, we can get human judgement of verb similarity both in a specific domain and in general domain.

We evaluate the models of verb meaning in terms of their Spearman correlation to the human rating benchmarks. Cosine similarity is used to measure the similarity between two verbs in these vector models. In order to tune the parameter α , the general domain benchmark was divided into development set and test set, each contain half of the data. We found the optimal value of α is around 0.5 on the development set. We set $\alpha = 0.5$ for the experiments. Tabel 3.3 reports the evaluation results. No significant differences were observed between results on two benchmarks. As expected, the concatenation model (DM+Causality) clearly outperforms the DM model on both benchmarks. This illustrates the effectiveness of causality information in capturing verb meaning.

Model	General Domain	Cooking Domain
DM	0.4460	0.4382
DM+Causality	0.5554	0.5328

Table 3.3: Results of verb similarity judgement task using Distributional Memory (DM) model, and concatenation model (DM+CoS). (Pearson’s correlation ρ , all values are significant with $p < 0.001$.)

3.1.4.2 Thematic Fit Estimation

First we define the *causality vector* for a verb and the *affordance vector* for a noun. The causality vector for a verb is the same vector defined in the last application, which is calculated by averaging all causality vectors of the sentences that contain this verb. This vector shows the possible changes of state to the physical world caused by the verb. The affordance vector for a noun is calculated through averaging all causality vectors of the sentences that contain this noun as patient role. This vector shows the possible change of state for the corresponding object as a consequence of actions. Thus, we can estimate the possibility of a noun being the patient role (or direct object) of a verb by calculating the similarity between the two vectors. For example, the causality vector of “cut” has a heavy weight on the causality attribute “Quantity”, and the affordance vector of “carrot” also has a heavy weight on the same attribute, thus we can tell “carrot” fits well as the object of verb “cut”.

I implemented *typeDM* [10] for comparison, since it has shown state of the art performance in thematic fit tasks [10, 46]. In *typeDM* model, to determine how well a noun fits the patient role of a verb, we first find out 20 most popular nouns for the patient role of the verb, by counting the syntactic dependence links of *object*. Then a centroid is calculated through normalizing and averaging the DM vectors of the 20 nouns. The thematic fit score is the cosine similarity between the DM vector of the target noun and the centroid.

To show the advantage of including the causality information in thematic fit estimation task, again we adopted the concatenation model from Equation 3.3 to integrate DM and CoS information ($\alpha = 0.5$). The cosine similarity between vectors is used to measure the possibility of a noun being the object of a verb.

Since there is no existing thematic fit benchmark that has a good coverage on the verbs and nouns we study, we created a new dataset of human judgements on thematic fit of patient role, following previous work on thematic fit estimation [89, 104]. 32 verbs and 36 nouns were sampled from the TACoS Multilevel dataset. These verbs and nouns were used to randomly construct 520 verb-noun pairs. Each verb-noun pair was rated by 5 different annotators from Amazon Mechanical Turk. They rated the pair based on the plausibility of the noun as patient of the verb. The rating was

on a scale 1 to 5, and judgement were then averaged from 5 annotators (e.g., *cook-broccoli* receives a high average rating 4.8, *cut-salt* receives a low rating 1.0). Tabel 3.4 reports the evaluation results. The concatenation model significantly outperforms DM model, indicating that causality information play an important role in measuring thematic fitness.

Model	Pearson's ρ
DM	0.3007
DM+Causality	0.3732

Table 3.4: Results of thematic fitness estimation using Distributional Memory (DM) model and concatenation model (DM+Causality). (Pearson's correlation ρ , all values are significant with $p < 0.001$.)

The above two applications of modeling physical causality of verbs illustrate that this kind of knowledge is an important complement to the distributional semantics of verbs. It can be used not only to measure similarity among words, but also to capture more abstract semantic relations.

3.2 Modeling Causality Knowledge via Embedding Methods

Previous discussions have shown the potential of modeling verb causality knowledge using pre-defined categories. However, the most natural way for humans to communicate and pass knowledge is through open-ended language. In this section, our goal is to directly model causality knowledge from human natural language, instead of manually translating natural language into pre-defined categories.

3.2.1 Cause-Effect Data Collection

As mentioned earlier, the commonsense causality knowledge associated with concrete action verbs is often pre-supposed and not explicitly stated in language. It is difficult to extract cause-effect data from existing text collections. Therefore, we applied human computing and collected a set of cause-effect data through crowd-sourcing.

We started with the top 1000 frequent English verbs from the Corpus of Contemporary American English. By querying these verbs from two dictionaries (LDOCE dictionary and the American

Cause Text	Effect Text
slice bread	The bread went from being a solid loaf to several pieces.
file nails	The nails became smooth.
fry potato	The potatoes become crisp and golden and go from raw to cooked.
stain carpet	There is a visible soiled mark on the carpet.

Table 3.5: Example cause and effect text from our collected data.

Heritage 3rd edition) using patterns provided by the dictionary (e.g., transitive verbs with direct object), we identified a subset of verbs which take concrete nouns as their *patient* (in another word, *direct object*). We then extracted all example sentences for this subset of verbs from the dictionaries. Finally, two undergraduate students manually extracted the verb and its patient (i.e., the noun that serves as direct object) from each example sentence to form a *verb-noun* pair (or referred to as *verb-patient* pair). Only those *verb-noun* pairs where the verb has a clear effect on the state of the world related to the noun are chosen for our crowd-sourcing data collection. This process has resulted in a total of 558 verb-noun pairs with 251 different verbs and 356 different nouns.

The crowd-sourcing data collection was carried out on Amazon Mechanical Turk. Annotators were shown a verb-noun pair, and they were asked to use their own words to describe what changes might occur to the object (denoted by the noun) as a result of the action (denoted by the verb). Each verb-noun pair was annotated by 10 different annotators, which has led to a total of 5580 effect descriptions. Table 3.5 shows some examples of collected effect descriptions.

3.2.2 Causality Embedding Models

We propose a text embedding method to model verb causality knowledge. The structure of our model is shown in Figure 3.2. It is composed of two sub-networks: one for verb-noun pairs (i.e., *cause*) and the other one for effect descriptions (i.e., *effect*). The *cause* and *effect* can be represented either by words or phrases (as explained later) using their pre-trained embeddings v_c and v_e . The pre-trained embedding is fed to a fully-connected layer and transformed into a new (or adapted) cause embedding \hat{v}_c and a new effect embedding \hat{v}_e . The adapted embeddings are learned by

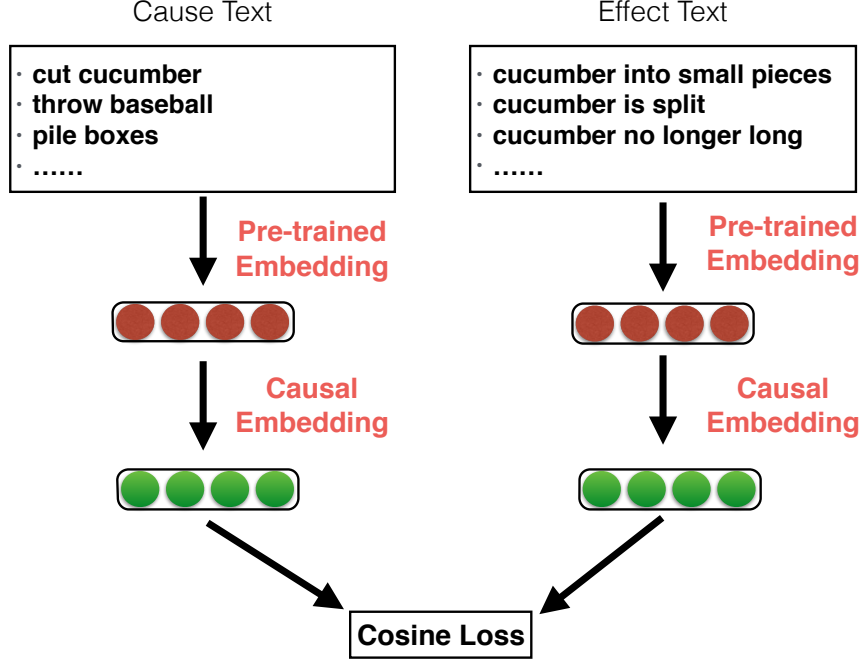


Figure 3.2: Architecture of the verb causality embedding model.

minimizing the following loss function l :

$$l = [s(\hat{v}_c, \hat{v}_e) - \gamma]^2, \quad (3.4)$$

where

$$\gamma = \begin{cases} 1, & \text{if } (c, e) \in C \\ s(v_c, v_e), & \text{if } (c, e) \notin C \end{cases} \quad (3.5)$$

$s(\cdot, \cdot)$ is the cosine similarity between vectors. C is the set of cause-effect tuples in our collected data. Suppose c is an input for *cause* and e is an input for *effect*, this loss function will learn a new cause and effect space that maximizes the similarities between c and e if they have a cause-effect relation (i.e., $(c, e) \in C$) while maintaining their original similarity if they don't have a cause-effect relation (i.e., $(c, e) \notin C$). Essentially this approach learns an adaptation of the original embedding space, which is encoded by two nonlinear transforms. To prevent the overfitting problem, we also add a dropout layer with 0.5 probability to the input of the adapted embedding layer.

As mentioned earlier, *cause* and *effect* can be represented by either words or phrases as follows.

Example patterns	Example extracted <i>State Phrases</i> (bold)
VP with a verb $\in \{\text{be, become, turn, get}\}$	The ship is destroyed .
VP + PRT	The wall is knocked off .
VP + ADVP	The door swings forward .
ADJP	The window would begin to get clean .
PP + NP	The eggs are divided into whites and yolks .

Table 3.6: Example patterns that are used to extract state phrases (bold) from sample sentences.

Word Causality Embedding (cEmbedWord). In this setting, the cause and effect text are first broken into words. For a verb-noun pair (as *cause*) and one of its effect description (as *effect*), after filtering out stop words, each word in the verb-noun pair is coupled with each word in the effect description to generate a cause-effect tuple. In this setting, we use the 300-dimension Word2Vec [90] weights pre-trained on Google News corpus.

Phrase Causality Embedding (cEmbedPhrase). In this setting, we first apply chunking (shallow parsing) using the SENNA software [20] to break an effect description into phrases such as noun phrases (NP), verb phrases (VP), prepositional phrases (PP), adjectives (ADJP), adverbs (ADVP), etc. After examining the syntactic structure of the collected effect descriptions, we found that most of the descriptions follow simple syntactic patterns. For a *verb-noun* pair, around 80% of its effect descriptions start with the same noun as the subject. In an effect description, the change of state associated with the noun is mainly captured by some key phrases. For example, an adjective phrase usually describes a physical state; verbs like *be, become, turn, get* often indicate a description of change of the state. Based on these observations, we defined a set of patterns to identify phrases that describe physical states of an object. We call these phrases *state phrases*. Table 3.6 shows some example patterns to identify state phrases and example state phrases that were extracted based on the patterns. Besides, if an effect sentence begins with a noun phrase as the subject, we also concatenate that noun phrase with each of the extracted state phrases.

After extracting state phrases from an effect description, we couple the corresponding verb-noun phrase with each of the extracted state phrases to form a (cause, effect) tuple. If no phrase is extracted from an effect description, we treat the whole description as a long phrase to form the

tuple. We encode phrases into vector representations using *skip-thought*, an RNN pre-trained on a large-scale book corpus [67].

3.2.3 Evaluation: Causality Embedding in Causal QA

The learned causality embedding can be applied to Causal Question Answering (cQA). This cQA is different from traditional question answering that involves cause-effect relations between high-level events. It was mainly designed to test machines/artificial agents’ ability in causal reasoning related to concrete actions. More specifically, we evaluate two types of questions.

Cause-to-Effect (Cause2Effect) questions. Given a verb-noun phrase, the question is: “what would likely happen to the object denoted by the noun as a result of the action denoted by the verb?” The answer would be an effect description describing the potential effect.

Effect-to-Cause (Effect2Cause) questions. Given a description illustrating a state of the world, the question is: “what action would likely cause the state of the world described in the text?” The answer would be a verb-noun pair that can potentially serve as the cause.

3.2.3.1 Ranking Algorithm

We adopt a simple ranking algorithm to retrieve answers for a question. Given a query q (i.e., either a verb-noun pair for Cause2Effect questions or a description of the state for Effect2Cause questions), we rank all candidate answers a based on their similarity score with the query in the embedded space as in the following.

$$score(q, a) = \frac{1}{|q|} \sum_{c \in q} \max_{e \in a} s(\hat{v}_c, \hat{v}_e) \quad (3.6)$$

where $s(\hat{v}_c, \hat{v}_e)$ is the cosine similarity between two words (or phrases) c, e in the causality embedding space, $|q|$ is the number of words (or phrases) in the query.

3.2.3.2 Dataset

We use our collected data (described in Section 3.2.1) for this task. The verb-noun phrases were split into 70%, 10%, and 20% for training, development, and testing respectively. The model parameters were selected based on the performance on the development set. Note that each unique verb-noun pair only appears in one of the training, validation and testing sets. The goal here is to evaluate whether the learned causality model can be applied to answer questions related to *unknown* verb-noun pairs.

3.2.3.3 Models for Comparison

We compare the following models:

- (1) The word embedding model (cEmbedWord) described in Section 3.2.2. The dimension of pre-trained Word2Vec embeddings is 300. The dimension of new word embeddings is set to 100. During training the negative sampling ratio is set to five. That is, for each positive cause-effect sample, five negative samples are created through random sampling.
- (2) The phrase model (cEmbedPhrase) described in Section 3.2.2. The dimension of pre-trained skip-thoughts embeddings is 4800. The dimension of new phrase embeddings is set to 800. The negative sampling ratio is set to five during training.
- (3) A baseline causal alignment model (cAlign). Alignment models have been successfully used in traditional QA tasks [146, 166, 130]. Here we use IBM Model 1 [15] and GIZA++ tool [101]. This baseline model is trained to “translate” questions to answers, using the question-answer training set.
- (4) A random baseline to show the absolute lower bound.

3.2.3.4 Evaluation Results

The above models were first trained using the training data. As a ranked list of answers is retrieved, we apply mean average precision (MAP) as an evaluation metric.

	Cause2Effect	Effect2Cause
cEmbedPhrase	0.7274	0.8132
cEmbedWord	0.7909	0.6478
cAlign	0.4498	0.4723
random	0.0144	0.0494

Table 3.7: MAP results for verb causality question answering task.

Table 3.7 shows the evaluation results. In general, our embedding models demonstrate good performance, considering that all the verb-noun pairs in the test data have never been seen in the training and validation data before. Both models significantly outperform the baseline (cAlign). This suggests that embedding models have a good potential in modeling physical causality knowledge.

CHAPTER 4

PHYSICAL CAUSALITY MODELING FOR LANGUAGE GROUNDING TASK

4.1 Introduction

Although recent years have seen an increasing amount of work on grounding language to perception [171, 150, 79, 98, 78], no previous work has investigated the link between physical causality denoted by action verbs and the change of state visually perceived. In this chapter, we intend to address this limitation and examine whether the causality denoted by action verbs can provide top-down information to guide visual processing and improve language grounding.

In the language grounding task, the input is parallel language and visual data, and the goal is to ground language components to entities in the visual data. Our expectation is that the categorization of physical causality can provide guidance for visual processing: once a parallel language and visual data about an action is given, the potential causality of the verb or the verb-noun pair can trigger some visual detectors that mainly focus on the potential state changes caused by this action. And applying these visual detectors to the visual data can potentially improve the performance of grounded language understanding.

Based on the categorization of physical causality attributes, we designed a set of change-of-state detectors to detect the corresponding changes from video data. We further applied two approaches, a knowledge-driven approach and a learning-based approach, to incorporate causality modeling in grounding. The empirical results have demonstrated that both of these approaches achieve significantly better performance compared to previous approaches. Moreover, we have shown that causality knowledge for verbs can be generalized to novel verbs through simple learned models.

This chapter has been published in the following paper: Qiaozi Gao, Malcolm Doering, Shao-hua Yang, and Joyce Chai. Physical causality of action verbs in grounded language understanding. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 1814-1824. 2016.

4.2 Visual Detectors based on Physical Causality

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.

The changes of state associated with the eighteen attributes can be detected from the physical world using various sensors. In this work, 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 4.1. 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.

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 4.1: Causality detectors applied to *patient* of a verb.

4.3 Verb Causality in Language Grounding

In this section, we demonstrate how verb causality modeling and visual detectors can be used together for a language grounding task. As shown in Figure 4.1, 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 [162]. 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 $\lambda_1, \lambda_2, \dots, \lambda_k$ to object (or hand) tracks $\gamma_1, \gamma_2, \dots, \gamma_n$, in the video¹. We applied two approaches to this problem.

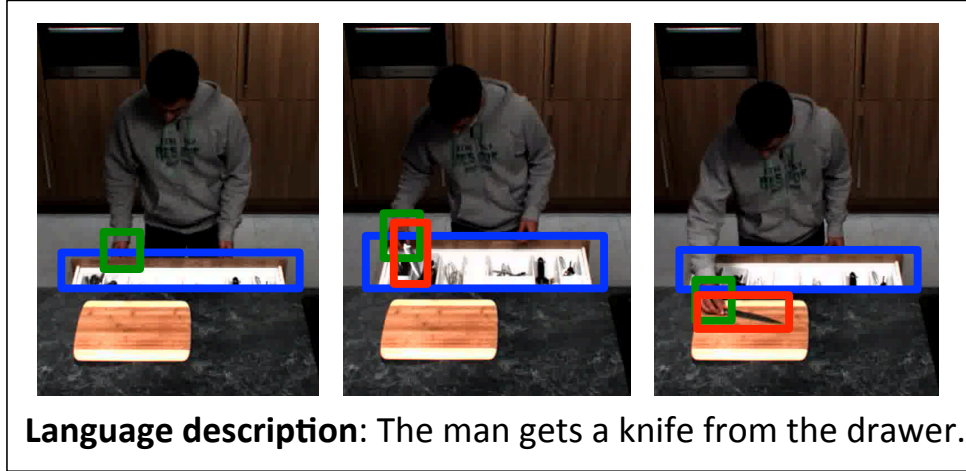
4.3.1 Knowledge-driven Approach

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.

4.3.1.1 Acquiring Knowledge

To acquire knowledge of verb causality, we collected a larger dataset of causality annotations based on sentences from the TACoS Multilevel corpus [121], 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.

¹For 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.



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

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* $\mathbf{c}(v)$ for each verb v by averaging all causality vectors of the sentences that contain this verb v .

4.3.1.2 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., $\mathbf{c}(v)$, for the verb. For each candidate track in the video, we can generate a causality detection vector $\mathbf{d}(\gamma_i)$, using the pre-defined 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 $\mathbf{c}(v)$ and $\mathbf{d}(\gamma_i)$.²

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

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 4.2. 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

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

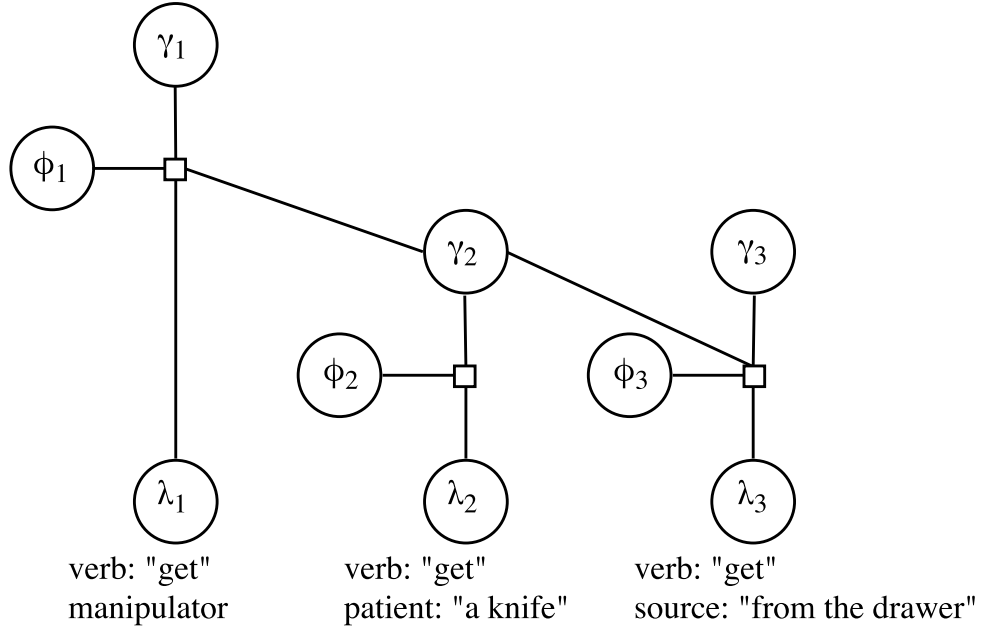


Figure 4.2: The CRF factor graph of the sentence: *the man gets a knife from the drawer.*

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.

4.3.2 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 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 4.2. 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 [147] and [162], where ϕ_1, \dots, ϕ_4 are binary random variables, indicating whether the grounding is correct. In the learning

stage, we use the following objective function:

$$p(\Phi|\lambda_1, \dots, \lambda_k, \gamma_1, \dots, \gamma_k, v) = \frac{1}{Z} \prod_i \Psi_i(\phi_i, \lambda_i, \gamma_1, \dots, \gamma_k, v) \quad (4.1)$$

where Φ is the binary random vector $[\phi_1, \dots, \phi_k]$, and v is the verb. Z is the normalization constant.

Ψ_i is the potential function that takes the following log-linear form:

$$\Psi_i(\phi_i, \lambda_i, \Gamma, v) = \exp\left(\sum_l w_l f_l(\phi_i, \lambda_i, \Gamma, v)\right) \quad (4.2)$$

where f_l is a feature function, w_l is feature weight to be learned, and $\Gamma = [\gamma_1, \dots, \gamma_k]$ 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 4.1 and Table 4.2. Note that the causality detectors shown in Table 4.2 capture relations between groundings of different semantic roles.

During parameter learning, we use gradient ascent with L2 regularization.

Compared to [147] and [162], a key difference in our model is the incorporation of causality detectors. These previous works [147, 162] 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 $\gamma_1, \dots, \gamma_k$ that maximize the probability as in Equation 4.1. To reduce the search space we apply beam search to ground in the following order: *patient, source, destination, agent*.

4.3.3 Experiments and Results

We conducted our experiments using the dataset from [162]. This dataset was developed from a subset of the TACoS corpus [117]. It contains parallel video clips and natural language descriptions.

The videos capture human performing two cooking tasks “cutting cucumber” and “cutting bread”. Each cooking task has 5 different people performing it, and all the videos were split into pairs of video clips and corresponding sentences. For each video clip, objects are annotated with bounding boxes, tracks, and labels (e.g. “cucumber”, “cutting board” etc). For each sentence, the semantic roles of a verb are extracted using PropBank [66] 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 study. 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 4.3. The results of grounding all four semantic roles are shown in Table 4.4. 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

	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 4.3: 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 4.4: Grounding accuracy on four semantic roles

the object tracks. In this case, only geometric information of tracked objects is available. Table 4.3 and Table 4.4 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 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 4.4 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 [164, 91]. We intend to apply our algorithms with automated tracking on real scenes in the future.

4.4 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

$$\arg \max_{\mathbf{c}'} p(\mathbf{c}' | \mathbf{v}), \quad (4.3)$$

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*) [10] vector. The class label indicates whether the corresponding attribute is associated with the verb.

	All	<i>take</i>	<i>put</i>	<i>get</i>	<i>cut</i>	<i>open</i>	<i>wash</i>	<i>slice</i>	<i>rinse</i>	<i>place</i>	<i>peel</i>	<i>remove</i>
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 4.5: Grounding accuracy on *patient* role using predicted causality knowledge.

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 4.5 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.

4.5 Conclusion

In this Chapter, we have applied the category-based 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.

CHAPTER 5

VISUAL CAUSALITY REASONING

5.1 Introduction

We humans rely on a vast amount of commonsense causality knowledge to understand and reason about the changing world states caused by actions. However, machines do not have such knowledge, which hinders their capability to reason, learn, and perform actions. To address this problem, we introduce a new task on naive physical action-effect prediction, which models the relations between concrete actions (expressed in the form of verb-noun pairs) and their effects on the state of the physical world as depicted by images. This task includes both *cause prediction*: given an image which describes a state of the world, identify the most likely action (in the form of a verb-noun pair, from a set of candidates) that can result in that state; and *effect prediction*: given an action in the form of a verb-noun pair, identify images (from a set of candidates) that depicts the most likely effects on the state of the world caused by that action.

We develop an approach that utilizes natural language effect descriptions as side knowledge to help acquiring web image data and bootstrap training. The empirical results have shown that, using a simple bootstrapping strategy, our approach can combine the noisy web data with a small number of seed examples to improve action-effect prediction. In addition, for a new verb-noun pair, our approach can infer its effect descriptions and predict action-effect relations only based on several image examples.

This chapter has been published in the following paper: Qiaozi Gao, Shaohua Yang, Joyce Chai, and Lucy Vanderwende. What action causes this? towards naive physical action-effect prediction. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 934-945. 2018.

5.2 Action-Effect Data Collection

First we collected a dataset to support the investigation on physical action-effect prediction. This dataset consists of actions expressed in the form of verb-noun pairs, effects of actions described in language, and effects of actions depicted in images. Note that, as we would like to have a wide range of possible effects, language data and image data are collected separately.

5.2.1 Actions (verb-noun pairs)

We selected 40 nouns that represent everyday life objects, most of them are from the COCO dataset [77], with a combination of food, kitchen ware, furniture, indoor objects, and outdoor objects. We also identified top 3000 most frequently used verbs from Google Syntactic N-gram dataset [41] (Verbargs set). And we extracted top frequent verb-noun pairs containing a verb from the top 3000 verbs and a noun in the 40 nouns which hold a *dobj* (i.e., direct object) dependency relation. This resulted in 6573 candidate verb-noun pairs. As changes to an object can occur at various dimensions (e.g., size, color, location, attachment, etc.), we manually selected a subset of verb-noun pairs based on the following criteria: (1) changes to the objects are visible (as opposed to other types such as temperature change, etc.); and (2) changes reflect one particular dimension as opposed to multiple dimensions (as entailed by high-level actions such as “cook a meal”, which correspond to multiple dimensions of change and can be further decomposed into basic actions). As a result, we created a subset of 140 verb-noun pairs (containing 62 unique verbs and 39 unique nouns) for our investigation.

5.2.2 Effects Described in Language

The basic knowledge about physical action-effect is so fundamental and shared among humans. It is often presupposed in our communication and not explicitly stated. Thus, it is difficult to extract naive action-effect relations from the existing textual data (e.g., web). This kind of knowledge is also not readily available in commonsense knowledge bases such as ConceptNet [143]. To overcome

Action	Effect Text
ignite paper	The paper is on fire.
soak shirt	The shirt is thoroughly wet.
fry potato	The potatoes become crisp and golden.
stain shirt	There is a visible mark on the shirt.

Table 5.1: Example action and effect text from our collected data.

this problem, we applied crowd-sourcing (Amazon Mechanical Turk) and collected a dataset of language descriptions describing effects for each of the 140 verb-noun pairs. The annotators were shown a verb-noun pair, and were asked to use their own words and imaginations to describe what changes might occur to the corresponding object as a result of the action. Each verb-noun pair was annotated by 10 different annotators, which has led to a total of 1400 effect descriptions. Table 5.1 shows some examples of collected effect descriptions. These effect language descriptions allow us to derive *seed effect knowledge* in a symbolic form.

5.2.3 Effects Depicted in Images

For each action, three students searched the web and collected a set of images depicting potential effects. Specifically, given a verb-noun pair, each of the three students was asked to collect at least 5 positive images and 5 negative images. Positive images are those deemed to capture the resulting world state of the action. And negative images are those deemed to capture some state of the related object (i.e., the nouns in the verb-noun pairs), but are not the resulting state of the corresponding action. Then, each student was also asked to provide positive or negative labels for the images collected by the other two students. As a result each image has three positive/negative labels. We only keep the images whose labels are agreed by all three students. In total, the dataset contains 4163 images. On average, each action has 15 positive images, and 15 negative images. Figure 5.1 shows several examples of positive images and negative images of the action *peel-orange*. The positive images show an orange in a *peeled* state, while the negative images show oranges in different states (orange as a whole, orange slices, orange juice, etc.).



Figure 5.1: Positive images (top row) and negative images (bottom row) of the action *peel-orange*.

5.3 Action-Effect Prediction

Action-effect prediction is to connect actions (as causes) to the effects of actions. Specifically, given an image which depicts a state of the world, our task is to predict what concrete actions could cause the state of the world. This task is different from traditional action recognition as the underlying actions (e.g., human body posture/movement) are not captured by the images. In this regard, it is also different from image description generation.

We frame the problem as a few-shot learning task, by only providing a few human-labelled images for each action at the training stage. Given the very limited training data, we attempt to make use of web-search images. Web search has been adopted by previous computer vision studies to acquire training data [32, 61, 11, 103]. Compared with human annotations, web-search comes at a much lower cost, but with a trade-off of poor data quality. To address this issue, we apply a bootstrapping approach that aims to handle data with noisy labels.

The first question is what search terms should be used for image search. There are two options. The first option is to directly use the action terms (i.e., verb-noun pairs) to search images and the downloaded web images are referred to as *action web images*. As desired images should

Example patterns	Extracted <i>Effect Phrases</i> (bold)
VP with a verb $\in \{\text{be, become, turn, get}\}$	The ship is destroyed .
VP + PRT	The wall is knocked off .
VP + ADVP	The door swings forward .
ADJP	The window would begin to get clean .
PP + NP	The eggs are divided into whites and yolks .

Table 5.2: Example patterns that are used to extract effect phrases (bold) from sample sentences.

be depicting effects of an action, terms describing effects become a natural choice. The second option is to use the key phrases extracted from language effect descriptions to search the web. The downloaded web images are referred to as *effect web images*.

5.3.1 Extracting Effect Phrases from Language Data

We first apply chunking (shallow parsing) using the SENNA software [20] to break an effect description into phrases such as noun phrases (NP), verb phrases (VP), prepositional phrases (PP), adjectives (ADJP), adverbs (ADVP), etc. After some examination, we found that most of the effect descriptions follow simple syntactic patterns. For a *verb-noun* pair, around 80% of its effect descriptions start with the same noun as the subject. In an effect description, the change of state associated with the noun is mainly captured by some key phrases. For example, an adjective phrase usually describes a physical state; verbs like *be, become, turn, get* often indicate a description of change of the state. Based on these observations, we defined a set of patterns to identify phrases that describe physical states of an object. In total 1997 *effect phrases* were extracted from the language data. Table 5.2 shows some example patterns and example effect phrases that are extracted.

5.3.2 Downloading Web Images

The purpose of querying search engine is to retrieve images of objects in certain effect states. To form image searching keywords, the effect phrases are concatenated with the corresponding noun phrases, for example, “apple + into thin pieces”. The image search results are downloaded and used as supplementary training data for the action-effect prediction models. However, web images can



Figure 5.2: Examples of image search results.

be noisy. First of all, not all of the automatically extracted effect phrases describe visible state of objects. Even if a phrase represents visible object states, the retrieved results may not be relevant. Figure 5.2 shows some example image search results using queries describing the object name “book”, and describing the object state such as “book is on fire”, “book is set aflame”. These state phrases were used by human annotators to describe the effect of the action “burn a book”. We can see that the images returned from the query “book is set aflame” are not depicting the physical effect state of “burn a book”. Therefore, it’s important to identify images with relevant effect states to train the model. To do that, we applied a bootstrapping method to handle the noisy web images as described in Section 5.3.3. For an action (i.e., a verb-noun pair), it has multiple corresponding effect phrases, and all of their image search results are treated as training images for this action.

Since both the human annotated image data (Section 5.2) and the web-search image data were obtained from Internet search engines, they may have duplicates. As part of the annotated images are used as test data to evaluate the models, it is important to remove duplicates. We designed a simple method to remove any images from the web-search image set that has a duplicate in the human annotated set. We first embed all images into feature vectors using pre-trained CNNs. For each web-search image, we calculate its cosine similarity score with each of the annotated images. And we simply remove the web images that have a score larger than 0.95.

5.3.3 Models

We formulate the action-effect prediction task as a multi-class classification problem. Given an image, the model will output a probability distribution \mathbf{q} over the candidate actions (i.e., verb-noun pairs) that can potentially cause the effect depicted in the image.

Specifically for model training, we are given a set of human annotated seeding image data $\{\mathbf{x}, \mathbf{t}\}$ and a set of web-search image data $\{\mathbf{x}', \mathbf{t}'\}$. Here \mathbf{x} and \mathbf{x}' are the images (depicting effect states), and \mathbf{t} and \mathbf{t}' are their classification targets (i.e., actions that cause the effects). Each target vector is the observed image label, $\mathbf{t} \in \{0, 1\}^C$, $\sum_i t_i = 1$, and C is the number of classes (i.e., actions). The human annotated targets \mathbf{t} can be trusted. But the targets of web-search images \mathbf{t}' are usually very noisy. Bootstrapping method has been shown to be an effective method to handle noisy labelled data [123, 154, 114]. The objective of the cross-entropy loss is defined as follows:

$$\mathcal{L}(\mathbf{t}, \mathbf{q}) = \sum_{i=1}^C t_i \log(q_i), \quad (5.1)$$

where \mathbf{q} are the predicted class probabilities, and C is the number of classes. To handle the noisy labels in the web-search data $\{\mathbf{x}', \mathbf{t}'\}$, we adopt a bootstrapping objective following Reed's work [114]:

$$\mathcal{L}(\mathbf{t}', \mathbf{q}) = \sum_{i=1}^C [\beta t'_i + (1 - \beta) z_i] \log(q_i), \quad (5.2)$$

where $\beta \in [0, 1]$ is a model parameter to be assigned, \mathbf{z} is the one-hot vector of the prediction \mathbf{q} , $z_i = 1$, if $i = \arg\max_k q_k, k = 1 \dots C$.

The model architecture is shown in Figure 5.3. After each training batch, the current model will be used to make predictions \mathbf{q} on images in the next batch. And the target probabilities is calculated as a linear combination of the current predictions \mathbf{q} and the observed noisy labels \mathbf{t}' . The idea behind this bootstrapping strategy is to ensure the consistency of the model's predictions. By first initializing the model on the seeding image data, the bootstrapping approach allows the model to trust more on the web images that are consistent with the seeding data.

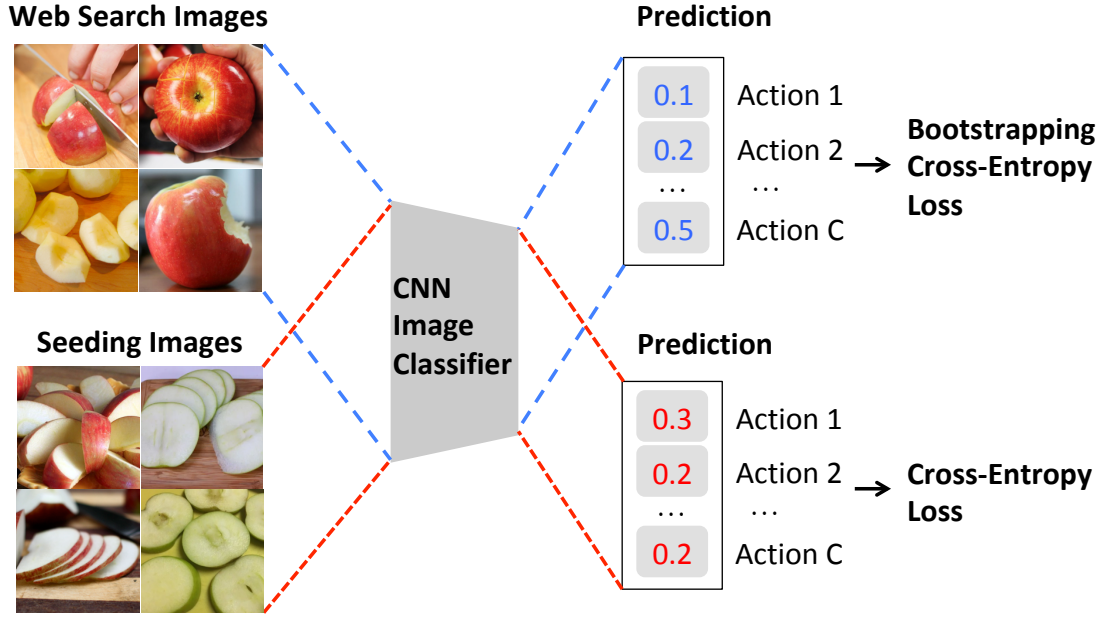


Figure 5.3: Architecture for the action-effect prediction model with bootstrapping.

5.3.4 Evaluation

We evaluate the models on the action-effect prediction task. Given an image that illustrates a state of the world, the goal is to predict what action could cause that state. Given an action in the form of a verb-noun pair, the goal is to identify images that depict the most likely effects on the state of the world caused by that action.

For each of the 140 verb-noun pairs, we use 10% of the human annotated images as the seeding image data for training, and use 30% for development and the rest 60% for test. The seeding image data set contains 408 images. On average, each verb-noun pair has less than 3 seeding images (including positive images and negative images). The development set contains 1252 images. The test set contains 2503 images. The model parameters were selected based on the performance on the development set.

As a given image may not be relevant to any effect, we add a background class to refer to images where effects are not caused by any action in the space of actions. So the total of classes for our evaluation model is 141. For each verb-noun pair and each of the effect phrases, around 40 images were downloaded from the Bing image search engine and used as candidate training examples. In

total we have 6653 action web images and 59575 effect web images.

5.3.4.1 Methods for Comparison

All the methods compared are based on one neural network structure. We use ResNet [51] pre-trained on ImageNet [22] to extract image features. The extracted image features are fed to a fully connected layer with rectified linear units and then to a softmax layer to make predictions. More specifically, we compare the following configurations:

(1) *BS+Seed+Act+Eff*. The bootstrapping approach trained on the seeding images, the action web images, and the effect web images. During the training stage, the model was first trained on the seeding image data using vanilla cross-entropy objective (Equation 5.1). Then it was further trained on a combination of the seeding image data and web-search data using the bootstrapping objective (Equation 5.2). In the experiments we set $\beta = 0.3$.

(2) *BS+Seed+Act*. The bootstrapping approach trained in the same fashion as (1). The only difference is that this method does not use the effect web images.

(3) *Seed+Act+Eff*. A baseline method trained on a combination of the seeding images, the web action images, and the web effect images, using the vanilla cross-entropy objective.

(4) *Seed+Act*. A baseline method trained on a combination of the seeding images and the action web images, using the vanilla cross-entropy objective.

(5) *Seed*. A baseline method that was only trained on the seeding image data, using the vanilla cross-entropy objective.

5.3.4.2 Evaluation Results

We apply the trained classification model to all of the test images. Based on the matrix of prediction scores, we can evaluate action-effect prediction from two angles: (1) given an action class, rank all the candidate images; (2) given an image, rank all the candidate action classes. Table 5.3 and 5.4 show the results for these two angles respectively. We report both mean average precision (MAP) and top prediction accuracy.







	Top Action Predictions	Top Effect Predictions		Top Action Predictions	Top Effect Predictions
	bite apple background cut apple peel apple	apple is eaten apple is being cut apple is chewed apple in tiny pieces		fry egg background crack egg mix eggs	egg into a harder substance cup into smaller pieces egg edible
	background chop carrot grate carrot peel carrot	carrot into tiny pieces carrot is being cut carrot into many smaller pieces		background insert key close drawer fasten door	key in the keyhole drawer without a key door is locked door is being bolted
	background cut potato fry potato mash potato	potato into a pot potato is being sliced potato for potato edible		pile books background wrap book roll paper	books in a stack book on books in a large stack books in a pile

Figure 5.4: Several example test images and their predicted actions and predicted effect descriptions. The actions in blue are ground-truth labels.

	MAP	Top 1	Top 5	Top 20
BS+Seed+Act+Eff	0.290	0.414	0.750	0.921
BS+Seed+Act	0.252	0.414	0.721	0.893
Seed+Act+Eff	0.247	0.314	0.679	0.886
Seed+Act	0.241	0.371	0.650	0.814
Seed	0.182	0.329	0.629	0.807

Table 5.3: Results for the action-effect prediction task (given an action, rank all the candidate images).

	MAP	Top 1	Top 5	Top 20
BS+Seed+Act+Eff	0.660	0.523	0.843	0.954
BS+Seed+Act	0.642	0.508	0.802	0.924
Seed+Act+Eff	0.289	0.176	0.398	0.625
Seed+Act	0.481	0.301	0.724	0.926
Seed	0.634	0.520	0.765	0.892

Table 5.4: Results for the action-effect prediction task (given an image, rank all the actions).

Overall, $BS+Seed+Act+Eff$ gives the best performance. By comparing the bootstrap approach with baseline approaches (i.e., $BS+Seed+Act+Eff$ vs. $Seed+Act+Eff$, and $BS+Seed+Act$ vs. $Seed+Act$), the bootstrapping approaches clearly outperforms their counterparts, demonstrating its ability in handling noisy web data. Comparing $BS+Seed+Act+Eff$ with $BS+Seed+Act$, we can see that $BS+Seed+Act+Eff$ performs better. This indicates the use of effect descriptions can bring more relevant images to train better models for action-effect prediction.

In Table 5.4, the poor performance of $Seed+Act+Eff$ and $Seed+Act$ shows that it is risky to fully rely on the noisy web search results. These two methods had trouble in distinguishing the background class from the rest.

We further trained another multi-class classifier with web effect images, using their corresponding effect phrases as class labels. Given a test image, we apply this new classifier to predict the effect descriptions of this image. Figure 5.4 shows some example images, their predicted actions based on our bootstrapping approach and their predicted effect phrases based on the new classifier. These examples also demonstrate another advantage of incorporating seed effect knowledge from language data: it provides state descriptions that can be used to better explain the perceived state. Such explanation can be crucial in human-agent communication for action planning and reasoning.

5.4 Generalizing Effect Knowledge to New Verb-Noun Pairs

In real applications, it is very likely that we do not have the effect knowledge (i.e., language effect descriptions) for every verb-noun pair. And annotating effect knowledge using language (as shown in Section 5.2) can be very expensive. In this section, we describe how to potentially generalize seed effect knowledge to new verb-noun pairs through an embedding model.

5.4.1 Action-Effect Embedding Model

The structure of our model is shown in Figure 5.5. This model is based on the causality embedding model in Chapter 3.2.2. It is composed of two sub-networks: one for verb-noun pairs (i.e., *action*) and the other one for effect phrases (i.e., *effect*). The *action* or *effect* is fed into an LSTM encoder

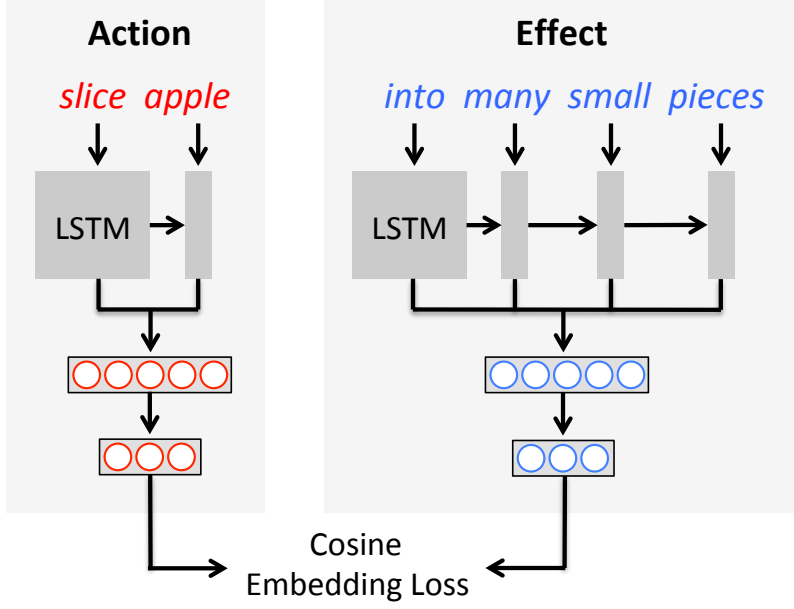


Figure 5.5: Architecture of the action-effect embedding model.

and then to two fully-connected layers. The output is an action embedding \mathbf{v}_c and effect embedding \mathbf{v}_e . The networks are trained by minimizing the following cosine embedding loss function:

$$\mathcal{L}(\mathbf{v}_c, \mathbf{v}_e) = \begin{cases} 1 - s(\mathbf{v}_c, \mathbf{v}_e), & \text{if } (c, e) \in T \\ \max(0, s(\mathbf{v}_c, \mathbf{v}_e)), & \text{if } (c, e) \notin T \end{cases}$$

$s(\cdot, \cdot)$ is the cosine similarity between vectors. T is a collection of action-effect pairs. Suppose c is an input for *action* and e is an input for *effect*, this loss function will learn an action and effect semantic space that maximizes the similarities between c and e if they have an action-effect relation (i.e., $(c, e) \in T$). During training, the negative action-effect pairs (i.e., $(c, e) \notin T$) are randomly sampled from data. In the experiments, the negative sampling ratio is set to 25. That is, for each positive action-effect pair, 25 negative pairs are created through random sampling.

At the inference step, given an unseen verb-noun pair, we embed it into the action and effect semantic space. Its embedding vector will be used to calculate similarities with all the embedding vectors of the candidate effect phrases.

	MAP	Top 1	Top 5
BS+Seed+Act+Eff	0.529	0.643	0.928
BS+Seed+Act+pEff	0.507	0.642	0.893
BS+Seed+Act	0.435	0.643	0.964
Seed	0.369	0.678	0.786

Table 5.5: Results for the action-effect prediction task (given an action, rank all the candidate images).

	MAP	Top 1	Top 5
BS+Seed+Act+Eff	0.733	0.574	0.947
BS+Seed+Act+pEff	0.729	0.551	0.961
BS+Seed+Act	0.724	0.557	0.933
Seed	0.705	0.557	0.898

Table 5.6: Results for the action-effect prediction task (given an image, rank all the actions).

5.4.2 Evaluation

We divided the 140 verb-noun pairs into 70% training set (98 verb-noun pairs), 10% development set (14) and 20% test set (28). For the action-effect embedding model, we use pre-trained GloVe word embeddings [108] as input to the LSTM. The embedding model was trained using the language effect data corresponding to the training verb-noun pairs, and then it was applied to predict effect phrases for the unseen verb-noun pairs in the test set. For each unseen verb-noun pair, we collected its top five predicted effect phrases. Each predicted effect phrase was then used as query keywords to download web effect images. This set of web images are referred to as *pEff* and will be used in training the action-effect prediction model.

For each of the 28 test (i.e., new) verb-noun pairs, we use the same ratio 10% (about 3 examples) of the human annotated images as the seeding images, which were combined with downloaded web images to train the prediction model. The remaining 30% and 60% are used as the development set, and the test set. We compare the following different configurations:

(1) *BS+Seed+Act+pEff*. The bootstrapping approach trained on the seeding images, the action web images, and the web images downloaded using the predicted effect phrases.

(2) *BS+Seed+Act+Eff*. The bootstrapping approach trained on the seeding images, the action web images, and the effect web images (downloaded using ground-truth effect phrases).

Action Text	Predicted Effect Text
chop carrot	carrot into sandwiches, carrot is sliced, carrot is cut thinly, carrot into different pieces, carrot is divided
ignite paper	paper is being charred , paper is being burned, paper is set, paper is being destroyed, paper is lit
mash potato	potato into chunks, potato into sandwiches, potato into slices, potato is chewed, potato into smaller pieces

Table 5.7: Example predicted effect phrases for new verb-noun pairs. Unseen verbs and nouns are shown in bold.

(3) *BS+Seed+Act*. The bootstrapping approach trained on the seeding images and the action web images.

(4) *Seed*. A baseline only trained on the seeding images.

Table 5.5 and 5.6 show the results for the action-effect prediction task for unseen verb-noun pairs. From the results we can see that *BS+Seed+Act+pEff* achieves close performance compared with *BS+Seed+Act+Eff*, which uses human annotated effect phrases. Although in most cases, *BS+Seed+Act+pEff* outperforms the baseline, which seems to point to the possibility that semantic embedding space can be employed to extend effect knowledge to new verb-noun pairs. However, the current results are not conclusive partly due to the small testing set. More in-depth evaluation is needed in the future.

Table 5.7 shows top predicted effect phrases for several new verb-noun pairs. After analyzing the action-effect prediction results we notice that generalizing the effect knowledge to a verb-noun pair that contains an unseen verb tends to be more difficult than generalizing to a verb-noun pair that contains an unseen noun. Among the 28 test verb-noun pairs, 12 of them contain unseen verbs and known nouns, 7 of them contain unseen nouns and known verbs. For the task of ranking images

given an action, the mean average precision is 0.447 for the unseen verb cases and 0.584 for the unseen noun cases. Although not conclusive, this might indicate that, verbs tend to capture more information about the effect states of the world than nouns.

5.5 Discussion and Conclusion

When robots operate in the physical world, they not only need to perceive the world, but also need to act to the world. They need to understand the current state, to map their goals to the world state, and to plan for actions that can lead to the goals. All of these point to the importance of the ability to understand causal relations between actions and the state of the world. To address this issue, this work introduces a new task on action-effect prediction.

Particularly, we focus on modeling the connection between an action (a verb-noun pair) and its effect as illustrated in an image and treat natural language effect descriptions as side knowledge to help acquiring web image data and bootstrap training. Our current model is very simple and performance is yet to be improved. We plan to apply more advanced approaches in the future, for example, attention models that jointly capture actions, image states, and effect descriptions. We also plan to incorporate action-effect prediction to human-robot collaboration, for example, to bridge the gap of commonsense knowledge about the physical world between humans and robots.

This chapter presents an initial investigation on action-effect prediction. There are many challenges and unknowns, from problem formulation to knowledge representation; from learning and inference algorithms to methods and metrics for evaluations. Nevertheless, we hope this work can motivate more research in this area, enabling physical action-effect reasoning, towards agents which can perceive, act, and communicate with humans in the physical world.

CHAPTER 6

UNDERSTANDING PHYSICAL ACTIONS THROUGH NATURAL LANGUAGE STORIES

6.1 Introduction

To further investigate machines’ ability in reasoning about cause-effect of physical actions as part of language understanding, we create a new language benchmark. This benchmark contains short stories created by human annotators. Each story describes a short sequence of human physical actions in our daily lives. For example, a story could describe the actions sequence of making a sandwich in the kitchen, or packing a suitcase in the bedroom. Based on the collected stories, we create two tasks to evaluate machine reading systems. The first task is to select the correct sentence from two alternatives to fill in the blank in a story, and it is called the cloze task. The second task is to select the correct order of sentences in a story, and it is called the ordering task.

Although the proposed tasks are easy for humans to solve, they are very challenging for machines. An analysis shows that understanding the stories and solving these tasks requires various types of commonsense knowledge, e.g., knowledge about action verbs, objects, and naïve physics rules. Therefore, we believe this benchmark will be a valuable resource for evaluating machines’ capability of acquiring and applying physical commonsense knowledge. Further, the setting of two sub-tasks can be naturally used to evaluate a model’s generalization ability, via training on one task and evaluating on the other task. If a model can successfully learn the fundamental knowledge and the reasoning abilities via training on the data of one sub-task, it can potentially perform well on the other sub-task. By doing this, we encourage models that focus on learning underlying knowledge instead of over-fitting to shallow statistical cues.

To tackle the commonsense reasoning tasks, we present a new neural network model. This model solves both the cloze task and the ordering task via explicitly examining the compatibility of each action with its context in those stories. Since the action-effect knowledge plays an essential role

in understanding these commonsense stories, we further incorporated physical causality knowledge into the proposed model. Experiments were designed to compare the proposed model with several state-of-the-art models for machine comprehension tasks. The results demonstrate the effectiveness of the proposed model, and further show the improvement introduced by external physical causality knowledge. The results also suggests that this benchmark is challenging for current approaches, and better solving this task requires a wider range of commonsense knowledge and richer semantic representation of actions and objects.

6.2 Physical Commonsense Reasoning Tasks

The proposed benchmark includes two subtasks. The **cloze task** is to select the correct sentence to fill in the blank in a story. The **ordering task** is to select the correct order of sentences in a story. In both tasks, each story describes a short sequence of human physical actions in our daily lives. Examples are shown in Figure 6.1.

Cloze Task:	Ordering Task:
<p><i>Select a sentence to fill the blank in the story.</i></p> <p>Story:</p> <ol style="list-style-type: none"> 1. John took a donut out of the cabinet. 2. John ate the donut. 3. John turned on the stove. 4. John melted butter in a pan. 5. _____ <p>Candidates:</p> <p>A. John stirred the butter in the pan.</p> <p>B. John poured melted butter on his donut.</p>	<p><i>Select the correct order of sentences.</i></p> <p>Story sentences:</p> <ol style="list-style-type: none"> 1. Ann put some yogurt into the blender. 2. Ann cut the banana into pieces and put it in the blender. 3. Ann washed and peeled a banana. 4. Ann poured a cup of water into the blender. 5. Ann blended the smoothie. <p>Candidates:</p> <p>A. 12345 B. 13245</p>

Figure 6.1: Example story data for the cloze task and the ordering task. Candidates in red are correct answers.

For the **cloze task**, one sentence in an original story is replaced with a blank. That sentence is then put together with a distraction sentence to form the candidates set. Given the story with a blank, a system needs to select the correct sentence from the candidates to fill in the blank. The distraction sentences are created in such a way that they describe very common human actions in

the corresponding environment, but adding them to the story will make the story irrational in the physical world.

For the **ordering task**, two sentences in an original story are chosen and their positions are switched. These sentences are selected in a manner that if we switch their positions, the story becomes irrational in the physical world. Given the original story and the reordered story, a system needs to determine which story makes more sense.

In our data, the cloze task and the ordering task are different in their setups, but they are also closely related, since both of them rely on the knowledge of action prerequisites and effects, and the capability of tracking the state changes introduced by human actions. The design of including two different but closely related subtasks is motivated by the recent criticisms on data biases introduced to natural language benchmarks during data collection [128, 49]. For example, Schwartz et al. [128] have shown that the Story Cloze Test [95] (which has a similar setting with our cloze task) can be solved with up to 75% accuracy by only exploiting stylistic features, even without looking at the story context. With two parallel tasks in this benchmark, we can use them to evaluate a model’s generalization ability, via training on one task and evaluating on the other task. If a model can successfully acquire the underlying commonsense knowledge and learn the reasoning abilities via training on the data of one task, it is very likely to perform well on the other task. By doing this, we encourage models that focus on learning underlying knowledge instead of overfitting to shallow language patterns.

6.2.1 Data Collection through Crowdsourcing

We collected a set of human-written stories via Amazon Mechanical Turk. Each story describes a sequence of physical actions in human daily lives. During data collection, the annotators were shown a person’s name and a location name, and they were asked to use their imagination to write a short story describing a sequence of physical actions the person takes in that location. Possible locations includes *kitchen, living room, bathroom, garage, bathroom, office, park*. Several requirements were given to the annotators: 1) All described actions should be entirely realistic; 2)

Story-03:

- 1 Tom opened the fridge and grabbed a pre-made breakfast sandwich.
- 2 Tom put the sandwich on a microwavable pan, put it in the microwave oven and turned it on.
- 3 Tom poured himself a cup of coffee and opened the fridge again.
- 4 Tom took out the watermelon he had cut up the night before along with a banana.
- 5 Tom put the fruit in his lunch bag and ate his breakfast sandwich.

Replace an original sentence and make the story NOT possible to happen in the physical world.

Sentence number (simply write down the digit):

Your new sentence:

Figure 6.2: Interface used for annotating stories for the cloze task.

Story-052:

- 1 Ann went into the bathroom and turned on the shower faucet.
- 2 Ann washed her hair in the shower, using the last of the shampoo from the shampoo bottle.
- 3 Ann got out of the shower and used her hair dryer.
- 4 Ann threw her sheet and blanket in the washer but found she was out of laundry detergent.
- 5 Ann got the shampoo bottle and poured a cup of shampoo into the washer.

Write down the indexes of two sentences that if we switch their positions, the story does not make sense.

Figure 6.3: Interface used for annotating stories for the ordering task.

The actions should be carried out in a short time period; 3) The story must include at least five sentences.

After collecting the original stories, we asked a different group of annotators to read the stories and prepare them for the cloze task and the ordering task. Specifically, to prepare data for the cloze task, we asked annotators to write a new sentence to replace an original sentence in the story, such that the story after replacement is not likely to happen in the physical world. The annotation

interface is shown in Figure 6.2. The new sentences will be used as distraction alternatives in the cloze task. To make this task more challenging, we asked the annotator to come up with sentences that are entirely realistic in real life. For example, sentences like “Mary fried eggs on the printer”, or “Tom ate the spoon”, are not acceptable, since they are not realistic. In this way, one can not determine which is the correct sentence to fill the blank by only looking at the candidate sentences. You always need to put the sentences back into the story context to determine.

To prepare data for the ordering task, we asked annotators to switch two sentences in the original story, so that the story after switching is not likely to happen in the physical world. The annotation interface is shown in Figure 6.3. After data collection, we also filtered out words like “the”, “a”, “an”, just to get rid of some trivial cues for the correct order between some sentences. Since “the” is usually used to refer something mentioned before, while “a” and “an” are usually used to refer something not mentioned before. For example, a system can easily determine the order of the following two sentences, “Tom got an apple out of the fridge” and “Tom peeled the apple with a knife”, only by looking at the usage of “an apple” and “the apple”.

In total, we have collected 727 human-written stories. And based on these original stories, we created 1,672 instances for the cloze task and 4,577 instances for the ordering task.

6.2.2 Underlying Commonsense Knowledge

After data collection, we analyzed the task data and discovered several categories of commonsense knowledge that are essential to solve the tasks. Here we list the knowledge categories and also show task examples that require them to make prediction.

1. Verb Causality Knowledge describes how a physical action changes the involving objects’ physical states. For example, the key point of solving the following cloze problem is knowing that the action *bake the potato* causes the potato to become hot.

- Story:

1. Tom preheated the oven.

2. Tom took out a potato from the fridge.

3. Tom put the potato in a metal pan.

4. Tom baked the potato in the oven.

5. _____.

- Select the correct sentence to fill in the blank:

A. Tom sprinkled some grated cheese on the potato.

B. Tom ate the cold potato.

- Correct Answer: A

2. Action Precondition is the requirement that must be satisfied before an action happens. For example, you can *cut* a solid object instead of liquid, or you can *stir* liquid instead of a solid object. To solve the following cloze problem, one needs to know that the butter is in liquid form after *melting* (this information belongs to verb causality knowledge), and you cannot *cut* liquid (action precondition knowledge).

- Story:

1. Tom took the potato out of the oven.

2. Tom mashed the potato.

3. Tom melted butter in the microwave.

4. _____.

5. Tom ate the mashed potato with a spoon.

- Select the correct sentence to fill in the blank:

A. Tom put the mashed potato and butter in a bowl.

B. Tom cut the butter into cubes.

- Correct Answer: A

3. Object Functionality involves information about specific functions of objects, especially for tools. For example, a *microwave oven* can be used to heat objects, and a *wrench* can be used to repair cars. In the following ordering task, it is critical to infer that the *wrench* was used to *tighten the bolt*.

- Story:

1. John opened the toolbox.
2. John took out the wrench.
3. John tightened a bolt on his bicycle.
4. John put the wrench back in the box.
5. John rode the bicycle to the store.

- Select the correct order:

A. 12345

B. 13245

- Correct Answer: A

4. Intuitive Physics is human's common understanding about basic physical phenomena. For example, a solid object can not pass through another solid object; an existing object continues to exist unless being moved away or destroyed; an object is in a container, if the container moves, the object also moves along. Psychological studies have shown that human naive physics rules usually develop at a very young age, even before the development of language ability [6]. Thus, we rarely explicitly express this kind of information in our communication. We simply assume everyone

knows that. If an AI system ever deeply understand human natural language, it needs to acquire this kind of knowledge. For the following cloze problem, the key is to infer that the hammer is not accessible since it is locked in the trunk.

- Story:

1. John took off a bucket from the shelf.
2. John picked up a hammer and a rope from the floor.
3. John put the hammer and rope into the bucket.
4. John locked the bucket in his car trunk.
5. _____.

- Select the correct sentence to fill in the blank:

- A. John used the hammer to repair the bike.
- B. John drove his car into the street.

- Correct Answer: B

6.2.3 Comparison with Existing Tasks

The proposed tasks are similar to existing machine comprehension tasks, for example, bAbI [153], SQuAD [113], and the Story Cloze Test [95]. Since in all of these tasks, a model needs to make predictions based on its understanding of the provided supporting text data. However the proposed tasks are also different from them. In bAbI [153] and SQuAD [113], the evaluation is done in a question answering setting, where the input includes a supporting document together with a question, and the model needs to select words from the supporting document as answer. In the Story Cloze Test [95], the input includes a short story and two alternative endings to the story, and the model needs to predict which ending is the correct one. For our proposed cloze task, it is very

similar to the Story Cloze Test task, except that the blanks are not always at the end of the stories. In the Story Cloze Test, most of the stories focus on human’s emotions, intentions and attitudes (i.e., naive psychology), while our tasks have a specific focus on human’s physical actions.

6.3 Methods

For the cloze task, a model needs to make a selection between two candidate sentence to fill in the blank. For the ordering task, a model needs to make a selection between two sequences of sentences. In order to unify these the two proposed tasks, we treat them as a story ranking task: given two candidate stories, predicting which one is more rational. For the cloze task, we get two candidate stories via replacing the blank with each candidate sentence. For the ordering task, we can also get two candidate stories via treating each sequence as a candidate story.

To tackle the proposed commonsense reasoning tasks, we propose a neural network model that explicitly examines each action in terms of its compatibility with the actions happening before it and actions happening after it. This model is motivated by the fact that the order of sentences are very important in understanding the narrative.

As discussed earlier, solving these tasks requires a lot of commonsense knowledge. Given the fact that commonsense knowledge is not usually explicitly stated in natural language, and also given the limited size of our data set, it is not practical to acquire all the commonsense knowledge from training data. So in this work, we also explore methods that can leverage external commonsense knowledge for better understanding and reasoning about the stories.

6.3.1 The Attentive-Reader Model

The architecture of the Attentive-Reader is shown in Figure 6.4. In this model, we adopt sentence-level representations. We first use a bi-directional gated recurrent unit (Bi-GRU) to embed the sentences into vectors. Then for each sentence, we examine its compatibility with the rest parts of the story. Taking sentence 3 as the target sentence, we use its embedding vector e_3 to attend every sentence before it (e_1 and e_2) and every sentence after it (e_4 and e_5), separately. For sentence

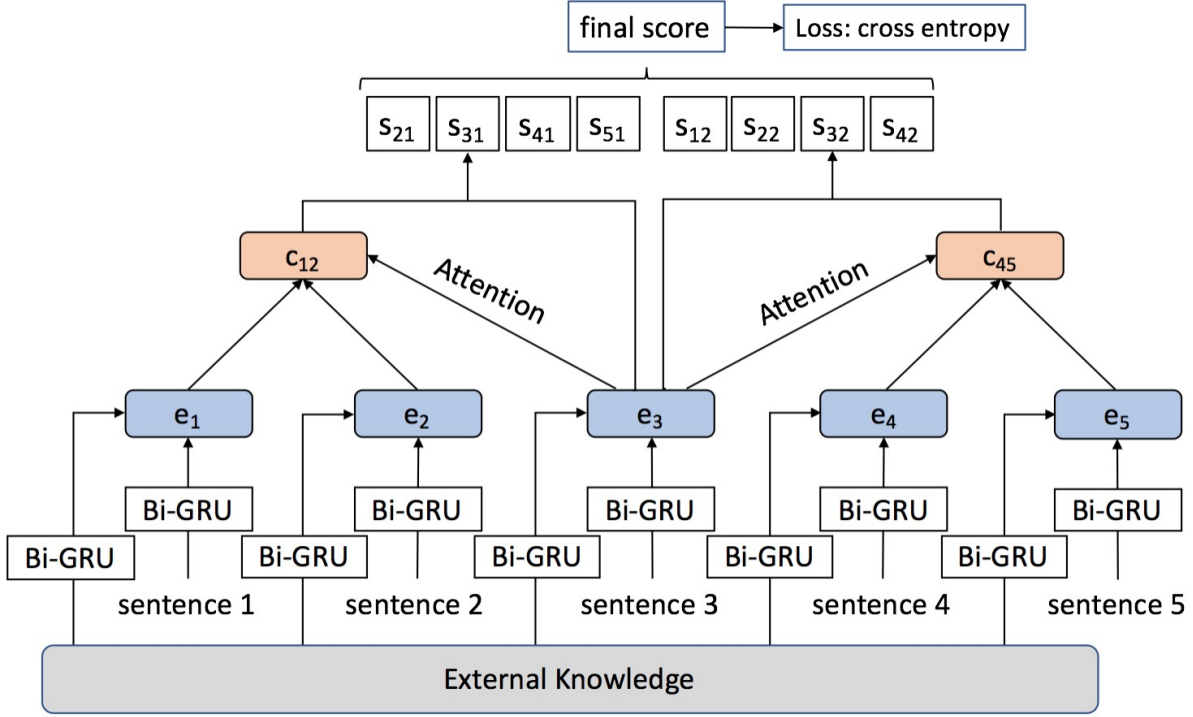


Figure 6.4: Network architecture for the Attentive-Reader. Note that this architecture only shows the computation structure for the anomaly scores corresponding to sentence 3 (score s_{31} and score s_{32}). The anomaly scores for other sentences are computed via similar processes.

before it, we calculate the attentions with

$$\alpha_i = \frac{\exp(e_3 W_a e_i)}{\sum_{i < 3} \exp(e_3 W_a e_i)}, \quad (6.1)$$

where W_a is the parameter matrix to be learned. Then we represent the before-context with a weighted sum

$$c_{12} = \sum_{i < 3} \alpha_i e_i. \quad (6.2)$$

Then the context representation is used to calculate the anomaly score s_{31} between the target sentence and the context before it.

$$s_{31} = \tanh(W_1 [c_{12} : e_3] + b_1) \quad (6.3)$$

Here W_1 and b_1 are parameters to be learned, and “:” denotes concatenation. The after-context representation c_{45} and the anomaly score s_{32} are computed in a similar way. After calculating the

anomaly scores for every target sentence, we apply a max/mean-pooling on them and generate the final score. We use a cross entropy loss at the final layer.

6.3.1.1 Leveraging Physical Causality Knowledge

Since the action-effect knowledge plays an essential role in understanding these commonsense stories, we further incorporated physical causality knowledge into the proposed model. To inject the verb causality knowledge, we introduce an external knowledge module in the Attentive-Reader architecture. This module takes external knowledge in the form of natural language sentences. As shown in Figure 6.4, we use the same bi-directional gated recurrent unit (Bi-GRU) to embed the knowledge into vector representations. Later they will be concatenated with story sentence vectors to form knowledge-aware sentence representations.

6.3.1.2 Typed Physical Causality Knowledge

To form the external knowledge base, we start with the category-based knowledge data in Chapter 4.3.1. Given a verb, this knowledge data only tells us which state categories are likely to changes, but can not tell us how will they change. Theoretical linguistic studies on verbs have shown that result verbs often specify movement along a scale [60]. Inspired by this, we introduce state changing directions (or types) to the categories. Specifically, we selected a subset (*presence, integrity, location, containment, temperature, wetness*) of the 18 attributes from Chapter 3.1, and added types to them (shown in Table 6.1). Then, we manually annotated the 329 transitive verbs from the current story dataset based on the typed causality attributes. When applying the knowledge to a story sentence, we first run dependency parsing on this sentence to find out the *verb*, *direct object* and *location* (if exists). After extracting the typed attribute values for this *verb*, we replace the terms *object* and *location* with the corresponding terms in the sentence. For example, the external knowledge for “Tom put the potato in the fridge” will be “*potato* be in *fridge*”.

State Attributes	Typed attribute values
Presence	object be present; object be not present
Location	object be in location; object be out of location
Integrity	object be broken; object be integral
Containment	object be full; object be empty
Wetness	object be wet; object be dry
Temperature	object be cold; object be hot

Table 6.1: Typed state attributes for physical causality knowledge.

6.3.2 Models for Comparison

The EntNet-Reader Model is based on the Recurrent Entity Network (EntNet) [52, 80], a neural framework with external memory chains. Neural models with long-term memory and attention mechanism have exhibited good reasoning capabilities in machine comprehension tasks [145, 45, 52, 80]. And particularly, the EntNet model has been proven to be very effective on similar reasoning tasks like bAbI [52], CBT (Children’s Book Test) [53, 52] and the Story Cloze Test [80].

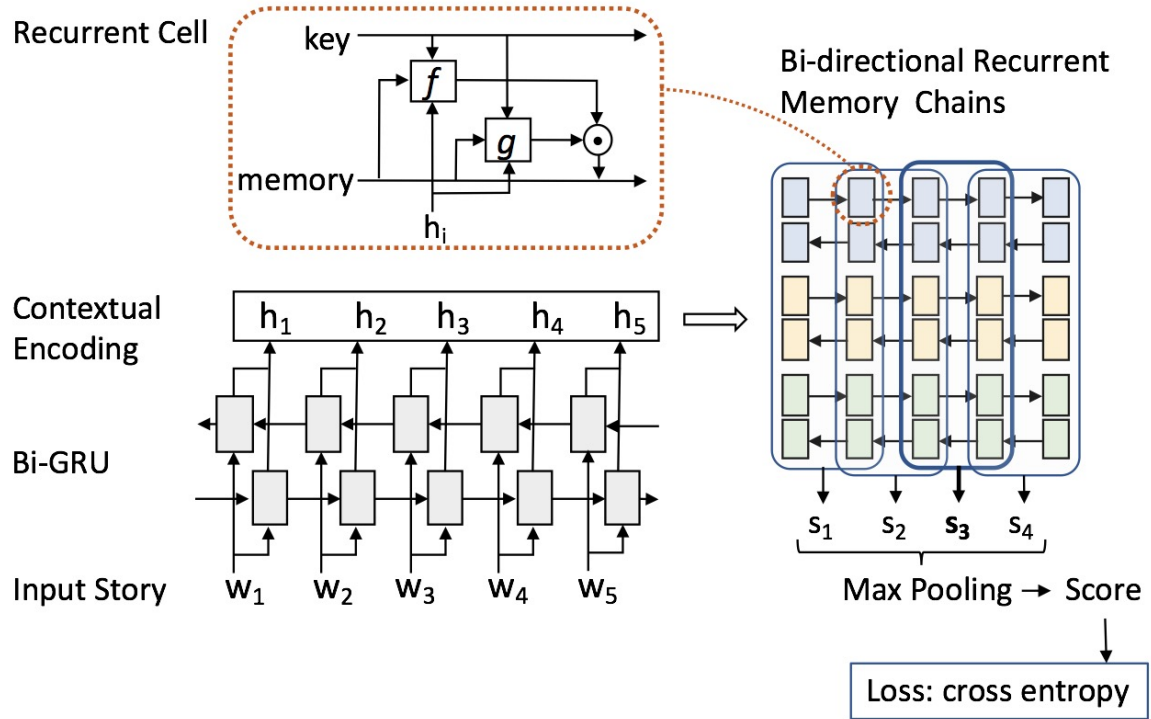


Figure 6.5: Network architecture for the EntNet-based approach.

The architecture of the EntNet-Reader is shown in Figure 6.5. First, a bi-directional gated recurrent unit (Bi-GRU) is used to embed the context information of the story at word level. Then the context-dependent word representations are taken as input to a bi-directional Recurrent Entity Network (EntNet) [52, 80], where the model tracks the state of world with memory chains. Each memory chain is a special RNN network, where there is a key governing what kind of information can pass the gates and be stored in the memory. The state representations of all memory chains are then gathered into a 2D array and a convolution filter is applied on top of it. Specifically, the filter covers the memory states from two adjacent time points, and outputs an anomaly score. This score basically tells us given the world state of the previous time point, how irrational is the next state/action. Lastly, a max/mean pooling layer will output the final anomaly score. We adopt a cross entropy loss on the final score.

The Bi-GRU Baseline. For comparison, we also introduce a baseline model that uses a Bi-GRU to embed the whole story into one single vector representation and then generates the final prediction score with MLP (multilayer perceptron). Again, we use the cross entropy as loss function.

6.4 Experiments

6.4.1 Experimental Settings

For both the cloze task and the ordering task, we randomly divided the data for training (20%), validation (20%) and test (60%). The task instances derived from one original story all appear in the same set (either in training set, validation set or test set). This data split strategy is to prevent a trivial solution that memorizes positive action sequences from the training data.

For all the models, pre-trained GloVE embeddings [108] with 300 dimension are used as input word embeddings. The hidden size for Bi-GRU and EntNet are set to 300. Training is carried out with the Adam optimizer [65] and a batch size of 32.

Using 100% of the training data				
	Bi-GRU	EntNet	Attentive	Attentive+KB
Cloze Task	0.634	0.668	0.681	0.701
Ordering Task	0.662	0.648	0.682	0.687
Using 67% of the training data				
	Bi-GRU	EntNet	Attentive	Attentive+KB
Cloze Task	0.622	0.647	0.660	0.688
Ordering Task	0.644	0.623	0.653	0.684
Using 33% of the training data				
	Bi-GRU	EntNet	Attentive	Attentive+KB
Cloze Task	0.565	0.585	0.597	0.630
Ordering Task	0.619	0.619	0.628	0.656

Table 6.2: Prediction accuracy results on the physical commonsense reasoning tasks.

6.4.2 Results and Analysis

Table 6.2 shows the evaluation results for different models on the cloze task and the ordering task. We vary the training size to evaluate the models’ performance with different training sizes. Here EntNet refers to the EntNet-Reader, Attentive refers to the Attentive-Reader, and KB denotes the use of external physical causality knowledge.

Overall the Attentive-Reader performs better than the EntNet-Reader and the Bi-GRU model. This might suggest that the sentence-level representation works better on the proposed tasks. After introducing external causality knowledge, the Attentive-Reader achieves the best performance on both tasks. This indicates the effectiveness of the external knowledge module in Figure 6.4, together with the typed physical causality knowledge base.

Given that a random guessing method could achieve 0.5 accuracy, these results also suggest that this benchmark is very challenging for current approaches. More advanced approaches are required, with better coverage of commonsense knowledge and richer semantic representation of actions and world states.

As mentioned earlier, although the cloze task and the ordering task are different in their task setups, they require very similar reasoning processes to solve. The key to tackle both tasks includes reliably tracking object state changes, and successfully detecting anomaly about action

	Bi-GRU	EntNet	Attentive	Attentive+KB
Cloze - Ordering	0.535	0.550	0.545	0.579
Ordering - Cloze	0.543	0.561	0.560	0.607

Table 6.3: Prediction accuracy results of training on one task and evaluating on the other task.

preconditions and effects. To evaluate how different models generalize to an unseen task, we carried out a new set of experiments of training these models on one task and testing them on the other task.

The results of evaluating on a new task are shown in Table 6.3. Again the best performing model is the Attentive-Reader with external causality knowledge. This suggests that the typed physical causality knowledge helps the model better tracking object state changes instead of overfitting to shallow language patterns.

6.4.3 Predicting Breakpoints in Negative Stories

For both the EntNet-Reader model and the Attentive-Reader model, their designed network structures enable them to make predictions about which part of stories does not make sense. For example, each of the intermediate anomaly scores $s_{21}, s_{31}, \dots, s_{42}$ indicates the compatibility of the corresponding sentence with the sentences before it or after it. Therefore, a sentence with the maximum anomaly score basically suggests that this sentence is most likely to be conflicting with its context, according to the trained model.

Note that for each negative story (selecting the wrong candidate sentence in the cloze task, or selecting the wrong order in the ordering task), there are at least two sentences conflicts with each other. A manual analysis of the results (models trained using 100% training data) shows that, both the Att-Reader and the EntNet-Reader have a good chance (around 80%) to successfully find out at least one of the conflicting sentences for negative stories. The following are several examples of the Attentive-Reader’s breakpoint predictions on negative stories.

Negative story 1 (cloze):

1. Mary took pan from cupboard.

2. Mary put pan on stove.
3. Mary took out bowl.
4. Mary cracked open egg.
5. Mary made hardboiled egg.

Conflicting sentences: 4 and 5

Model's prediction: 5

Negative story 2 (cloze):

1. John locked window.
2. John unplugged tv.
3. John turned off fan.
4. John closed up his suitcase.
5. John stayed in and watched tv.

Conflicting sentences: 2 and 5

Model's prediction: 2

Negative story 3 (ordering):

1. John opened freezer and took out ice cream.
2. John scooped out some ice cream and put it in blender.
3. John cleaned up his mess with broom.
4. John knocked over bowl of butter.
5. John put fruit in blender and made some shake.

Conflicting sentences: 3 and 4

Model's prediction: 3

6.5 Summary

In this chapter we propose a new benchmark for physical commonsense reasoning. To the best of our knowledge, this is the first crowdsourced natural language story dataset specifically targeted

for evaluating machines’ capability of understanding and reasoning about human physical actions. This benchmark contains two sub-tasks: a cloze task and an ordering task. The setting of two sub-tasks in this benchmark can be naturally used to evaluate a model’s generalization ability, via training on one task and evaluating on the other task. We believe this benchmark will serve as an valuable resource for physical commonsense reasoning.

As the first attempt to tackle the proposed tasks, we present a neural network architecture together with an external knowledge module. This model solves both the cloze task and the ordering task via explicitly examining the compatibility of each action with its context in those stories. The experimental results demonstrate the best performing setup is the proposed model with typed verb causality annotation as external knowledge. The relative low performance of all the tested models suggests that the proposed tasks are far from being solved by current approaches. A careful analysis of the task data suggests that future investigates could focus on modeling a wider range of commonsense knowledge and providing richer semantic representation of actions and objects.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation presents a series of investigation on collecting, modeling and utilizing physical causality knowledge of action verbs. First, physical causality knowledge were collected from human contributors through crowdsourcing. Two representation methods were adopted to model physical causality knowledge, one is based on pre-defined categories, and the other one is based on natural language embedding models. Both approaches have demonstrated their potential on modeling verb semantics and connecting language to the physical world. We further incorporated causality modeling in solving several challenging tasks: language grounding, visual causality reasoning, and commonsense story understanding.

In Chapter 4, we applied the category-based causality modeling to the task of grounding semantic roles from sentences to visual perceptions using two approaches: a knowledge-based approach and a learning-based approach. The empirical evaluations have demonstrated that both of the proposed approaches outperform previous work, indicating that causality categorization provides a good guideline for designing intermediate visual features. Moreover, we have shown that physical causality knowledge can be generalized to novel verbs using simple learned models.

In Chapter 5, we introduced a novel task of visual causality reasoning, which focuses on the connection between an action (a verb-noun pair) and its effect as illustrated in an image. We have developed an approach that applies distant supervision to harness web data for bootstrapping action-effect prediction models. The empirical results have shown that, using a simple bootstrapping strategy, our approach can combine the noisy web data with a small number of seed examples to improve action-effect prediction. Furthermore, our approach can infer effect descriptions for new verb-noun pairs and thus to facilitate the training of action-effect prediction. This opens up the possibility for humans to teach robots new tasks through language communication and small number of examples.

In Chapter 6, we introduced a new benchmark for physical commonsense reasoning, which

contains two sub-tasks, a cloze task and an ordering task. This benchmark evaluates a system’s capability of understanding and reasoning about human physical actions from story data. We presented a novel neural network model that explicitly examines the compatibility of each sentence with its context. Experimental results have demonstrated the effectiveness of the proposed model, and further show the improvement introduced by incorporating external physical causality knowledge.

Apart from the studies shown in this dissertation, there are many interesting and promising directions left for future exploration. Here we mention several:

1. Building a general purpose knowledge base for physical causality of action verbs. Our studies have shown that the proposed physical causality knowledge datasets are good supplements to current verb meaning models and resources. Thus, a natural extension is to build a large-scale physical causality knowledge base for open-domain tasks.
2. Connecting verb causality knowledge with other types of intuitive physics knowledge. As mentioned in Chapter 6.2, there are different types of commonsense knowledge closely related to the understanding and reasoning about physical actions. Therefore, one interesting research topic is to explore how to acquire these different types of knowledge and model them with a unified framework.
3. Exploring methods that better ground language to perceptions. In Chapter 5 we present an initial investigation on connecting language with action effect images. There are many challenges and unknowns in grounding language to more complex forms of perceptions, from video data to live human-robot interactions.
4. Extending verb causality knowledge to metaphorical uses. The studies we have done are only focused on literal uses of verbs, i.e., mainly about concrete actions applied on concrete objects. However, metaphorical uses of verb are very common in human natural language. For example, “Two planes were shot down” is a literal usage of the verb “shot”, while “The proposals were shot down” is a metaphorical usage. Given the value of verb causality

knowledge on literal verb uses, one would anticipate that extending verb causality knowledge to metaphorical uses will help comprehending and reasoning about natural language in more general situations.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826, 2013.
- [2] Jacob Andreas and Dan Klein. Grounding language with points and paths in continuous spaces. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 58–67, 2014.
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.
- [4] Yoav Artzi and Luke Zettlemoyer. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62, 2013.
- [5] David Bailey, Nancy Chang, Jerome Feldman, and Srinu Narayanan. Extending embodied lexical development. In *Proceedings of the Twentieth Conference of the Cognitive Science Society*, pages 84–89, 1998.
- [6] Renée Baillargeon. Infants’ physical world. *Current directions in psychological science*, 13(3):89–94, 2004.
- [7] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.
- [8] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *Journal of machine learning research*, 3(Feb):1107–1135, 2003.
- [9] Kobus Barnard and David Forsyth. Learning the semantics of words and pictures. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 408–415. IEEE, 2001.
- [10] Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.
- [11] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010.
- [12] Eduardo Blanco, Nuria Castell, and Dan I Moldovan. Causal relation extraction. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC’08)*, 2008.

- [13] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM, 2008.
- [14] Benjamin Börschinger, Bevan K Jones, and Mark Johnson. Reducing grounded learning tasks to grammatical inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1416–1425. Association for Computational Linguistics, 2011.
- [15] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.
- [16] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics, 2012.
- [17] Elia Bruni, Giang Binh Tran, and Marco Baroni. Distributional semantics from text and images. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 22–32. Association for Computational Linguistics, 2011.
- [18] Craig G Chambers, Michael K Tanenhaus, and James S Magnuson. Actions and affordances in syntactic ambiguity resolution. *Journal of experimental psychology: Learning, memory, and cognition*, 30(3):687, 2004.
- [19] Stephen V Cole, Matthew D Royal, Marco G Valtorta, Michael N Huhns, and John B Bowles. A lightweight tool for automatically extracting causal relationships from text. In *SoutheastCon, 2006. Proceedings of the IEEE*, pages 125–129. IEEE, 2005.
- [20] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.
- [21] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer, 2005.
- [22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
- [23] Robert MW Dixon and Alexandra Y Aikhenvald. *Adjective Classes: A Cross-linguistic Typology*. Explorations in Language and Space C. Oxford University Press, 2006.
- [24] Quang Xuan Do, Yee Seng Chan, and Dan Roth. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics, 2011.

- [25] Malcolm Doering. *Verb semantics as denoting change of state in the physical world*. Michigan State University, 2015.
- [26] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.
- [27] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 277–285. Association for Computational Linguistics, 2010.
- [28] Curt J Ducasse. On the nature and the observability of the causal relation. *The Journal of Philosophy*, 23(3):57–68, 1926.
- [29] Bastianelli Emanuele, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. Textual inference and meaning representation in human robot interaction. In *Joint Symposium on Semantic Processing.*, page 65, 2013.
- [30] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785. IEEE, 2009.
- [31] Alireza Fathi and James M Rehg. Modeling actions through state changes. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2579–2586. IEEE, 2013.
- [32] Robert Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman. Learning object categories from google’s image search. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1816–1823. IEEE, 2005.
- [33] Richard E. Fikes and Nils J. Nilsson. 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, 1971. Morgan Kaufmann Publishers Inc.
- [34] Amy Fire and Song-Chun Zhu. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):23, 2015.
- [35] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.
- [36] Qiaozi Gao, Malcolm Doering, Shaohua Yang, and Joyce Y Chai. Physical causality of action verbs in grounded language understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1814–1824, 2016.

- [37] Qiaozi Gao, Shaohua Yang, Joyce Chai, and Lucy Vanderwende. What action causes this? towards naive physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 934–945, 2018.
- [38] Peter Gardenfors. Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, 2(2):9–27, 2004.
- [39] Malik Ghallab, Dana Nau, and Paolo Traverso. *Automated planning: theory & practice*. Elsevier, 2004.
- [40] Roxana Girju, Dan I Moldovan, et al. Text mining for causal relations. In *FLAIRS Conference*, pages 360–364, 2002.
- [41] Yoav Goldberg and Jon Orwant. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 241–247, 2013.
- [42] Eugenia Goldvarg and Philip N Johnson-Laird. Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive science*, 25(4):565–610, 2001.
- [43] Dave Golland, Percy Liang, and Dan Klein. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 410–419. Association for Computational Linguistics, 2010.
- [44] Alison Gopnik, Laura Schulz, and Laura Elizabeth Schulz. *Causal learning: Psychology, philosophy, and computation*. Oxford University Press, 2007.
- [45] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471, 2016.
- [46] Clayton Greenberg, Asad Sayeed, and Vera Demberg. Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics–Human Language Technologies, Denver, USA*, 2015.
- [47] Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Go, Yangqing Jia, Dan Klein, Pieter Abbeel, Trevor Darrell, et al. Grounding spatial relations for human-robot interaction. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1640–1647. IEEE, 2013.
- [48] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.
- [49] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

- [50] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.
- [51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [52] Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. Tracking the world state with recurrent entity networks. *arXiv preprint arXiv:1612.03969*, 2016.
- [53] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children’s books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015.
- [54] Malka Rappaport Hovav and Beth Levin. Reflections on Manner / Result Complementarity. *Lexical Semantics, Syntax, and Event Structure*, pages 21–38, 2010.
- [55] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016.
- [56] Gerhard Jäger. Natural color categories are convex sets. In *Logic, language and meaning*, pages 11–20. Springer, 2010.
- [57] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In *Advances in neural information processing systems*, pages 3464–3472, 2014.
- [58] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.
- [59] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.
- [60] Christopher Kennedy and Louise McNally. Scale structure and the semantic typology of gradable predicates. *Language*, 81(2)(0094263):345–381, 2005.
- [61] Lyndon S Kennedy, Shih-Fu Chang, and Igor V Kozintsev. To search or to label?: predicting the performance of search-based automatic image classifiers. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 249–258. ACM, 2006.
- [62] Casey Kennington, Spyros Kousidis, and David Schlangen. Situated incremental natural language understanding using a multimodal, linguistically-driven update model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1803–1812, 2014.
- [63] Casey Kennington and David Schlangen. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. 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)*, volume 1, pages 292–301, 2015.
- [64] Joohyun Kim and Raymond J Mooney. Unsupervised pcfg induction for grounded language learning with highly ambiguous supervision. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 433–444. Association for Computational Linguistics, 2012.
 - [65] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
 - [66] Paul Kingsbury and Martha Palmer. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, 2002.
 - [67] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.
 - [68] Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 259–266. IEEE Press, 2010.
 - [69] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.
 - [70] Jayant Krishnamurthy and Thomas Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206, 2013.
 - [71] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.
 - [72] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.
 - [73] Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.
 - [74] Beth Levin and Malka Rappaport Hovav. Lexicalized scales and verbs of scalar change. In *46th Annual Meeting of the Chicago Linguistics Society*, 2010.
 - [75] Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. Learning to rank for plausible plausibility. *arXiv preprint arXiv:1906.02079*, 2019.

- [76] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2657–2664, 2014.
- [77] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [78] Changsong Liu and Joyce Y. Chai. Learning to mediate perceptual differences in situated human-robot dialogue. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI’15)*, pages 2288–2294, Austin, TX, 2015.
- [79] Changsong Liu, Lanbo She, Rui Fang, and Joyce Y. Chai. 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, 2014.
- [80] Fei Liu, Trevor Cohn, and Timothy Baldwin. Narrative modeling with memory chains and semantic supervision. *arXiv preprint arXiv:1805.06122*, 2018.
- [81] Hugo Liu and Push Singh. Conceptnet - a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.
- [82] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.
- [83] Wei Lu, Hwee Tou Ng, Wee Sun Lee, and Luke S Zettlemoyer. A generative model for parsing natural language to meaning representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 783–792. Association for Computational Linguistics, 2008.
- [84] Zhiyi Luo, Yuchen Sha, Kenny Q Zhu, Seung-won Hwang, and Zhongyuan Wang. Commonsense causal reasoning between short texts. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2016.
- [85] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- [86] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning. *arXiv preprint arXiv:1206.6423*, 2012.
- [87] Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. Learning to parse natural language commands to a robot control system. In *Experimental Robotics*, pages 403–415. Springer, 2013.

- [88] Brian McMahan and Matthew Stone. A bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115, 2015.
- [89] Ken McRae, Michael J Spivey-Knowlton, and Michael K Tanenhaus. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312, 1998.
- [90] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [91] Anton Milan, Stefan Roth, and Kaspar Schindler. Continuous energy minimization for multitarget tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(1):58–72, 2014.
- [92] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [93] Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. *Robotics: Science and Systems (RSS)*, 2014.
- [94] Dipendra Kumar Misra, Kejia Tao, Percy Liang, and Ashutosh Saxena. 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)*, volume 1, pages 992–1002, 2015.
- [95] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, 2016.
- [96] Tanmoy Mukherjee and Timothy Hospedales. Gaussian visual-linguistic embedding for zero-shot recognition. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 912–918, 2016.
- [97] Rutu Mulkar-Mehta, Christopher Welty, Jerry R Hoobs, and Eduard Hovy. Using granularity concepts for discovering causal relations. In *Proceedings of the FLAIRS conference*, 2011.
- [98] Iftekhhar Naim, Young C. Song, Qiguang Liu, Liang Huang, Henry Kautz, Jiebo Luo, and Daniel Gildea. Discriminative unsupervised alignment of natural language instructions with corresponding video segments. In *Proceedings of NAACL HLT 2015*, pages 164–174, Denver, Colorado, May–June 2015. Association for Computational Linguistics.
- [99] Ad Neeleman, Hans Van de Koot, et al. The linguistic expression of causation. *The Theta System: Argument Structure at the Interface*, page 20, 2012.

- [100] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*, 2019.
- [101] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.
- [102] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [103] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision*, pages 651–667. Springer, 2016.
- [104] Ulrike Padó. *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. PhD thesis, Universitätsbibliothek, 2007.
- [105] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.
- [106] Megha Pandey and Svetlana Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *2011 International Conference on Computer Vision*, pages 1307–1314. IEEE, 2011.
- [107] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.
- [108] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [109] Sameer S Pradhan, Wayne Ward, Kadri Hacioglu, James H Martin, and Daniel Jurafsky. Shallow semantic parsing using support vector machines. In *HLT-NAACL*, pages 233–240, 2004.
- [110] J Pustejovsky. The syntax of event structure. *Cognition*, 41(1-3):47–81, 1991.
- [111] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, pages 909–918. ACM, 2012.
- [112] Nazneen Fatema Rajani and Raymond J Mooney. Combining supervised and unsupervised ensembles for knowledge base population. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, 2016.
- [113] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

- [114] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.
- [115] Terry Regier and Laura A Carlson. Grounding spatial language in perception: an empirical and computational investigation. *Journal of experimental psychology: General*, 130(2):273, 2001.
- [116] Terry Regier, Paul Kay, and Richard S Cook. Focal colors are universal after all. *Proceedings of the National Academy of Sciences of the United States of America*, 102(23):8386–8391, 2005.
- [117] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzels, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36, 2013.
- [118] Mehwish Riaz and Roxana Girju. 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, 2014.
- [119] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011.
- [120] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.
- [121] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition*, pages 184–195. Springer, 2014.
- [122] Raquel Ros, Séverin Lemaignan, E Akin Sisbot, Rachid Alami, Jasmin Steinwender, Katharina Hamann, and Felix Warneken. Which one? grounding the referent based on efficient human-robot interaction. In *19th International Symposium in Robot and Human Interactive Communication*, pages 570–575. IEEE, 2010.
- [123] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *Application of Computer Vision, 2005. WACV/MOTIONS’05 Volume 1. Seventh IEEE Workshops on*, volume 1, pages 29–36. IEEE, 2005.
- [124] Deb Roy. Grounding words in perception and action: computational insights. *Trends in cognitive sciences*, 9(8):389–396, 2005.
- [125] Deb Roy and Alex Pentland. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146, 2002.
- [126] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2010.

- [127] Karin Kipper Schuler. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2005.
- [128] Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. *arXiv preprint arXiv:1702.01841*, 2017.
- [129] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 357–360. ACM, 2007.
- [130] Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. Creating causal embeddings for question answering with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 138–148, 2016.
- [131] Lanbo She and Joyce Chai. Incremental acquisition of verb hypothesis space towards physical world interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 108–117, 2016.
- [132] Lanbo She and Joyce Chai. Interactive learning of grounded verb semantics towards human-robot communication. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1634–1644, 2017.
- [133] Lanbo She, Yu Cheng, Joyce Y Chai, Yunyi Jia, Shaohua Yang, and Ning Xi. Teaching robots new actions through natural language instructions. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 868–873. IEEE, 2014.
- [134] Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Chai, and Ning Xi. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 89–97, 2014.
- [135] Dan Shen and Mirella Lapata. Using semantic roles to improve question answering. In *EMNLP-CoNLL*, pages 12–21, 2007.
- [136] Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 1223–1237. Springer, 2002.
- [137] Jeffrey M Siskind. Naive physics, event perception, lexical semantics, and language acquisition. Technical report, DTIC Document, 1993.
- [138] Jeffrey Mark Siskind. Grounding language in perception. *Artificial Intelligence Review*, 8(5-6):371–391, 1994.
- [139] Jeffrey Mark Siskind. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *J. Artif. Intell. Res.(JAIR)*, 15:31–90, 2001.

- [140] Marjorie Skubic, Dennis Perzanowski, Samuel Blisard, Alan Schultz, William Adams, Magda Bugajska, and Derek Brock. Spatial language for human-robot dialogs. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2):154–167, 2004.
- [141] Grace Song and Phillip Wolff. Linking perceptual properties to the linguistic expression of causation. *Language, culture and mind*, pages 237–250, 2003.
- [142] Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [143] Robert Speer and Catherine Havasi. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686, 2012.
- [144] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007.
- [145] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.
- [146] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers to non-factoid questions from web collections. *Computational linguistics*, 37(2):351–383, 2011.
- [147] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, volume 1, page 2, 2011.
- [148] Stefanie Tellex, Pratiksha Thaker, Joshua Joseph, and Nicholas Roy. Learning perceptually grounded word meanings from unaligned parallel data. *Machine Learning*, 94(2):151–167, 2014.
- [149] Yao-Hung Hubert Tsai and Ruslan Salakhutdinov. Improving one-shot learning through fusing side information. *arXiv preprint arXiv:1710.08347*, 2017.
- [150] Matthew R Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. Learning semantic maps from natural language descriptions. In *Robotics: Science and Systems*, 2013.
- [151] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.
- [152] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

- [153] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.
- [154] Max Whitney and Anoop Sarkar. Bootstrapping via graph propagation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 620–628. Association for Computational Linguistics, 2012.
- [155] Phillip Wolff. Direct causation in the linguistic coding and individuation of causal events. *Cognition*, 88(1):1–48, 2003.
- [156] Phillip Wolff and Grace Song. Models of causation and the semantics of causal verbs. *Cognitive Psychology*, 47(3):276–332, 2003.
- [157] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4622–4630, 2016.
- [158] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. *arXiv*, 1603, 2016.
- [159] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, 123(3):309–333, 2017.
- [160] Xun Xu, Timothy M Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *European Conference on Computer Vision*, pages 343–359. Springer, 2016.
- [161] Zhongwen Xu, Linchao Zhu, and Yi Yang. Few-shot object recognition from machine-labeled web images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [162] Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Y. Chai. Grounded semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, San Diego, CA, 2016.
- [163] Xuefeng Yang and Kezhi Mao. Multi level causal relation identification using extended features. *Expert Systems with Applications*, 41(16):7171–7181, 2014.
- [164] Yezhou Yang, Cornelia Fermüller, and Yiannis Aloimonos. Detection of manipulation action consequences (mac). In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2563–2570. IEEE, 2013.
- [165] Yezhou Yang, Anupam Guha, C Fermuller, and Yiannis Aloimonos. A cognitive system for understanding human manipulation actions. *Advances in Cognitive Sysstems*, 3:67–86, 2014.

- [166] Xuchen Yao and Benjamin Van Durme. Semi-markov phrase-based monolingual alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 590–600. Association for Computational Linguistics, 2013.
- [167] Mark Yatskar, Vicente Ordonez, and Ali Farhadi. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of NAACL-HLT*, pages 193–198, 2016.
- [168] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542, 2016.
- [169] Chen Yu and Dana H Ballard. On the integration of grounding language and learning objects. In *AAAI*, volume 4, pages 488–493, 2004.
- [170] Haonan Yu, N Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. A compositional framework for grounding language inference, generation, and acquisition in video. *Journal of Artificial Intelligence Research*, 52:601–713, 2015.
- [171] Haonan Yu and Jeffrey Mark Siskind. 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, 2013.
- [172] John M Zelle and Raymond J Mooney. Learning semantic grammars with constructive inductive logic programming. In *AAAI*, pages 817–822, 1993.
- [173] John M Zelle and Raymond J Mooney. Inducing deterministic prolog parsers from treebanks: A machine learning approach. In *AAAI*, pages 748–753, 1994.
- [174] John M Zelle and Raymond J Mooney. Learning to parse database queries using inductive logic programming. In *Proceedings of the national conference on artificial intelligence*, pages 1050–1055, 1996.
- [175] Jie Zhou and Wei Xu. End-to-end learning of semantic role labeling using recurrent neural networks. 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 1127–1137, Beijing, China, July 2015. Association for Computational Linguistics.