PIQA: Reasoning about Physical Commonsense in Natural Language

Yonatan Bisk[1,2,3,4] **Rowan Zellers**[1,4] **Ronan Le Bras**[1] **Jianfeng Gao**[2] **Yejin Choi**[1,4]
[1]Allen Institute for Artificial Intelligence [2]Microsoft Research AI [3]Carnegie Mellon University
[4]Paul G. Allen School for Computer Science and Engineering, University of Washington
http://yonatanbisk.com/piqa

Abstract

To apply eyeshadow without a brush, should I use a *cotton swab or a toothpick*? Questions requiring this kind of **physical commonsense** pose a challenge to today's natural language understanding systems. While recent pretrained models (such as BERT) have made progress on question answering over more *abstract* domains – such as news articles and encyclopedia entries, where text is plentiful – in more *physical* domains, text is inherently limited due to reporting bias. Can AI systems learn to reliably answer physical commonsense questions without experiencing the physical world?

In this paper, we introduce the task of physical commonsense reasoning and a corresponding benchmark dataset **Physical Interaction: Question Answering** or **PIQA**. Though humans find the dataset easy (95% accuracy), large pretrained models struggle (∼77%). We provide analysis about the dimensions of knowledge that existing models lack, which offers significant opportunities for future research.

Introduction

Before children learn language, they already start forming categories and concepts based on the physical properties of objects around them (Hespos and Spelke 2004). This model of the world grows richer as they learn to speak, but already captures *physical commonsense* knowledge about everyday objects: their physical properties, affordances, and how they can be manipulated. This knowledge is critical for day-to-day human life, including tasks such as problem solving (what can I use as a pillow when camping?) and expressing needs and desires (bring me a hard*er* pillow). Likewise, we hypothesize that modeling physical commonsense knowledge is a major challenge on the road to true AI-completeness, including robots that interact with the world and understand natural language.

Much of physical commonsense can be expressed in language, as the versatility of everyday objects and common concepts eludes other label schemes. However, due to issues of reporting bias, these commonsense properties - facts like 'it is a bad idea to apply eyeshadow with a toothpick' are rarely directly reported. Although much recent progress has been made in Natural Language Processing through a shift towards large-scale pretrained representations from unlabeled text (Radford et al. 2018; Devlin et al. 2019; Liu et al. 2019), the bulk of the success of this paradigm has been on core *abstract* tasks and domains. State-of-the-art models can reliably answer questions given an encyclopedia article (Rajpurkar et al. 2016) or recognize named entities (Tjong Kim Sang and De Meulder 2003), but it is not clear whether they can robustly answer questions that require physical commonsense knowledge.

To study this question and begin bridging the representational gap, we introduce **Physical Interaction: Question Answering**, or **PIQA** to evaluate language representations on their knowledge of physical commonsense. We focus on everyday situations with a preference for atypical solutions. Our dataset is inspired by instructables.com, which provides users with instructions on how to build, craft,

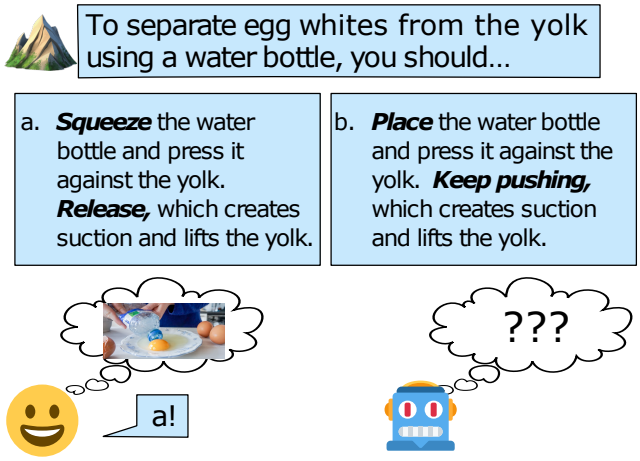

Figure 1: **PIQA**: Given a physical **goal** expressed in natural language, like 'to separate egg whites...,' a model must choose the most sensible **solution**. Our dataset tests the ability of natural language understanding models to link text to a robust intuitive-physics model of the world. Here, humans easily pick answer **a)** because separating the egg requires *pulling* the yolk out, while machines are easily fooled.

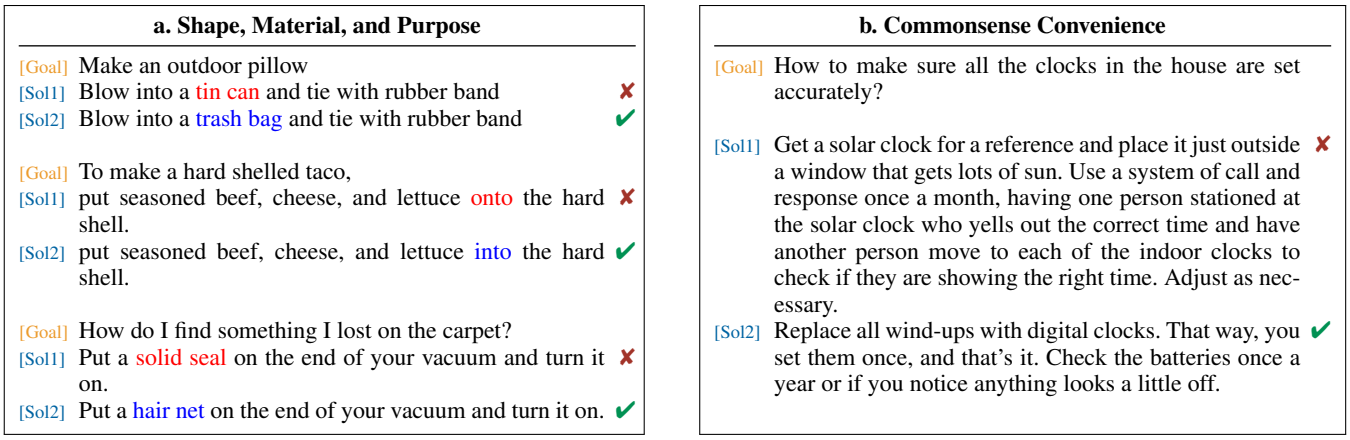

Figure 2: **PIQA** covers a broad array of phenomena. Above are two categories of example QA pairs. **Left** are examples that require knowledge of basic properties of the objects (flexibility, curvature, and being porous), while on the **Right** both answers may be technically correct but one is more convenient and preferable.

bake, or manipulate objects using everyday materials. We asked annotators to provide semantic perturbations or alternative approaches which are otherwise syntactically and topically similar to ensure physical knowledge is targeted. The dataset is further cleaned of basic artifacts using the `AFLite` algorithm introduced in (Sakaguchi et al. 2020; Sap et al. 2019) which is an improvement on adversarial filtering (Zellers et al. 2018; Zellers et al. 2019b).

Throughout this work we first detail the construction of our new benchmark for physical commonsense. Second, we show that popular approaches to large-scale language pretraining, while highly successful on many *abstract* tasks, fall short when a physical model of the world is required. Finally, our goal is to elicit further research into building language representations that capture details of the real world. To these ends, we perform error and corpora analyses to provide insights for future work.

Dataset

We introduce a new dataset, **PIQA**, for benchmarking progress in physical commonsense understanding. The underlying task is multiple choice question answering: given a question q and two possible solutions s_1, s_2, a model or a human must choose the most appropriate solution, of which exactly one is correct. We collect data with how-to instructions as a scaffold, and use state-of-the-art approaches for handling spurious biases, which we will discuss below.

Instructables as a source of physical commonsense

Our goal is to construct a resource that requires concrete physical reasoning. To achieve this, we provide a prompt to the annotators derived from instructables.com. The instructables website is a crowdsourced collection of instructions for doing everything from cooking to car repair. In most cases, users provide images or videos detailing each step and a list of tools that will be required. Most goals are simultaneously rare and unsurprising. While an annotator is unlikely to have built a UV-Flourescent steampunk lamp or made a backpack out of duct tape, it is not surprising that someone interested in home crafting would create these, nor will the tools and materials be unfamiliar to the average person. Using these examples as the seed for their annotation, helps remind annotators about the less prototypical uses of everyday objects. Second, and equally important, is that instructions build on one another. This means that any QA pair inspired by an instructable is more likely to explicitly state assumptions about what preconditions need to be met to start the task and what postconditions define success.

Collecting data through goal-solution pairs

Unlike traditional QA tasks, we define our dataset in terms of Goal and Solution pairs (see Figure 2 for example Goal-Solution pairs and types of physical reasoning). The Goal in most cases can be viewed as indicating a post-condition and the solutions indicate the procedure for accomplishing this. The more detailed the goal, the easier it is for annotators to write both correct and incorrect solutions. As noted above, the second component of our annotation design is reminding people to think creatively. We initially experimented with asking annotators for (task, tool) pairs via unconstrained prompts, but found that reporting bias swamped the dataset. In particular, when thinking about how to achieve a goal, people most often are drawn to prototypical solutions and look for tools in the kitchen (e.g. forks and knives) or the garage (e.g. hammers and drills). They rarely considered the literal hundreds of other everyday objects that might be in their own homes (e.g. sidewalk chalk, shower curtains, etc).

To address this, and flatten the distribution of referenced objects (see Figure 5), we prompt the annotations with links to instructables. Specifically, annotators were asked to glance at the instructions of an instructable and pull out or have it inspire them to construct two component tasks. They would then articulate the goal (often centered on atypical materials) and how to achieve it. In addition, we asked them to provide a permutation to their own solution which makes it invalid, often subtly (Figure 3). To further assist diversity

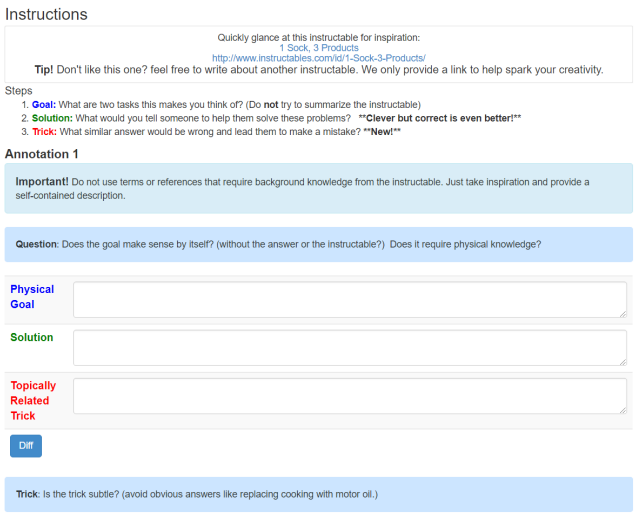

Figure 3: In the HIT design the instructable provides inspiration to think out-of-the-box (*1 Sock, 3 Products*) and annotators are asked for 1. a *physical* goal, 2. a valid solution, and 3. a trick. The trick should sound reasonable, but be wrong often due to a subtle misunderstanding of preconditions or physics. Additional HITs (not shown) were run for qualification prior to this stage and validation afterwards.[2]

we seed annotators with instructables drawn from six categories (costume, outside, craft, home, food, and workshop). We asked that two examples be drawn per instructable to encourage one of them to come later in the process and require precise articulation of pre-conditions.

During validation, examples with low agreement were removed from the data. This often meant that correct examples were removed that required expert level knowledge of a domain (e.g. special woodworking terminology) which should not fall under the umbrella of "commonsense." Because, we focus on human generated tricks, annotators were free to come up with clever ways to hide deception. Often, this meant making very subtle changes to the solution to render it incorrect. In these cases, the two solutions may differ by as little as one word. We found that annotations used both simple linguistic tricks (e.g. negation and numerical changes) and often swapped a key action or item for another that was topically similar but not helpful for completing the given goal. For this reason, our interface also includes a `diff` button which highlights where the solutions differ. This improved annotator accuracy and speed substantially. Annotator pay averaged $>$ 15$/hr according to both self-reporting on `turkerview.com` and our timing calculations.

[2]In addition to this design, we also include a qualification HIT which contained well constructed and underspecified (goal, solution) pairs. Annotators had to successfully (>80%) identify which were well formed to participate in the main HIT. Data was collected in batches of several thousand triples and validated by other annotators for correctness. Users will low agreement were de-qualed.

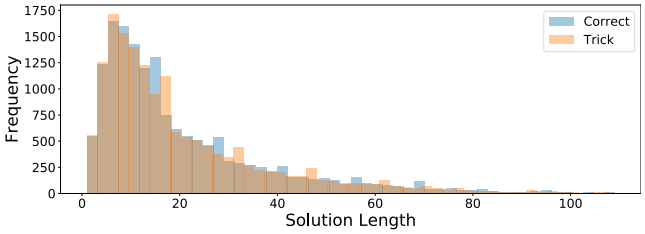

Figure 4: Sentence length distributions for both correct solutions and tricks are nearly identical across the training set.

Statistics

In total our dataset is comprised of over 16,000 training QA pairs with an additional ∼2K and ∼3k held out for development and testing, respectively. Our goals, as tokenized by Spacy,[3] average 7.8 words and both correct and incorrect solutions average 21.3 words. In total, this leads to over 3.7 million lexical tokens in the training data.

Figure 4 shows a plot of the correct and incorrect sequence lengths (as tokenized by the GPT BPE tokenizer), with the longest 1% of the data removed. While there are minor differences, the two distributions are nearly identical.

We also analyzed the overlap in the vocabulary and find that in all cases (noun, verb, adjective, and adverb) we see at least an 85% overlap between words used in correct and incorrect solutions. In total we have 6,881 unique nouns, 2,493 verbs, 2,263 adjectives, and 604 adverbs in the training data.. The most common of each are plotted in Figure 5 alongside their cumulative distributions. Again, this helps verify that the dataset revolves very heavily around physical phenomena, properties, and manipulations. For example, the top adjectives include state (*dry, clean, hot*) and shape (*small, sharp, flat*); adverbs include temporal conditions (*then, when*) and manner (*quickly, carefully, completely*). These properties often differentiate correct from incorrect answers, as shown in examples throughout the paper. We also color words according to their concreteness score (Brysbaert, Warriner, and Kuperman 2014), though many "abstract" words have concrete realizations in our dataset.

Removing Annotation Artifacts

As noted previously, we use `AFLite` (Sakaguchi et al. 2020) to remove stylistic artifacts and trivial examples from the data, which have been shown to artificially inflate model performance on previous NLI benchmarks (Poliak et al. 2018; Gururangan et al. 2018). The `AFLite` algorithm performs a systematic data bias reduction: it discards instances whose given feature representations are collectively highly indicative of the target label. In practice, we use 5,000 examples from the original dataset to fine-tune BERT-Large for this task and compute the corresponding embeddings of all remaining instances. `AFLite` uses an ensemble of linear classifiers trained on random subsets of the data to determine whether these pre-computed embeddings are strong

[3]`https://spacy.io` – all data was collected in English.

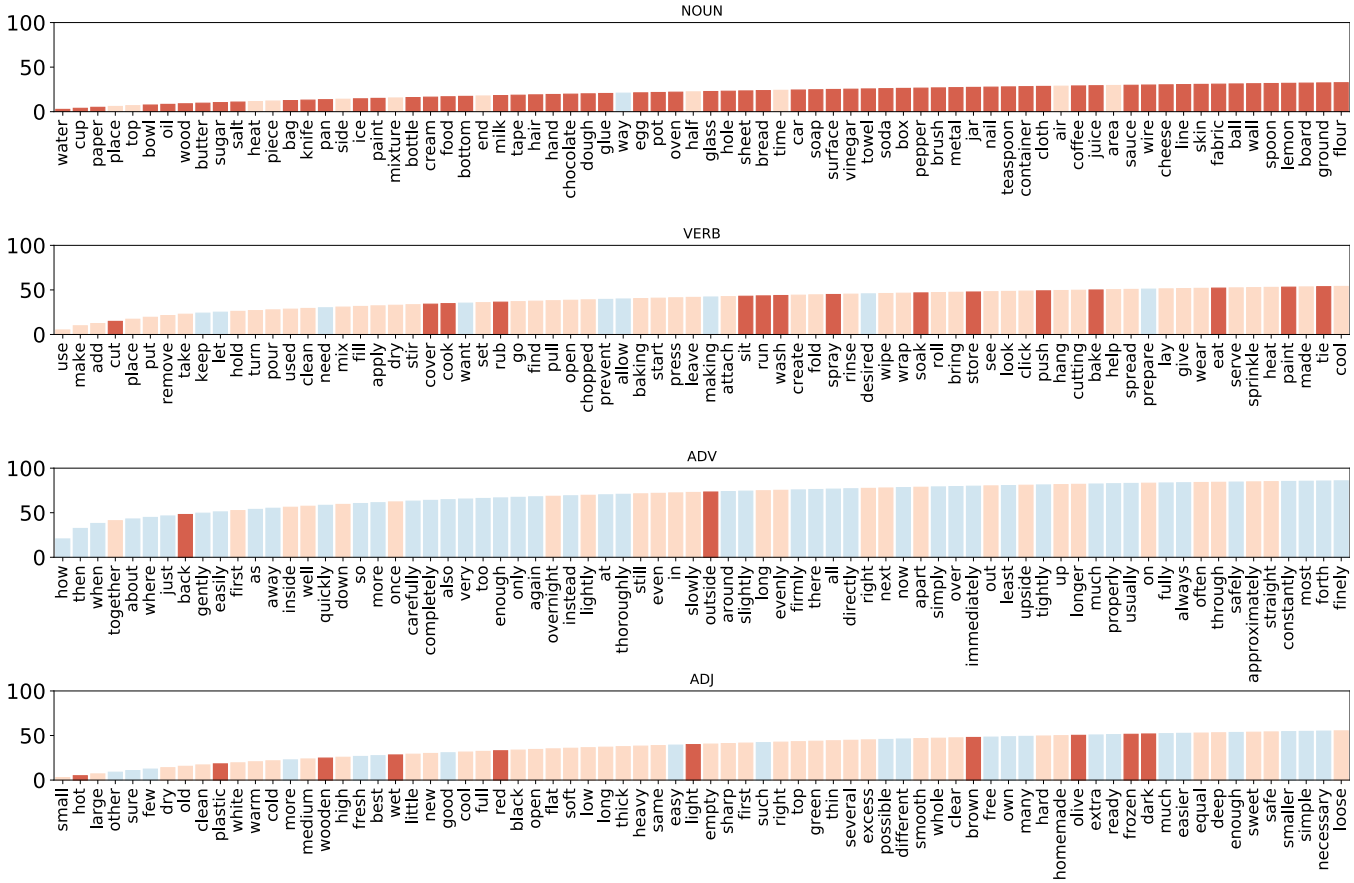

Figure 5: Here we show the frequency distributions for the top seventy-five words tagged by Spacy as noun, verb, adverb or adjective. We see that the vast majority of concepts focus on physical properties (*e.g. small, hot, plastic, wooden*) and how objects can be manipulated (*e.g. cut, cover, soak, push*). Additionally, we see strongly zipfian behavior in all tags but the adverbs. Words are colored by the average concreteness scores presented by (Brysbaert, Warriner, and Kuperman 2014).

indicators of the correct answer option. Instead of having to specifically identify the possible sources of biases, this approach enables unsupervised data bias reduction by relying on state-of-the-art methods to uncover undesirable annotation artifacts. For more information about `AFLite`, please refer to (Sakaguchi et al. 2020).

Experiments

In this section, we test the performance of state-of-the-art natural language understanding models on our dataset, **PIQA**. In particular, we consider the following three large-scale pretrained transformer models:

a. **GPT** (Radford et al. 2018) is a model that processes text left-to-right, and was pretrained using a language modeling objective. We use the original 124M parameter GPT model.

b. **BERT** (Devlin et al. 2019) is a model that process text bidirectionally, and thus was pretrained using a special masked language modeling objective. We use BERT-Large with 340M parameters.

c. **RoBERTa** (Liu et al. 2019) is a version of the BERT model that was made to be significantly more robust through pretraining on more data and careful validation of the pretraining hyperparameters. We use RoBERTa-Large, which has 355M parameters.

We follow standard best practices in adapting these models for two-way classification. We consider the two solution choices independently: for each choice, the model is provided the goal, the solution choice, and a special `[CLS]` token. At the final layer of the transformer, we extract the hidden states corresponding to the positions of each `[CLS]` token. We apply a linear transformation to each hidden state and apply a softmax over the two options: this approximates the probability that the correct solution is option A or B. During finetuning, we train the model using a cross-entropy loss over the two options. For GPT, we follow the original implementation and include an additional language modeling loss, which improved training stability.

Generally, we found that finetuning was often unstable with some hyperparameter configurations leading to validation performance around chance, particularly for BERT. We follow best practices in using a grid search over learning rates, batch sizes, and the number of training epochs for each model, and report the best-scoring configuration as was found on the validation set. For all models and experiments,

Model	Size	Accuracy (%) Validation	Test
Random Chance		50.0	50.0
Majority Class		50.5	50.4
OpenAI GPT	124M	70.9	69.2
Google BERT	340M	67.1	66.8
FAIR RoBERTa	355M	79.2	77.1
Human		94.9	

Table 1: Results of state-of-the-art natural language understanding models on **PIQA**, compared with human performance. The results show a significant gap between model and human performance, of roughly 20 absolute points.

we used the `transformers` library and truncated examples at 150 tokens, which affects 1% of the data.

Manual inspection of the development errors show that some "mistakes" are actually correct but required a web-search to verify. Human performance was calculated by a majority vote. Annotators were chosen to participate that achieved ≥90% on the qualification HIT from before. It is therefore, completely reasonable that automated methods trained on large web crawls may eventually surpass human performance here. Human evaluation was performed on development data, and the train, development, and test folds were automatically produced by `AFLite`.

Results

We present our results in Table 1. As the dataset was constructed to be adversarial to BERT, it is not surprising that it performs the worst of three models despite generally outperforming GPT on most other benchmarks. Comparing GPT and RoBERTa we see that despite more training data, a larger vocabulary, twice the number of parameters and careful construction of robust training, there is only a 8pt performance gain and RoBERTa still falls roughly 18 points short of human performance on this task. As noted throughout, exploring this gap is precisely the purpose for **PIQA** existing and which facets of the dataset fool RoBERTa is the focus of the remainder of this paper.

Analysis

In this section, we unpack the results of state-of-the-art models on **PIQA**. In particular, we take a look at the errors made by the top-performing model RoBERTa, as a view towards the physical commonsense knowledge that can be learned through language alone.

PIQA as a diagnostic for physical understanding

The setup of **PIQA** allows us to use it to probe the inner workings of deep pretrained language models, and to determine the extent of their physical knowledge. In this way, our dataset can augment prior work on studying to what extent models such as BERT understand syntax (Goldberg 2019). However, while syntax is a well studied problem within linguistics, physical commonsense does not have as rich a literature to borrow from, making its dimensions challenging to pin down.

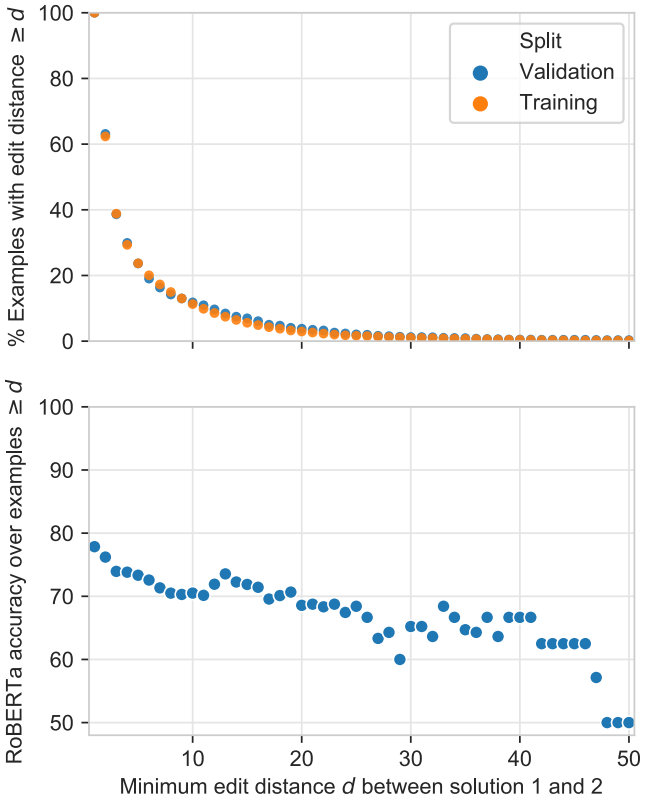

Figure 6: Breaking down **PIQA** by edit distance between solution choices. **Top**: Cumulative histogram of examples in the validation and training sets, in terms of minimum edit distance d between the two solution choices. The majority of the dataset consists of small tweaks between the two solution pairs; nevertheless, this is enough to confuse state-of-the-art NLP models. **Bottom**: RoBERTa accuracy over validation examples with a minimum edit distance of d. Dataset difficulty increases somewhat as the two solution pairs are allowed to drift further apart.

Simple concepts. Understanding the physical world requires a deep understanding of simple concepts, such as "water" or "ketchup," and their affordances and interactions with respect to other concepts. Though our dataset covers *interactions* between and with common objects, we can analyze the space of concepts in the dataset by performing a string alignment between solution pairs. Two solution choices that differ by editing a single phrase must by definition test the commonsense understanding of that phrase.

In Figure 6 we show the distribution of the edit distance between solution choices. We compute edit distance over tokenized and lowercased strings with punctuation removed. We use a cost of 1 for edits, insertions, and deletions. Most of the dataset covers simple edits between the two solution choices: roughly 60% of the dataset in both validation and training involves a 1-2 word edit between solutions. In the bottom of Figure 6, we show that the dataset complexity

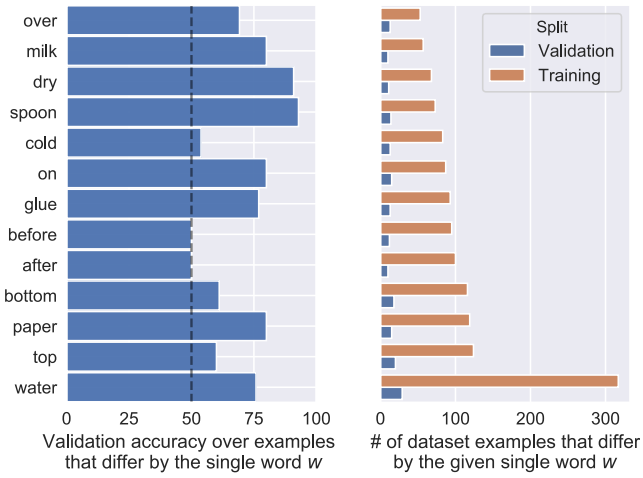

Figure 7: Common concepts as a window to RoBERTa's understanding of the physical world. We consider validation examples $(\boldsymbol{q}, \boldsymbol{s}_1, \boldsymbol{s}_2)$ wherein $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ differ from each other by a given word w. **Left**, we show the validation accuracy for common words w, while the number of dataset examples are shown **right.** Though certain concepts such as *water* occur quite frequently, RoBERTa nevertheless finds those concepts difficult, with 75% accuracy. Additionally, on common relations such as 'cold', 'on', 'before', and 'after' RoBERTa performs roughly at chance.

generally increases with the edit distance between the solution pairs. Nevertheless, the head of the distribution represents a space that is simple to study.

Single-word edits. In Figure 7, we plot the accuracy of RoBERTa among dataset examples that differ by a single word. More formally, we consider examples $(\boldsymbol{q}, \boldsymbol{s}_1, \boldsymbol{s}_2)$ whereby moving from $\boldsymbol{s}_1$ to $\boldsymbol{s}_2$, or vice versa, requires editing a given word w.[4] We show examples of words w that occur frequently in both the training and validation splits of the dataset, which allows RoBERTa to refine representations of these concepts during training and gives us a large enough sample size to reliably estimate model performance.

As shown, RoBERTa struggles to understand certain highly flexible relations. In particular, Figure 7 highlights the difficulty of correctly answering questions that differ by the words 'before,' 'after', 'top', and 'bottom': RoBERTa performs nearly at chance when encountering these.

Interestingly, the concepts shown in Figure 7 suggest that RoBERTa also struggles to understand many common, more versatile, physical concepts. Though there are 300 training examples wherein the solution choices $\boldsymbol{s}_1, \boldsymbol{s}_2$ differ by the word 'water.' RoBERTa performs worse than average on these replacements. On the other hand, RoBERTa does much better at certain nouns, such as 'spoon.'

Common replacements in PIQA. We dig into this further in Figure 8, where we showcase the most common replacements for three examples: 'water,' 'spoon,' and 'freeze.' While 'water' is prevalent in the training set, it is also highly versatile. One can try to substitute it with a variety of different household items, such as 'milk' or 'alcohol,' often to disastrous effects. However, 'spoons' have fewer challenging properties. A spoon cannot generally be substituted with a utensil that is sharp or has prongs, such as a fork, a knife, or a toothpick. RoBERTa obtains high accuracy on 'spoon' examples, which suggests that it might understand this simple affordance, but does not capture the long tail of affordances associated with 'water.'

[4]We additionally allow for an additional insertion; this helps to capture simple phrases like going from 'water' to 'olive oil.' Nevertheless, these multiword expressions tend to be less common, which is why we omit them in Figure 7.

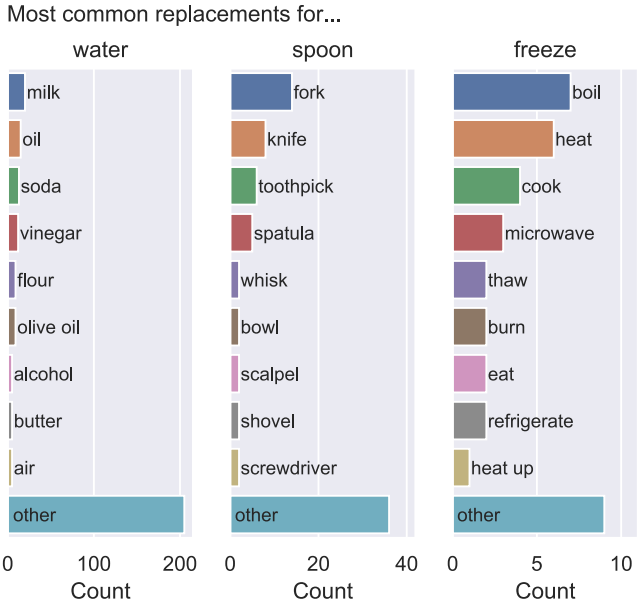

Figure 8: The most common replacements for three selected words: 'water,' 'spoon,' and 'freeze.' These cover several key dimensions: 'water' is a broad noun with many properties and affordances, whereas 'spoons' are much narrower in scope. Perhaps as a result, RoBERTa performs much butter at examples where 'spoon' is the pivot word (90%) versus 'water' (75%). Freeze has an accuracy of 66% on the validation set, and shows that verbs are challenging as well.

Qualitative results

Our analysis thus far has been on simple-to-analyze single word expressions, where we have shown that the state-of-the-art language model, RoBERTa, struggles at a nuanced understanding of key commonsense concepts, such as relations. To further probe the knowledge gap of these strong models, we present qualitative examples in Figure 9. The examples are broadly representative of larger patterns: RoBERTa can recognize clearly ridiculous generations (Figure 9, top left) and understands differences between some commonsense concepts (bottom left). It's important to note, that in both cases the correct answer is prototypical and something we might expect the models to have seen before.

However, it struggles to tell the difference between sub-

Correct examples

[Goal] Best way to pierce ears.
[Sol1] It is best to go to a professional to get your ear pierced to avoid medical problems later. ✔
[Sol2] The best way to pierce your ears would be to insert a needle half inch thick into the spot you want pierced. ✘

[Goal] How do you reduce wear and tear on the nonstick finish of muffin pans?
[Sol1] Make sure you use ***paper liners*** to protect the nonstick finish when baking muffins and cupcakes in muffin pans. ✔
[Sol2] Make sure you use ***grease and flour*** to protect the nonstick finish when baking muffins and cupcakes in muffin pans. ✘

Incorrect examples

[Goal] How can I quickly and easily remove strawberry stems?
[Sol1] Take a straw and from the ***top*** of the strawberry push the straw through the center of the strawberry until the stem pops off. ✘
[Sol2] Take a straw and from the ***bottom*** of the strawberry push the straw through the center of the strawberry until the stem pops off. ✔

[Goal] how to add feet to a coaster.
[Sol1] cut four slices from a glue stick, and attatch to the coaster with glue. ✔
[Sol2] place a board under the coaster, and secure with zip ties and a glue gun. ✘

Figure 9: Qualitative analysis of RoBERTa's predictions with. **Left**: Two examples that RoBERTa gets right. **Right**: two examples that RoBERTa gets incorrect. Short phrases that differ between solution 1 and solution 2 are shown in **bold and italics**.

tle relations such as top and bottom (top right of Figure 9). Moreover, it struggles with identifying non-prototypical situations (bottom right). Though using a gluestick as feet for a coaster is uncommon, to a human familiar with these concepts we can visualize the action and its result to verify that the goal has been achieved. Overall, these examples suggest that physical understanding – particularly involving novel combinations of common objects – challenges models that were pretrained on text only.

Related Work

Physical understanding is broad domain that touches on everything from scientific knowledge (Schoenick et al. 2016) to the interactive acquisition of knowledge by embodied agents (Thomason et al. 2016). To this end, work related to the goals of our benchmark span the NLP, Computer Vision and Robotics communities.

Language. Within NLP, in addition to large scale models, there has also been progress on reasoning about cause and effect effects/implications within these models (Bosselut et al. 2019), extracting knowledge from them (Petroni et al. 2019), and investigating where large scale language models fail to capture knowledge of tools and elided procedural knowledge in recipes (Bisk et al. 2019). The notion of procedural knowledge and instruction following is a more general related task within vision and robotics. From text alone, work has shown that much can be understood about the implied physical situations of verb usage (Forbes and Choi 2017) and relative sizes of objects (Elazar et al. 2019).

Vision. Physical knowledge can be discovered and evaluated within the visual world. Research has studied predicting visual relationships in images (Krishna et al. 2016) and as well as actions and their dependent objects (Yatskar, Zettlemoyer, and Farhadi 2016). Relatedly, the recent HAKE dataset (Li et al. 2019) specifically annotates which object/body-parts are essential to completing or defining an action. Image data also allows for studying the concreteness of nouns and provides a natural path forward for further investigation (Hessel, Mimno, and Lee 2018). Related to physical commonsense, research in *visual commonsense* has studied intuitive physics (Wu et al. 2017), cause-effect relationships (Mottaghi et al. 2016), and what can be reasonably inferred beyond a single image (Zellers et al. 2019a).

Robotics. Learning from interaction and intuitive physics (Agrawal et al. 2016) can also be encoded as priors when exploring the world (Byravan et al. 2018) and internal models of physics, shape, and material strength enable advances in tool usage (Toussaint et al. 2018) or construction (Nair, Balloch, and Chernova 2019). Key to our research aims in this work is helping to build language tools which capture enough physical knowledge to speed up the bootstrapping of robotic-language applications. Language tools should provide strong initial priors for learning (Tellex et al. 2011; Matuszek 2018) that are then refined through interaction and dialogue (Gao et al. 2016).

Conclusion

We have evaluated against large-scale pretrained models as they are in vogue as the de facto standard of progress within NLP, but are primarily interested in their performance and failings as a mechanism for advancing the position that learning about the world from language alone, is limiting. Future research, may "match" humans on our dataset by finding a large source of in-domain data and fine-tuning heavily, but this is very much *not the point*. Philosophically, knowledge should be learned from interaction with the world to eventually be communicated with language.

In this work we introduce the **Physical Interaction: Question Answering** or **PIQA** benchmark for evaluating and studying physical commonsense understanding in natural language models. We find the best available pretrained models lack an understanding of some of the most basic physical properties of the world around us. Our goal with **PIQA** is to provide insight and a benchmark for progress towards language representations that capture knowledge traditionally only seen or experienced, to enable the construction of language models useful beyond the NLP community.

Acknowledgements

We thank the anonymous reviewers for their insightful suggestions. This research was supported in part by NSF (IIS-1524371, IIS-1714566), DARPA under the CwC program through the ARO (W911NF-15-1-0543), DARPA under the MCS program through NIWC Pacific (N66001-19-2-4031), and the NSF-GRFP No. DGE-1256082. Computations on `beaker.org` were supported in part by Google Cloud.

References

Agrawal, P.; Nair, A.; Abbeel, P.; Malik, J.; and Levine, S. 2016. Learning to poke by poking: Experiential learning of intuitive physics. In *NeurIPS*.

Bisk, Y.; Buys, J.; Pichotta, K.; and Choi, Y. 2019. Benchmarking hierarchical script knowledge. In *NAACL-HLT*.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *ACL*.

Brysbaert, M.; Warriner, A. B.; and Kuperman, V. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods* (46):904–911.

Byravan, A.; Leeb, F.; Meier, F.; and Fox, D. 2018. Se3-pose-nets: Structured deep dynamics models for visuomotor planning and control. In *ICRA*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Elazar, Y.; Mahabal, A.; Ramachandran, D.; Bedrax-Weiss, T.; and Roth, D. 2019. How large are lions? inducing distributions over quantitative attributes. In *ACL*.

Forbes, M., and Choi, Y. 2017. Verb physics: Relative physical knowledge of actions and objects. In *ACL*.

Gao, Q.; Doering, M.; Yang, S.; and Chai, J. 2016. Physical causality of action verbs in grounded language understanding. In *ACL*, 1814–1824.

Goldberg, Y. 2019. Assessing BERT's Syntactic Abilities. *arXiv:1901.05287*.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation artifacts in natural language inference data. In *NAACL-HLT*, 107–112.

Hespos, S. J., and Spelke, E. S. 2004. Conceptual precursors to language. *Nature* 430:453–456.

Hessel, J.; Mimno, D.; and Lee, L. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In *NAACL-HLT*, 2194–2205.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M.; and Fei-Fei, L. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *arXiv:1602.07332*.

Li, Y.-L.; Xu, L.; Huang, X.; Liu, X.; Ma, Z.; Chen, M.; Wang, S.; Fang, H.-S.; and Lu, C. 2019. Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Matuszek, C. 2018. Grounded Language Learning: Where Robotics and NLP Meet. In *IJCAI*, 5687 – 5691.

Mottaghi, R.; Rastegari, M.; Gupta, A.; and Farhadi, A. 2016. "what happens if..." learning to predict the effect of forces in images. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *ECCV*, 269–285.

Nair, L.; Balloch, J.; and Chernova, S. 2019. Tool Macgyvering: Tool Construction Using Geometric Reasoning. In *ICRA*.

Petroni, F.; Rocktschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language models as knowledge bases? In *EMNLP*.

Poliak, A.; Naradowsky, J.; Haldar, A.; Rudinger, R.; and Van Durme, B. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Joint Conference on Lexical and Computational Semantics (StarSem)*.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2383–2392.

Sakaguchi, K.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*.

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019. Socialiqa: Commonsense reasoning about social interactions. In *EMNLP*.

Schoenick, C.; Clark, P.; Tafjord, O.; Turney, P.; and Etzioni, O. 2016. Moving beyond the turing test with the allen ai science challenge. *Communications of the ACM*.

Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M. R.; Banerjee, A. G.; Teller, S.; and Roy, N. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the National Conference on Artificial Intelligence*.

Thomason, J.; Sinapov, J.; Svetlik, M.; Stone, P.; and Mooney, R. J. 2016. Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy". In *IJCAI*, 3477–3483.

Tjong Kim Sang, E. F., and De Meulder, F. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *NAACL*, 142–147.

Toussaint, M.; Allen, K. R.; Smith, K. A.; and Tenenbaum, J. B. 2018. Differentiable physics and stable modes for tool-use and manipulation planning. In *RSS*.

Wu, J.; Lu, E.; Kohli, P.; Freeman, B.; and Tenenbaum, J. 2017. Learning to see physics via visual de-animation. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *NeurIPS*.

Yatskar, M.; Zettlemoyer, L.; and Farhadi, A. 2016. Sit-

uation recognition: Visual semantic role labeling for image understanding. In *CVPR*.

Zellers, R.; Bisk, Y.; Schwartz, R.; and Choi, Y. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *EMNLP*.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019a. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019b. HellaSwag: Can a Machine Really Finish Your Sentence? In *ACL*.