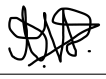
Saarland University
Department of Computer Science
Winter term 2020/2021
Seminar Common Sense Knowledge Extraction And Curation
Lecturer: Simon Razniewski

**Seminar Paper**

# Physical Common Sense

| | |
|---|---|
| Name: | Awantee Deshpande |
| Matriculation number: | 2581348 |
| Field of study: | Master of Computer Science, (semester 3) |
| Email: | s8awdesh@stud.uni-saarland.de |
| Date: | February 8, 2021 |

I hereby confirm that I have written the seminar paper with the title **Physical Common Sense** in the seminar **Common Sense Knowledge Extraction And Curation** (lecturer: **Simon Razniewski**) on my own and that I have not used any other media or materials than the ones referred to in this seminar paper.
I further confirm that I did not submit this or a very similar seminar paper in another seminar.

Saarbrücken, February 8, 2021 

# 1 Abstract

When asked a simple question like *"Can a baby lift a mountain?"* it is immediately obvious to humans that the answer is "No". It involves an implicit reasoning in our minds with respect to i) the sizes of a baby and a mountain, ii) the weights of a baby and a mountain, iii) the knowledge of what entities are animate, and iv) knowing that when $[size(entity_1), weight(entity_1)] << [size(entity_2), weight(entity_2)]$, $entity_1$ cannot lift $entity_2$.
While it is very easy for humans to formulate a logical thinking process to address such questions, this common sense understanding is not easily learned by natural language models. Modeling Physical Common Sense is one of the underlying challenges for today's Natural Language Processing and Natural Language Understanding systems. A primary reason for this is that the physical knowledge of the world is learned by humans through observation and intuition via various textual, visual, auditory and tactile stimuli. Most NLP models just learn the surface pattern relations between the input and output texts and therefore have no idea about the actual semantics of the representations. This lack of latent knowledge results in poor performance of language models despite the superior architecture, abundance of training data, and plenty of processing power. In this report, we give a brief overview of Physical Common Sense and present the problem of representation and learning it in NLP models. We acquaint the reader with the motivation and semantics for Physical Common Sense Reasoning. We also delve into the the research done by two prominent works in modelling Physical Common Sense. We further survey other areas of research that comprise data extraction, representation, and model architecture design for this task. Lastly, we present our own evaluation and discussion of the work studied and put forth the challenges and scope of researching Physical Common Sense in Natural Language Processing.

# Contents

## 2 Introduction

Recent advancements in NLP have made a lot of progress on Question-Answering on factual and contextual questions. Plenty of textual data is available on such domains in the form of newspaper articles, Wikipedia knowledge, knowledge bases etc. Humans may not always possess enough knowledge to answer such questions. On the other hand, there is another facet of questions that requires a common sense understanding to correctly answer them. Contrary to the above scenario, humans excel at answering such questions while NLP models struggle to come up with sensible answers. This can be illustrated with a simple example. On asking an average person, *"What is the capital of Bolivia?"* they will most likely not know the answer. However, if you ask someone, *"Can a goat drive a car?"* they will be able to answer this question, but it could easily stump an NLP model. The key to this difference is the underlying form of information and reasoning involved in answering the two questions.

There was a hiatus in common sense learning from between the 80s up to a few years ago. However, the situation today is different because we have more data, more processing power, crowdsourcing, stronger computational models, and better representations. In the recent years, efforts have been made to harness common sense knowledge and represent it in the form of knowledge graphs and concrete datasets [1].

As systems become more efficient and learn to obtain more data for extracting information, it becomes more and more important to imbibe in them a capacity to draw common sense inferences from the data they amass. Artificial intelligence can benefit a lot by incorporating common sense knowledge in the background such as **(ice-hasProperty-isCold)**, **(chewing-subEventOf-Eating)**, **(steeringWheel-locatedIn-car)** etc. Unlike most machine learning tasks, common sense intelligence is not just about pattern recognition. It is about modelling the world in terms of:

1. *Explaining* and *understanding* what is observed
2. *Imagining* things that haven't been observed yet
3. *Problem solving* and *planning* actions to make this real
4. *Building* newer models continuously

There are different kinds of common sense such as physical, social, temporal, generic etc. In this report, we focus on the problem of Physical Common

Sense Reasoning. Section 2.1 follows with the motivation behind researching this problem. Sections 3 and 4 give a brief introduction to some terms and concepts and how to approach the problem of Physical Common Sense modelling. A comprehensive literature review with respect to various approaches studied so far has been given in Section 5, and Sections 6 and 7 conclude by summarising the challenges and future scope of this research.

### 2.1 Motivation

Before children learn how to speak, they start perceiving the world around them through shapes, objects, and colours [2]. Thus, before learning to communicate, human beings acquire physical common sense knowledge about everyday objects. It is this physical common sense that enables people to solve tasks like putting letters in envelopes and electronic gadgets in larger cardboard boxes. Reading between the lines, localised knowledge, implicit understanding are all important aspects of using common sense reasoning. A critical part of this is to gain Physical Common Sense knowledge vis-à-vis the understanding of phenomena via naïve physical and psychological beliefs, and social norms. Common sense knowledge is especially useful when unfamiliar scenarios are encountered because humans can use a combination of imagination and reasoning to come up with rational decisions in the face of unknown situations. For instance, when humans understand that they can catch a cold from getting wet in the rain, they learn to dry themselves no matter how they get wet.

Natural language models should ideally be able to learn through textual representations what humans learn through perceptions. Given a sentence, a model should be able to determine if it is semantically plausible even if the represented text is not explicitly fed in the trained model. This Physical Common Sense understanding is important in downstream tasks like story understanding, paragraph reconstruction, and coreference resolution [3]. For instance, on given a sentence like *"She picked up the coin from the table and put it in her purse,"* an NLP model would need to resolve if *'it'* refers to the coin or the table. Physical Common Sense would dictate that a table cannot be put into a purse and thus would help to resolve this ambiguity.

Thus, natural language models would benefit largely from being able to model Physical Common Sense. Some research has been done in Computer Vision to solve the same problem [4, 5, 6, 7]. Images can give a good relative perspective to the physical dimensions of objects. However, properties

such as weight, strength, speed etc. cannot be determined through currently available image recognition techniques. Such examples can, however, be richly available in textual format, either directly (*"The cheetah is the fastest mammal"*) or indirectly (*"Muhammad Ali beat Sonny Liston"* $\Rightarrow$ Muhammad Ali $>^{strength}$ Sonny Liston). While natural language is a rich resource of broad knowledge in itself, compiling trivial Common Sense Knowledge from natural language is a nontrivial feat.
Hence, it becomes necessary to come up with new resources and techniques to harness Physical Common Sense knowledge and improve the current state-of-the-art Natural Language Understanding scene.

## 3 Terminology

In this section, we describe some terminology and notations that will enable the reader to understand the current research and explain the task of Physical Common Sense Reasoning better.

- **Physical Common Sense**: Any aspect of understanding and reasoning about the physical properties of the world such as size, weight, speed, height etc. comes under the umbrella of Physical Common Sense.

- **Reporting Bias**: Reporting bias is defined as the "selective revealing or suppression of information" by subjects. In simple terms, the information acquired in the world is through what has been explicitly stated and put down. In normal situations, people don't state what is evident, and if they do state something, it is usually as an exception [8, 9].
  e.g. One would never have to state *"Humans are larger than spiders."* But what might be mentioned is, *"That spider was as large as a cat!"* Reporting bias describes the discrepancy between what is frequent in text and what is likely in the world. This is in part because people do not describe the obvious.

- **Semantic Plausibility**: Determining whether a piece of text is semantically plausible even if it is syntactically correct. For example, the sentence pattern PERSON-ATE-NOUN is syntactically sound, but saying *"She ate an apple"* is semantically plausible, while *"I ate a chair"* is not.

- **Selectional Preference**: This is complementary to the notion of semantic plausibility, and has to do with the typicality of the event. Out of a possible set of semantically plausible notions, selectional preference chooses what is more likely to be observed.
  e.g. The cloze pattern MAN-RIDES-____ would predict "bicycle" over "gorilla" even if both are semantically plausible.

- **Annotation artifact**: An artifact is something observed in a scientific investigation or experiment that is not naturally present but occurs as a result of preparatory or investigative procedures. Human annotation on datasets can generate such annotation artifacts that can help a classifier detect the correct class without ever observing the premise. This is one of the reasons why models do not learn the latent representations of the text.

- **Self-supervision**: In self-supervision, human-annotated data is not used for learning. Examples of self-supervision in NLP include predicting the next word in the sentence, masking a word and predicting the blank, replacing words in sentences and checking whether the sentence makes sense etc.

- **Distributional Model**: Distributional models believe that the statistical distribution of linguistic items in a context plays a key role in characterising their semantic behavior. For instance, such a model would surmise that similar words occur together in text. While learning word embeddings, the word representations are learned from the distribution of their context.

- **Probing Model**: Probing methods train supervised models to predict linguistic properties from representations of the language. The probing task is designed in such a way to isolate some linguistic phenomena. If the probing classifier performs well on the probing task, we infer that the system has encoded the linguistic phenomena in question. Thus, probing models can be used to learn parts-of-speech, syntactical structure, chunking etc.

## 4 A Background of Physical Common Sense

In the recent years, the need for integrating reasoning capacities for downstream tasks like question-answering, smarter dialogue, text generation etc. has emerged. Advances in large pretrained models like BERT, GloVe, and

ELMo have pushed machines closer to humanlike understanding capabilities, calling into question whether machines should directly model common sense through symbolic integrations. When the Winograd Schema Challenge [10] was introduced, this spurred the research community to come up with better and more innovative solutions for modelling common sense, a rich spectrum of which is covered by Physical Common Sense. The research being done with Physical Common Sense Reasoning constitutes determining how language models can be best used to harness and represent Physical Common Sense knowledge w.r.t the amount of training data required, the means and methods for obtaining data, the format for data representation, the models used for training, and suitable evaluation metrics and benchmarks.
There are three primary questions to address when it comes to Physical Common Sense Reasoning.

1. How to represent Physical Common Sense data?
2. How to model the training task from the data and vice versa?
3. How to establish the model architecture for learning Physical Common Sense?

The research methodologies utilise a dual approach to tackle the issue of common sense learning. This approach studies:

1. Better methods to represent Physical Common Sense through knowledge graphs, triples etc.
2. Improved model architectures (e.g. deep learning networks) that can model the latent processes by abstracting away from the surface patterns.

### 4.1 Knowledge bases

At first, symbolic logic approaches were the main type of data representation [11, 12]. Today, the computational advantages allow for a more data-driven knowledge collection and representation.

General knowledge bases like DBPedia [13] and YAGO [14] model the real world in terms of entities and relations. However, they are very broad in nature and largely depend on tangible knowledge. There are specific knowledge bases catered to common sense knowledge, like ATOMIC [15], ConceptNet [16], WebChild [17] etc. However, there aren't many such knowledge

graphs that capture Physical Common Sense only, though Physical Common Sense is imbibed in projects like NELL [18] (e.g. relations like *Animal-Eats-Vegetable*, *Cloak-ClothingItem-GoesWithDress*) and has been used for such research.
There are also knowledge bases such as DEKO [19] and its extension, KNEWS [20] that are targeted to generating an information base for prototypical knowledge about objects. However, this research was designed for a robotics based perspective for regular household chores and uses, and has not been typically used for general Physical Common Sense research.
Knowledge graphs can be used as they are by directly retrieving triples from them. Some approaches have also used knowledge graphs to create word embeddings to use for further training [21, 22, 23].
These embeddings can be:

(a) Translational - The score of a triple $(e_h, r, e_t)$ is modelled as the distance between $e_h$ and $e_t$ via $r$.

(b) Semantic - The latent semantics of $e_h$ and $e_t$ are represented as tensors.

While translational embeddings suffice for surface tasks like factual question-answering, common sense reasoning requires the model to understand latent relationships, and would work better with semantic embeddings. Thus, one of the prominent scopes of research for Physical Common Sense Reasoning is the creation and development of a knowledge base modelling the physical properties and attributes of daily world objects. However, as Zhao et al. [24] succinctly list down in their paper, this is a challenge because:

1. Training facts are scarce - when labeling the properties of an object, people usually name the ones that are easiest to think of but cannot enumerate all properties.

2. Not all facts can be mined from existing texts - this stems from the problem of reporting bias stated in Section 3.

3. The number of relationships is small and all are $n-n$ relationships, which makes modelling relationships between entities more complicated.

### 4.2 Datasets

While there are quite a few datasets available for common sense representation, especially Physical Common Sense learning, the decision of how to

model the learning task greatly motivates the choice of dataset to use. A prequel of this is the creation of such a dataset if none are available. The most commonly used methods are either extractive in nature or require crowdsourcing. Extractive methods largely scrape a portion of the Web and other corpora and then process the extracted text to form structured training data. The three things to consider here are the source of extraction, method of extraction, and choice of consolidation.

Text extraction methods can also be used to construct labelled triples that are then used to train a model for physical plausibility. We will see this in more detail in Section 5.

Other works have also used existing general text-based and image datasets such as NELL, the McRae et al. [25] property norm dataset, and MS COCO object dataset for Computer Vision [26]. These datasets are filtered and preprocessed to produce appropriate data for a Physical Common Sense learning task by obtaining objects, their dimensions, physical properties, and action sets.

For the downstream task of Question-Answering, the PIQA dataset [3] is a commonly used benchmark. PIQA focuses on everyday situations, and is created by scraping the web and using crowdsourced annotations. Given a physical goal, the model should be able to choose a sensible solution from a given set of options. An example is shown in Figure 1.

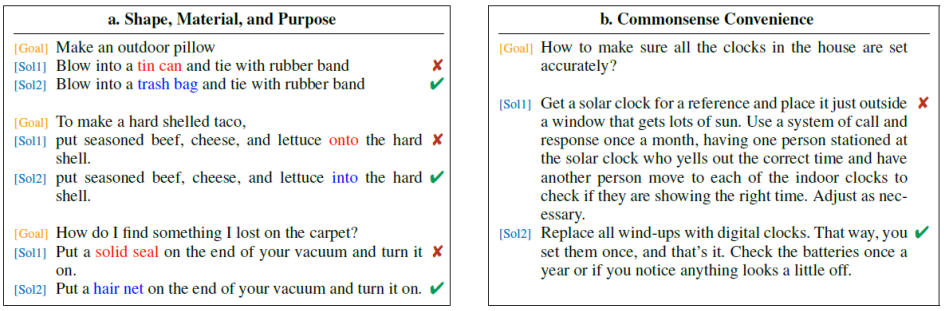



Figure 1: Source: Bisk et al. (2020) - **Left** are examples that require knowledge of basic properties of the objects (flexibility, curvature, and being porous), while on the **Right** both answers may be technically correct but one is more convenient and preferable.

The VerbPhysics dataset by Forbes et al. [27] uses an inference of relative physical knowledge to predict if $word_1 < / > / \cong word_2$ when compared on an attribute, for example, $bed >^{weight} hand$ or $mouth \cong^{size} fist$. This dataset is created by a combination of extraction from the Google Syntax Ngram Corpus and crowdsourcing .

### 4.3 Models and Architecture

Most language models now utilise the vast diversity of neural networks to encode linguistic phenomena for downstream tasks. While some approaches like Support Vector Machines (SVMs) have been used for Common Sense Classification [28], thanks to pretrained language models, the focus has largely shifted to deep learning architectures [29]. An entire neural network can replace the traditional NLP pipeline of tokenisation, POS tagging, parsing, coreference resolution etc. Pretrained word embeddings from BERT, GloVe etc. have shown promising results in the task of Physical Common Sense Reasoning. There are two types of Natural Language Understanding that come into consideration for NLP tasks:

1. Shallow NLU that has a strong alignment between the input and output and can perform extremely well only through surface pattern matching. e.g. Translating *"The ant kicked the football"* from English to German would give *"Die Ameise kickte den Fußball."*

2. Deep NLU has a weak alignment between the input and output. It is based more on abstraction, common sense, and reasoning. The neural network must learn that *"The ant kicked the football"* is not physically possible.

Various works have looked at directly encoding common sense knowledge from structured KBs as additional inputs to a neural network. Where knowledge bases are unavailable, corpus statistics pulled from unstructured text have been used to aggregate Common Sense Reasoning. More recently, rather than providing relevant common sense as an additional input to neural networks, researchers have looked into indirectly encoding common sense knowledge into the parameters of neural networks through pretraining on common sense KBs, explanations, or by using multi-task objectives with common sense relation prediction.

Despite the advances in neural network models, they can be challenging to work with. The main problem is that neural networks are not robust.

They fail the moment you give them unfamiliar, out-of-context, or adversarial examples. Since common sense largely relies on implicit knowledge and imagination, it is difficult to design deep learning models that can function well for such tasks.

# 5 Current Research

The previous sections provided the reader with a background and motivation for learning Physical Common Sense knowledge. The following section reviews recent work that studies what models capture during their pretraining, what they can be fine-tuned to capture, and what types of knowledge are not captured by them.

## 5.1 Do Neural Language Representations Learn Physical Commonsense?

The research by Forbes et al. [5] served as a baseline for further studies into neural understanding of Physical Common Sense. In this paper, the authors train latest neural models like BERT and ELMo on vast amounts of text to investigate the extent to which they can demonstrate Physical Common Sense Reasoning. The idea is to see that knowing properties and affordances of objects, can they be implicitly learned to map to each other? Properties are the static attributes of an object (e.g. *apple-is_fruit*), while affordances are the actions that can be applied to them (e.g. *apple-is_eaten*). Then, the plausibility of (*fruit-is_eaten*) can be determined knowing these object-property and object-affordance relations.
This requires a massive amount of text to train on.

(a) Task
The main task is to map the compatibility of two instances $(i_1, i_2) \rightarrow 0, 1$. The instances can be objects ($O$), absolute properties ($P$), or affordances ($A$), and 0 and 1 are labels for compatibility. Absolute properties do not rely on relative comparisons between objects (e.g. *"fit x into y"*$\Rightarrow x <^{size} y$), but fixed properties (e.g *look-through(x)* $\Rightarrow$ *transparent(x)*). The mapping is formed as object-property ($O \leftrightarrow P$), object-affordance ($O \leftrightarrow A$), and affordance-property ($A \leftrightarrow P$) compatibility.

(b) Datasets
The authors introduce two new datasets. The first is an *abstract*

dataset of objects and properties, drawn and filtered from the McRae et al. data [25] and the MS COCO Computer Vision dataset [26] and augmented with additional properties. The second is a *situated* dataset which uses situated images to resolve visual ambiguities in appearance and shape (e.g. a glass bottle is different from a plastic bottle). Objects sampled from these images are then annotated with properties introduced for the abstract dataset. They are also annotated by crowdsourcing with affordances from the verbs in the imSitu dataset. Some examples have been illustrated in Figure 2.

| **Objects** | **Properties** | **Affordances** |
|---|---|---|
| *harmonica, van* | *expensive, squishy* | *pick up, remove* |
| *potato, shovel* | *used as a tool for cooking* | *pet, talk to* |
| *cat, bed* | *decorative, fun* | *cook, throw out* |

Figure 2: Source: Forbes et al.(2019) - Examples of **objects**, **properties**, and **affordances** in the abstract and situated datasets.

(c) Model
For training, sentences are generated using the objects, properties, and affordances.
e.g.
O-P *(fork/light_ weight): "A fork is light."*
O-A *(elephant/photograph): "He photographed the elephant."*
A-P *(eat/expensive): "If you can eat something, then it is expensive."*
A sentence has the structure $\{s = w_1,\ w_2 \ldots\ w_n\}$. Four kinds of word embeddings are considered for training. These are GloVe embeddings [30], Dependency Based Word Embeddings [31] that are created from dependency parse trees, and ELMo [32] and BERT [33] contextualised representations. For any two words $w_i$, $w_j$ that occur together in the sentence the embeddings obtained are $r(w_i, w_j)$. Each of these models is then fine-tuned by adding a Multilayer Perceptron (MLP) after the input representations and using a mean squared loss with L2 regularisation.

$$\texttt{MLP Layer} : \hat{y}_{w_i,w_j} \propto \sigma(W_2 \times a(W_1 \times r(w_i, w_j) + b_1) + b_2)$$

$$\texttt{Loss function} : \mathcal{L}(w_i, w_j, y, \theta, \lambda) = (y - \hat{y}_{w_i,w_j})^2 + \lambda||\theta||_2^2$$

$\theta = \{W_1, W_2, b_1, b_2\}$ are the trainable parameters.

(d) Evaluation
The evaluation is done against a random baseline, majority baseline, and human performance. The results are shown in Figure 3.

| | **Abstract** | | | | | | | | **Situated** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | O ⟷ P | | | | O ⟷ P | | | | O ⟷ A | | | | A ⟷ P | | | |
| | *obj* | *prop* | *μF1* | *sig* | *obj* | *prop* | *μF1* | *sig* | *obj* | *aff* | *μF1* | *sig* | *aff* | *prop* | *μF1* | *sig* |
| RANDOM | 0.25 | 0.26 | 0.26 | *** | 0.24 | 0.25 | 0.22 | *** | 0.53 | 0.62 | 0.51 | *** | 0.24 | **0.26** | 0.23 | *** |
| MAJORITY | 0.34 | 0.11 | 0.31 | *** | 0.16 | 0.05 | 0.17 | *** | 0.82 | 0.68 | 0.82 | *** | 0.18 | 0.05 | 0.17 | *** |
| GLOVE | 0.63 | 0.47 | 0.63 | *** | 0.55 | 0.39 | 0.57 | | 0.85 | 0.73 | 0.86 | | 0.27 | 0.13 | 0.29 | |
| DEP-EMBS | 0.62 | 0.42 | 0.60 | *** | 0.54 | 0.36 | 0.54 | *** | 0.84 | 0.67 | 0.84 | * | 0.26 | 0.12 | 0.28 | |
| ELMO | 0.67 | 0.55 | 0.67 | ** | 0.58 | 0.44 | 0.58 | *** | 0.84 | 0.71 | 0.85 | ** | 0.31 | 0.17 | 0.34 | |
| BERT | **0.74** | **0.67** | **0.74** | ← | **0.64** | **0.59** | **0.67** | ← | **0.87** | **0.77** | **0.88** | ← | **0.36** | 0.25 | **0.37** | ← |
| HUMAN | 0.78 | 0.80 | 0.67 | | 0.70 | 0.69 | 0.61 | | 0.83 | 0.93 | 0.80 | | 0.65 | 0.67 | 0.40 | |

Figure 3: Source: Forbes et al. (2019) - Macro and micro F1 scores. Statistical significance (*sig*) is calculated by comparing the best-scoring model with each other model ($*$ for $p < 0.05$, $**$ for $p < 0.01$, $***$ for $p < 0.001$).

(e) Discussion
The $(O \leftrightarrow P)$ and $(O \leftrightarrow A)$ tasks are found to have a much better performance than the $(A \leftrightarrow P)$ task. The inference between affordances and properties requires multi-hop reasoning that is simply not present in the pretraining of large text-based models. The properties are classified by the authors as functional (e.g. *used_for_writing*), encyclopaedic (e.g. *is_a_car*), common sense (e.g. *comes_in_pairs*), and perceptual (e.g. *is_soft).* The system is best at classifying functional properties, since they are directly tied to affordances. Perceptual properties exhibit lower F1 scores. There is also a clear lack of correlation in the frequencies of properties and affordances in the corpus and the model performance. This leads the authors to conclude that current neural models are fundamentally limited in their capacity for physical reasoning, and that only new designs, and not more data, can allow them to acquire this skill.

We also evaluate this system based on the implementation provided by the authors[1]. The steps for evaluation along with the reproduced results are present in the repository[2]. The outputs show the types of sentences are generated for training, training steps over the epochs, and the losses and accuracies over the training and test sets. We find that the results are equivalent to those presented by the authors. As

[1] `https://github.com/mbforbes/physical-commonsense`
[2] `https://github.com/awanteedeshpande/seminar-physical-common-sense`

mentioned earlier, neural models do not easily capture the latent relationships between the objects, properties, and affordances, which is why massive amounts of training text still cannot capture implicit relationships. However, stating that neural architectures alone should be improved on to tackle this problem would be an understatement.

## 5.2 Can a Gorilla Ride a Camel?

Distributional models of words are based on the hypothesis that meanings of words can be inferred on the basis of the context they appear in. Word embeddings are generated from such models because they capture the semantic context. In their work, Porada et al. [34] show that in the supervised setting, pretrained language models are effective at modelling semantic plausibility. The approaches to selectional preference are also distributional and show limited performance for capturing semantic plausibility. Since semantic plausibility and selectional preference are closely related, even this should be factored in while learning the model.

(a) Task
This is the problem of determining if a given event, represented as an S-V-O triple, is semantically plausible (refer to Figure 4). The idea is to prove that distributional data does provide a strong cue for semantic plausibility in a supervised setting.

| Event | Plausible? |
|---|---|
| *bird-construct-nest* | ✓ |
| *bottle-contain-elephant* | ✗ |
| *gorilla-ride-camel* | ✓ |
| *lake-fuse-tie* | ✗ |

Figure 4: Source: Porada et al. (2019) - Example events for modelling semantic plausibility.

(b) Dataset
The crowdsourced dataset from Wang et al. [35] comprising S-V-O triples is used. The original dataset contains 3062 S-V-O triples, built from a vocabulary of 150 verbs and 450 nouns. As a further contribution, the work also studies the problem of learning to model semantic plausibility directly from text. For this, the authors construct two

training sets. In the first one, they parse English Wikipedia using the StanfordNLP neural pipeline to extract 6 million unique S-V-O triples. For the second one, 2.5 million triples are filtered out from the NELL dataset [18].

(c) Model and Evaluation
The baseline is a two-layer Artifical Neural Network (ANN) over static embeddings. In the supervised setting, GloVe embeddings are used. Because the supervised approach is likely to have the same words over the test and training sets, self-supervision is done on the textually extracted dataset. Attested events are considered to be plausible, while fake implausible events are created by sampling of individual words in S-V-O triples independent of their occurrence.
The authors focus on the usage of the masked language model BERT by treating this as a sequence classification task on the same supervised and self-supervised settings. The input is of the format:

$$[CLS][SUBJ]<subject>[/SUBJ][VERB]<verb>[/VERB][OBJ]<object>[/OBJ][SEP]$$

(d) Discussion
The results for the supervised and self-supervised settings are shown in Figure 5. The supervised accuracy scores on a smaller dataset are higher than the self-supervised scores on a large dataset. A point to note here is that supervised performance depends on the coverage of the training set vocabulary and is also affected by annotation artifacts. Secondly, the size of the training set in the supervised setting is limited, and results in overlapping training and test sets. Wang et al., in their paper, have tried to improve the performance by injecting world-knowledge about object properties with the help of crowdsourcing. However, this requires large manual interference. Self-supervision from extracted text helps to avoid this bias towards the training corpus. In their experiments, the authors find that BERT performs the best in all settings, though it is biased to labelling events as plausible.

There is no implementation or demo of this work available online. Nevertheless, with the help of this research, we can conclude that distributional models can be leveraged to learn semantic plausibility. At the same time, extra steps must be taken (e.g. learning from text, injecting world knowledge etc.) to ensure that they are not biased to annotation artifacts, and that the training set and test sets have minimal overlap.

| Model | Accuracy |
|---|---|
| Random | 0.50 |
| NN (Van de Cruys, 2014) | 0.68 |
| NN+WK (Wang et al., 2018) | 0.76 |
| Fine-tuned BERT | **0.89** |

(a) Supervised Setting

| Model | Wikipedia | | NELL | |
|---|---|---|---|---|
| | Valid | Test | Valid | Test |
| Random | 0.50 | 0.50 | 0.50 | 0.50 |
| NN | 0.53 | 0.52 | 0.50 | 0.51 |
| BERT | **0.65** | **0.63** | **0.57** | **0.56** |

(b) Self-supervised Setting

Figure 5: Source: Porada et al. (2019) - Results showing the accuracy of the models for classifying plausible events.

### 5.3 Other work

Section 4 mentions the ways to solve the problem of Physical Common Sense learning by finding better representations, or designing better models to learn the latent relations and implications in word contexts.
We give the reader a brief introduction to the ideas proposed in other popular research.

1. Physical Interaction: Question Answering (PIQA) - The focus of this research by Bisk et al. [3] is to create a benchmark targeted for Physical Common Sense understanding. The task is modelled as follows: Given a question goal $q$ and two possible solutions $s1$ and $s2$, the correct solution has to be picked. The inspiration is drawn from the *instructables.com*[3] site that contains instructions for doing everyday household tasks. Annotators were asked to provide semantic perturbations or alternative solutions that differed slightly by syntax or topic. The dataset was cleaned with the AFLite algorithm [36]. AFLite uses an ensemble of linear classifiers to determine if pre-computed embeddings of the data are strong indicators of the correct solution option. Anything that constituted as 'expert knowledge' was removed to ensure the data was strictly about common sense solutions. Thus, the final dataset obtained is in the form of Goal-Solution pairs[4] (Refer Figure 6). The PIQA dataset is also integrated into standard libraries such as HuggingFace and TensorFlow.

   After evaluating on models like GPT, BERT, and RoBERTa, it was found that GPT worked the best, because the dataset was designed to be adversarial to BERT. Striking observations made by this work

[3] `https://instructables.com`
[4] `https://yonatanbisk.com/piqa/`

```
{
"goal": "How do I ready a guinea pig cage for its new
    occupants?",
"sol1": "Provide the guinea pig with a cage full of a few
    inches of bedding made of ripped paper strips, you will
    also need to supply it with a water bottle and a food
    dish.",
"sol2": "Provide the guinea pig with a cage full of a few
    inches of bedding made of ripped jeans material, you
    will also need to supply it with a water bottle and a
    food dish."
}
```

Figure 6: Source: PIQA by Bisk et al. (2020) - Sample of a Goal-Solution pair.

include high versatility of words like 'water' in different applications, making it difficult when the two solutions differ only by words like these. On the contrary, words like 'spoon' have a specific meaning and get better accuracy. The authors provide a strong conviction that when two solution choices differ by editing a single phrase such as replacing 'water' by 'milk', 'before' by 'after' etc., the model must by definition check the common sense understanding of that phrase.

2. Verb Physics - This research by Forbes et al. [27] presents an approach to infer relative physical knowledge of actions and objects from unstructured text. The task focuses on learning the joint inference of two problems - relative physical knowledge of object pairs, and the physical implications of actions applied to object pairs. The principle behind this is that there is consistency in the way people describe how they react with the world, which provides clues to reverse engineer the knowledge. Saying *"x threw y"* usually implies that x is heavier and larger than y. Given a certain action, a probabilistic distribution is assumed over 5 dimensions (*size, weight, strength, rigidness, speed*) as an implication.
   $O^a_{x,y}$ denotes a random variable with range $\{>, <, \cong\}$ along dimension $a$. e.g. $\{O^{weight}_{x,y} = <\}$ indicates $x <^{weight} y$.
   $F^a_v$ denotes the implication of action word $v$ when applied to the arguments $x$ and $y$ w.r.t the dimension $a$.
   e.g. $F^{size}_{threw} \Rightarrow$ ``x threw y" $\Rightarrow x >^{size} y$.
   Since a verb can have many meanings $(v_1 \dots v_t)$ according to the frame

it is used in, the random variable is denoted by $F^a_{vt}$. For example, the phrase *'pick up'* has different relations in the sentences *"I will pick you up after work"*, *"He picked up the book from the floor"*, and *"The sales usually pick up around Christmas."*
The data is obtained through crowdsourcing and the tasks are modelled as a probabilistic inference over a factor graph of knowledge. The graph consists of object pairs and frame nodes. The subgraphs are connected by selectional preference. This means that if there is sufficient evidence that $x >^{size} y$, then "x threw y" should entail this. One such example is illustrated in Figure 7, obtained from the contributions that are available online [5].

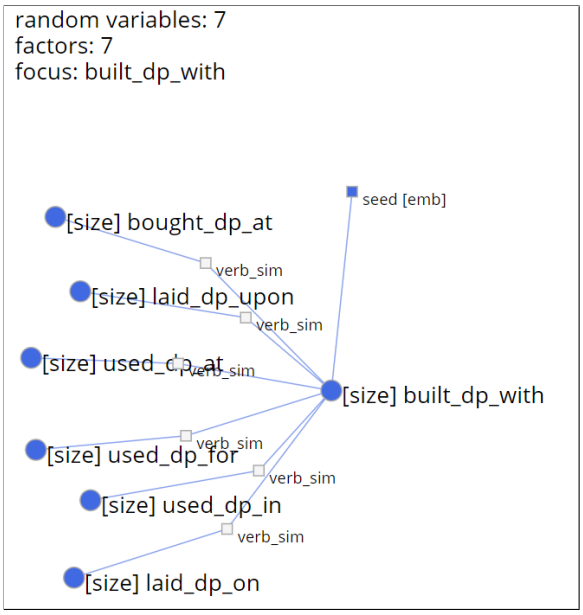



Figure 7: Source: Forbes et al. (2017) - Action frame for sentence type <person> built <object1> with <object2> w.r.t the size dimension.

3. Pretrained Embeddings for Common Sense Comparison - As mentioned in Section 3, probing models learn linguistic phenomena instead of just surface patterns, and have better accuracy than baselines which use dataset artifacts. In the paper by Goel et al. [37], the authors check if a probing model can learn to represent Physical Common Sense through its word embeddings, and analyse how such models compare

[5] `https://uwnlp.github.io/verbphysics/`

objects. A classification task is learned on the data where two objects with a property are passed as input and the output should state if the comparative physical property holds true.
e.g. *(car, aeroplane, bigger)* $\rightarrow 0$
The authors use a single layer fully connected network. The input is the word embeddings which are concatenated or subtracted and passed to the network (refer Figure 8). Both, GloVe and BERT embeddings are tested on the Verb Physics dataset by Forbes et al (2017) [27]. The authors evaluate this model against baselines like Majority-Class, Verb-Centric Frame Semantics using probabilistic graphical modelling, and Property-Comparisons from Embeddings.

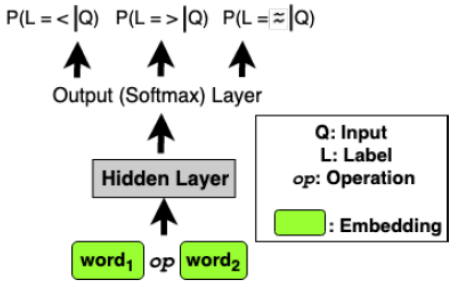



Figure 8: Source: Goel et al. (2019) - The Probing Model.

Interestingly, the research shows that the probing model learns an ordering of objects and uses it for comparison. This is observed in the ouput logit label for the objects. The results show that ELMo and GloVe embeddings are comparable to each other and both better than BERT embeddings, contrary to other NLP tasks (refer Figure 9).

4. Knowledge Graph Completion with BERT - In this study, Zhao et al. [24] try to extend the capability of BERT as studied by Forbes et. al (2019) [5] to learn Physical Common Sense interactions. They make two unique contributions. Firstly, they formulate the dataset as a knowledge graph using tensors. Secondly, they combine the pre-trained model architecture and knowledge graph embeddings to form a novel pipeline. The knowledge graph has the following characteristics: $(e_h, r, e_t)$ are the head entity, relation, and tail entity respectively. They are labelled as $(e_h, r, e_t) = 1$ denoting the fact is true in the training data, and $(e_h, r, e_t) = 0$ denoting the fact does not exist or is false in the training data. Extending the dataset from Forbes et al. (2019),

| | Majority Class Baseline | F&C | PCE | Probing Model (GloVe) | Probing Model (ELMo) | Probing Model (BERT-base) |
|---|---|---|---|---|---|---|
| **Size** | 0.66 | 0.75 | 0.80 | **0.82** | **0.82** | 0.80 |
| **Weight** | 0.67 | 0.74 | 0.81 | **0.82** | **0.82** | 0.80 |
| **Strength** | 0.66 | 0.71 | 0.77 | 0.78 | **0.79** | 0.75 |
| **Rigidness** | 0.60 | 0.68 | 0.71 | 0.71 | **0.72** | 0.71 |
| **Speed** | 0.59 | 0.66 | 0.72 | 0.72 | **0.76** | 0.71 |
| **Overall** | 0.64 | 0.71 | 0.76 | 0.77 | **0.78** | 0.75 |

Figure 9: Source: Goel et al. (2019) - Accuracy of the probing model compared with the baselines. The simple probing model achieves better accuracy indicating that the pre-trained representations capture common sense physical comparisons.

entities come from the set of all objects, properties, and affordances, and relationships come from the set of $O \leftrightarrow P$, $O \leftrightarrow A$, and $A \leftrightarrow P$. Since unknown facts cannot be assumed to be negative samples, for every $(i_1, i_2)$ relationship, $NOT(i_1, i_2)$ and reverse relations are explicitly added.
e.g. *(person, NOT-OP, a_ tool)*, *(cup, NOT-OA, twist)*, *(walk, NOT-AP, used_for_ eating)*.
The Tucker Factorisation proposed by Balazevic et al.(2019) [21] approximates the 3-way binary tensor of all $(e_h, r, e_t)$ tuples $X \epsilon \{0, 1\}^{n_e \mathbf{x} n_e \mathbf{x} r}$ and optimises the final objective jointly on these tasks. The details can be found in the paper by Zhao et al. [24]. With the learned joint objectives, the probabilities of $e_h$ and $e_t$ being related by $r$ can be determined. The results show an improved performance on triple classification as correct or incorrect, and link prediction of $e_t$ when $e_h$ and $r$ are given.

## 6 Challenges

Despite the different advances made in model architectures and data representation techniques, modelling Physical Common Sense is not an easy task. The major problem is the reporting bias associated with natural language texts. One way to put this is that "Not everything that is plausible occurs, and not everything that occurs is unbiased." The focus of all the papers mentioned in this report is the English language. Modelling Physical Common Sense would also be a considerable challenge for low-resource languages. Different techniques would be required to handle the morphological and syn-

tactical differences in languages other than English.
Most current language models are based on neural architectures that are not robust and do not have the "imaginative capacity" needed to reason for common sense scenarios. Explainability is also an important point of consideration with deep learning architectures. Explainable architectures would be useful to understand exactly how the latent semantics of the text are captured by language models to represent Physical Common Sense knowledge. The common issues across all natural language texts are summarised below:

1. Usage of dataset - As mentioned previously, the training corpus utilised can highly bias the model. The coverage of the training set limits the generalisation ability of supervised models. Because of reporting bias, atypical phrases can occur more often in the corpora than the typical ones. For instance, Ó Séaghdha.(2010)[38] has shown that the subject-verb bigram *carrot-laugh* occurs 855 times in a web corpus, while *manservant-laugh* occurs zero times.

2. Representation methods - The current research has focused on knowledge bases, distributional models, word embeddings, and contextualised embeddings to represent Physical Common Sense. But these methods do not encode the salient features of knowledge. Relative object comparisons are also not learned by methods like selectional preference.

3. Training methods - How to train the language models for Physical Common Sense Reasoning is another challenge. It is vital to consider the downstream task usage of the model when the training task is being formulated. This makes Physical Common Sense learning and reasoning important from the perspective of where in the model pipeline this understanding will be resolved.

# 7 Summary & Discussion

In this report, we present the problem of Physical Common Sense modelling. We take the reader through a background on Physical Common Sense, common problems associated with learning it, and the different approaches that have been utilised to tackle these problems. The current research focuses on representations through knowledge graphs, embeddings etc. and on improving model architectures by using probing models, distributional models, graph based embeddings, and optimisations like Tucker Factorisation.

We also get an understanding of the challenges involved in modelling common sense knowledge as semantic plausibility, selectional preference, and question-answering tasks because of factors like reporting bias, annotation artifacts, implicit knowledge etc.
Explainability and downstream usage are two important factors to consider while training language models. BERT-like models used in classifying OP, OA, or AP compatibility or semantic plausibility are not able to explain how exactly the latent relationships are modelled, making it difficult to decide how to improve these methods. The paper by Goel et al. [37] clearly showed that the probing model learns a global comparison of objects along physical properties. This is necessary for other techniques as well. While commendable research has been done in modelling semantic plausibility and physical compatibility in both the papers, they don't seem to have any well-defined downstream usage. One can't categorically state where to use this binary classification step in a standard model pipeline for solving any NLP task.

There is no doubt that this field has a tremendous scope for research. Physical Common Sense Reasoning can be directly correlated to tasks that involve modelling the physical world. These include the improvement of physics engines, prediction of physical phenomena through perceptual stimuli (such as,*"What would happen if a bowling ball is hurled against a stack of bottles vs. against a wall?"*) and so on. More ambitious explorations could emphasise the development of systems that use Physical Common Sense knowledge to learn on their own. Rather than providing external scenarios, such systems would be able to use common sense to create alternative scenarios independent of human interference and learn from them.
The ultimate aim is to to enable the construction of language models that provide common sense reasoning useful beyond the NLP community.

## References

[1] Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. Commonsense reasoning for natural language processing. *Association for Computational Linguistics*, pages 27–33, 2014.

[2] Spelke E Hespos S. Conceptual precursors to language. *Nature 430*, pages 453–456, 2004.

[3] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019.

[4] Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. Are elephants bigger than butterflies? reasoning about sizes of objects, 2016.

[5] Maxwell Forbes, Ari Holtzman, and Yejin Choi. Do neural language representations learn physical commonsense? *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, 2019.

[6] Hamid Izadinia, Fereshteh Sadeghi, Santosh K. Divvala, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. Segment-phrase table for semantic segmentation, visual entailment and paraphrasing. ICCV '15, page 10–18, USA, 2015. IEEE Computer Society.

[7] Fereshteh Sadeghi, Santosh K. Kumar Divvala, and Ali Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[8] Jonathan Gordon and Benjamin Van Durme. Reporting Bias and Knowledge Extraction. In *Automated Knowledge Base Construction (AKBC) 2013: The 3rd Workshop on Knowledge Extraction, at CIKM*, 2013.

[9] Benjamin Van Durme. *Extracting Implicit Knowledge from Text.* PhD thesis, University of Rochester, Rochester, NY 14627, 2010.

[10] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. KR'12, page 552–561. AAAI Press, 2012.

[11] Bruce D'Ambrosio. Qualitative process theory using linguistic variables. ., 01 1989.

[12] Douglas Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38, 12 1998.

[13] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015.

[14] 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*, WWW '07, page 697–706, New York, NY, USA, 2007. Association for Computing Machinery.

[15] Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning, 2019.

[16] R. Speer and Catherine Havasi. Conceptnet 5: A large semantic network for relational knowledge. In *The People's Web Meets NLP*, 2013.

[17] Niket Tandon, Gerard de Melo, Fabian Suchanek, and Gerhard Weikum. Webchild: harvesting and organizing commonsense knowledge from the web. pages 523–532, 02 2014.

[18] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015.

[19] Valerio Basile, Elena Cabrio, and Fabien Gandon. Building a general knowledge base of physical objects for robots. In *13th International Conference, ESWC 2016*, The Semantic Web. Latest Advances and New Domains, Helaklion, Greece, May 2016. Poster paper.

[20] Claudia Schon Valerio Basile, Elena Cabrio. Knews: Using logical and lexical semantics to extract knowledge from natural language. *ecai*, 2016.

[21] Ivana Balazevic, Carl Allen, and Timothy Hospedales. Tucker: Tensor factorization for knowledge graph completion. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[22] Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*, May 2015.

[23] Q. Wang, Z. Mao, B. Wang, and L. Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017.

[24] Zhenjie Zhao, Evangelos Papalexakis, and Xiaojuan Ma. Learning Physical Common Sense as Knowledge Graph Completion via BERT Data Augmentation and Constrained Tucker Factorization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3293–3298, Online, November 2020. Association for Computational Linguistics.

[25] Ken McRae, George Cree, Mark Seidenberg, and Chris Mcnorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37:547–59, 12 2005.

[26] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[27] Maxwell Forbes and Yejin Choi. Verb physics: Relative physical knowledge of actions and objects, 2017.

[28] Frank F. Xu, Bill Yuchen Lin, and Kenny Q. Zhu. Automatic extraction of commonsense locatednear knowledge, 2018.

[29] Joe Davison, Joshua Feldman, and Alexander Rush. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

*(EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China, November 2019. Association for Computational Linguistics.

[30] 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, Doha, Qatar, October 2014. Association for Computational Linguistics.

[31] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[32] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[34] Ian Porada, Kaheer Suleman, and Jackie Chi Kit Cheung. Can a gorilla ride a camel? learning semantic plausibility from text. *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, 2019.

[35] Su Wang, Greg Durrett, and Katrin Erk. Modeling semantic plausibility by injecting world knowledge, 2018.

[36] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019.

[37] Pranav Goel, Shi Feng, and Jordan Boyd-Graber. How pre-trained word representations capture commonsense physical comparisons. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 130–135, Hong Kong, China, November 2019. Association for Computational Linguistics.

[38] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics.