

Do Nuclear Submarines Have Nuclear Captains? A Challenge Dataset for Commonsense Reasoning over Adjectives and Objects

James Mullenbach^{1*}, Jonathan Gordon^{2*}, Nanyun Peng³, Jonathan May³

¹ ASAPP Inc.

² Department of Computer Science, Vassar College

³ Information Sciences Institute, University of Southern California

jmullenbach@asapp.com, jgordon@vassar.edu

npeng, jonmay@isi.edu

Abstract

How do adjectives project from a noun to its parts? If a motorcycle is red, are its wheels red? Is a nuclear submarine’s captain nuclear? These questions are easy for humans to judge using our commonsense understanding of the world, but are difficult for computers. To attack this challenge, we crowd-source a set of human judgments that answer the English-language question “Given a whole described by an adjective, does the adjective also describe a given part?” We build strong baselines for this task with a classification approach. Our findings indicate that, despite the recent successes of large language models on tasks aimed to assess commonsense knowledge, these models do not greatly outperform simple word-level models based on pre-trained word embeddings. This provides evidence that the amount of commonsense knowledge encoded in these language models does not extend far beyond that already baked into the word embeddings. Our dataset will serve as a useful testbed for future research in commonsense reasoning, especially as it relates to adjectives and objects.

1 Introduction

We investigate the commonsense inference of the transitivity of an attribute of a whole object to its component parts. To illustrate this targeted reasoning by example, “is a sharp knife’s handle sharp?” The ability to perform commonsense inference of this type enables a more complete understanding of the physical world and therefore may find use in a variety of tasks in pragmatics and at the interface of vision and language. Consider generating a story in which a slow car goes to the shop to get a new part. If the new part is a windshield, the car remains slow, whereas if the new part is an engine,

the car may now be fast. This knowledge may also help a visual agent reason about unseen objects: it knows a brick house does not have a brick door without needing to see the door.

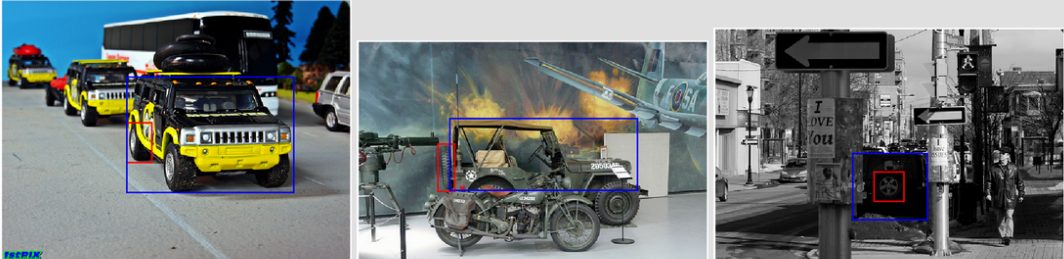
The past few years have seen a raft of data sets intended to test our ability to construct models with an understanding of commonsense knowledge. Standout examples are the Stanford Natural Language Inference (SNLI) and related Multi-Genre Natural Language Inference (MNLI) corpora (Bowman et al., 2015; Williams et al., 2018), the SemEval-2018 commonsense shared task (Ostermann et al., 2018), the Rochester Story Completion (ROCStories) corpus (Mostafazadeh et al., 2016), and the Situations with Adversarial Generations (SWAG) grounded inference corpus (Zellers et al., 2018). After their release, very large language models (LMs) were able to reach or surpass human-level performance on SNLI (Peters et al., 2018) and SWAG (Devlin et al., 2018).

However, researchers have found inadequacies in these datasets and the models trained on them. Despite the strong performance of recent systems on SNLI (e.g., Chen et al., 2017; Parikh et al., 2016), Glockner et al. (2018) show that by making trivial changes to the test set, these methods suffered. Further, Pavlick and Callison-Burch (2016) show that state-of-the-art models for natural language inference fail on a task requiring only reasoning over adjective-noun relations. Relatedly, in the shared task to predict sentence endings of ROCStories, Schwartz et al. (2017) show that by incorporating style features, with only the answer choices as input, it is possible to reach near state-of-the-art performance. These results point to implicit bias baked into the data sets.

Rudinger et al. (2017) demonstrate similar systematic and social bias in SNLI, attributing it to the fact that hypothesis sentences were written by crowd workers. The SWAG data set was specif-

*Research conducted while author was at USC/ISI

Each of the pictures below has a **jeep**, outlined with a blue box. Notice that each **jeep** also has a **tire**. The **tire** is outlined with a red box.



Now consider any **jeep**, not the particular ones pictured. If the **jeep** is **old**, which of the following is true?

- Because the **jeep** is **old**, the **tire** must also be **old**
- Because the **jeep** is **old**, the **tire** is probably **old**
- The **tire** might be **old**, but this is unrelated to the **jeep** being **old**
- Although the **jeep** is **old**, the **tire** is unlikely to be **old**
- Although the **jeep** is **old**, the **tire** cannot be **old**
- I don't know (check one of the boxes below)

I don't know what it means for a **jeep** to be **old** I don't know what one of these words means

I don't know what it means for a **tire** to be **old** I don't know what it means for a **jeep** to have a **tire**

Figure 1: Visual annotation interface, excluding overall instructions.

ically constructed in an adversarial way with this in mind, but may be disadvantaged by the fact that continuation sentences are generated by computers. This may lead to patterns that are hard to detect but can nevertheless be picked up by other language models. We avoid the issue of elicitation bias by first collecting candidates grounded in natural sources of text and images, and then gathering only scaled judgments from crowd workers, as was done by Zhang et al. (2017).

To understand how to build truly intelligent agents, we should strive to create datasets with as little exploitable bias as possible, and to further investigate the landscape of current performance. We contribute a dataset which provides a focused evaluation, based on a specific task in commonsense reasoning. Gathering and validating data from crowd workers, we evaluate a number of approaches to performing these inferences, a three-way lexical entailment problem. We find that simple word embedding-based models perform adequately, but beneath humans, on this task, with recent large LM approaches (Devlin et al., 2018; Radford et al., 2018) providing only slight improvement over the purely lexical approach.

2 Related Work

Other researchers have constructed datasets investigating similar ideas in commonsense reasoning. Forbes and Choi (2017) develop a dataset and methods for inferring physical commonsense knowledge from verb usage, showing it is possible to learn the physical implications of unseen verbs

from a small seed set. Zhang et al. (2017) create a large dataset for general commonsense inference in the form of premise-hypothesis pairs, equipped with ordinal labels ranging from “impossible” to “very likely”. We adopt much of their methodology but for a targeted subset of commonsense reasoning. The SemEval 2018 Task 10 on Capturing Discriminative Attributes (Krebs et al., 2018) describes a similar lexical reasoning task involving triplets of words, though it focuses on finding attributes that *distinguish* two concepts, while in our work the adjective may well apply to both part and whole.

Past work has also evaluated commonsense capabilities in neural models. Pavlick and Callison-Burch (2016) investigate the related problem of entailment in adjective-nouns, and show surprising negative results for neural NLI models. Wang et al. (2018) showed that models based on distributional semantics without explicit external knowledge perform poorly at predicting physical plausibility of actions.

Lucy and Gauthier (2017) investigate perceptual properties of distributional embeddings and suggest that part-whole properties like *has_legs* are well encoded by embeddings. This may help explain why the simple word-based MLP models perform well without other sources of context. Rei et al. (2018) introduce an effective neural architecture for learning word-embedding based models for graded lexical entailment. Prior work (Bulat et al., 2016; Fagarasan et al., 2015) utilizes embeddings to predict real-world perceptual proper-

Whole	Part	Adjective	Label	NLI premise	NLI hypothesis
armchair	arm	black	Probably	On the back of the president’s quaint black armchair there was emblazoned a half-sun, brilliant with its gilded rays.	The armchair’s arm is black.
vanity	mirror	white	Impossible	A door to a bathroom half open and a white vanity.	The vanity’s mirror is white.
bench	support	wooden	Unrelated	In front of me about five feet distance, stood a wooden bench.	The bench’s support is wooden.

Table 1: Example triples and retrieved premise sentences, with labels, used for training word embedding-based models and language model fine-tuning.

ties, and we expect an approach that leverages this will help solve this task, but we leave it to future work.

3 Candidate collection

We seek to annotate examples of (*whole*, *part*, *adjective*) triples with answers to the question: “Does an $\langle adjective \rangle$ $\langle whole \rangle$ have an $\langle adjective \rangle$ $\langle part \rangle$?” As a major part of our contribution, we provide an annotated dataset that is visually grounded, with relations mined from Visual Genome (Krishna et al., 2017) and Google Syntactic N-grams (Goldberg and Orwant, 2013). We provide an overview here, with details in Appendix A.

3.1 Part–whole relations

Visual Genome (VG) is a large dataset of images annotated with objects, their attributes, and the relations between them. We start by considering all relationships in the VG dataset where the predicate is an underspecified *has* relation. We count the number of images in which a pair of objects appear in a *has* relation, and keep only those pairs appearing in at least three distinct images.

3.2 Adjectives

We gather adjectives from both Google Syntactic N-grams and VG. From Syntactic N-grams, we count the occurrences of an adjective modifying a noun with the *amod* relation. We remove common non-attributive (e.g., *awake*) and non-descriptive (e.g., *first*) adjectives using manually constructed lexicons. Then, for each whole noun, we gather its five most common adjectival modifiers, as well as its five most common adjective attributes from

Visual Genome. Through pilot studies we observed that without further filtering, annotations were highly skewed towards non-entailment, thus we achieve a more balanced dataset by filtering out adjectives that are never observed attached to the part.

4 Collecting human annotations

We crowdsource annotations on Amazon Mechanical Turk (AMT) for each (*whole*, *part*, *adjective*) triple as follows:

4.1 Task overview

For each part–whole pair, we sample three random images from VG that contain the pair, and draw bounding boxes around both objects, provided by VG annotations. We present these to workers simply to provide context for the part–whole pair, since early tests showed that without visual cues workers often have trouble understanding the overall problem. Then, we ask a series of questions that each associates the pair with an adjective. To encourage the worker to imagine the prototypical version of the objects rather than the specific ones shown,¹ we use the template “Consider any $\langle whole \rangle$, not the particular ones pictured”. Specific questions have the form: “If the $\langle whole \rangle$ is $\langle adjective \rangle$, which of the following is true?” The answers describe whether it is “impossible”, “unlikely”, “unrelated”, “likely”, or “guaranteed” that the identified part is also described by the adjective. The answers use causal language to encourage “conditional plausibility” thinking, as described by Zhang et al. (2017). This also allows for the “unrelated” answer, which covers spu-

¹This is a necessary downside of displaying visual cues.

Label	Percentage
Guaranteed	44.2
Probably	19.8
Unrelated	23.7
Unlikely	5.6
Impossible	6.6

Table 2: Final label distribution

rious examples, such as a black guitar’s cord being black, where the cord is likely black, but not *as a result* of the guitar being black. We also give an option for the worker to mark that one of the pairwise relations is nonsensical.

4.2 Qualification task

After manually annotating some examples, and conducting two AMT pilot studies, we found a non-trivial margin between our own agreement and that of workers, as measured by the quadratic-weighted Cohen’s κ . To alleviate this, we followed Zhang et al. (2017) and conducted a pilot study to gather a pool of qualified workers. We launched a pilot task with two gold examples from each class on which our manual annotations agreed, and recruited 300 crowd workers to label them. By setting a κ threshold on agreement with the gold examples at 0.7, this resulted in 106 qualified workers, whom we requested to perform the rest of the annotations. We collected at least three annotations per triple. An example annotation interface is shown in Figure 1.

4.3 Filtering and statistics

From the total of 20,284 triples annotated, we filter out 4,040 (19.9%) that were reported to contain an invalid triple. We further remove instances without a majority vote from the workers. This results in a set of 13,684 triples with an inter-annotator agreement (quadratic-weighted Cohen’s κ) of 0.624. (For reference, Zhang et al. (2017) report $\kappa = 0.54$ for general commonsense inference.) The label distribution is shown in Table 2. The dataset has 728 unique *part* nouns, 873 unique *whole* nouns, and 553 unique adjectives.

5 Inference baselines

We now describe several basic approaches for solving these commonsense inference problems, which we intend as a baseline to be built upon by

future work. Formally, models answer the question: Given (1) a noun denoting a whole object that has (2), a component part also denoted by a noun, does (3), an adjective that describes 1 also describe 2? The data is first split into training, validation, and test sets consisting of 70%, 10%, and 20% of the data respectively. Model selection and tuning details are described in Appendix C.

5.1 Word embedding models

We approach the problem as categorical classification and train a multi-layer perceptron (MLP) model to classify inputs consisting of word embeddings for the whole, part, and adjective words. The MLP takes as input the concatenation of these three word embeddings, obtained from GloVe (Pennington et al., 2014), and applies a single hidden layer with ReLU activation before the final softmax layer which predicts the class label.

5.2 Adjective projection as NLI

As we want to evaluate strong yet simple pre-existing language understanding models on this task, we now describe a method for obtaining the direct prediction described above via conversion to a form suitable for inference in the style of the SNLI and MNLI datasets (Bowman et al., 2015; Williams et al., 2018), which consist of *premise* and *hypothesis* sentence pairs. We first form simple hypothesis sentences from the tuples using the fixed template “The $\langle whole \rangle$ ’s $\langle part \rangle$ is $\langle adjective \rangle$.” We then retrieve premise sentences that describe a $\langle whole \rangle$ $\langle adjective \rangle$. An example for (*bicycle*, *old*) is “He rode an old bicycle and brought fruits and vegetables home from Chinatown.” We retrieve context sentences from five resources: Project Gutenberg books², the Gigaword news corpus (Parker et al., 2011), SNLI, MNLI, and MSCOCO image captions (Lin et al., 2014); premise sentence selection is described fully in Appendix D and examples are shown in Table 1.

5.3 Fine-tuning language models

We apply transfer learning from two recently developed large contextualized LMs to this task. Both are state-of-the-art on NLI and commonsense tasks.

Specifically, we test OpenAI GPT (Radford et al., 2018), and BERT (Devlin et al., 2018).

²<https://gutenberg.org>

OpenAI GPT is a unidirectional model that predicts the next word, while BERT is bidirectional and predicts randomly missing words, as well as the next sentence. Both train on the BooksCorpus, with BERT additionally trained on English Wikipedia. Both models are fine-tuned to perform NLI by applying a linear layer to the model’s final output at one position of the input. The models are then trained in a multi-task way for the inference task and the language modeling objective(s), updating the whole network.

Method	Accuracy
Majority baseline	0.430
Majority-per-part baseline	0.485
GloVe embeddings	0.651
OpenAI GPT (Radford et al., 2018)	0.666
BERT (Devlin et al., 2018)	0.667
Human performance	0.785

Table 3: Test set results for multi-class prediction

Test set results for these methods are in Table 3. We also provide performance by two simple baselines, the first of which always predicts the majority class (“guaranteed”). To choose the second baseline, we evaluated choosing the majority class for the given whole, part, or adjective. Of these, predicting the majority-per-part had the best validation set performance, so we report that result on test.

We observe that the best model that operates on just word embeddings is within ≈ 0.02 of both language models in absolute accuracy points, and the best performing model still lags behind human accuracy³ by nearly 0.12 absolute points, suggesting work remains to be done on incorporating this variety of common sense into intelligent models.

6 Conclusion

Inspired by recent commonsense dataset construction efforts and the speed with which researchers develop highly performant models for them, we develop a dataset that evaluates a type of inference that is specific but that agents with commonsense should be able to solve. We show that state-of-the-art language models perform well, but that models using just pretrained word embeddings perform

³Measured using the triplet, not NLI, version of the data.

comparably, and both fall short of human accuracy. We release our dataset to provide a challenging commonsense reasoning task for the community.

Acknowledgements

Thanks to Sandeep Soni and Ian Stewart for helpful feedback on an initial draft. Thanks to Kevin Knight for initial inspiration and for the title. This work is supported by Contract W911NF-15-1-0543 with the US Defense Advanced Research Projects Agency (DARPA). Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. *A large annotated corpus for learning natural language inference*. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.
- Luana Bulat, Douwe Kiela, and Stephen Clark. 2016. Vision and feature norms: Improving automatic feature norm learning through cross-modal maps. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 579–88.
- Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. *Enhanced LSTM for natural language inference*. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–68. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- Luana Fagarasan, Eva Maria Vecchi, and Stephen Clark. 2015. From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 52–57.
- Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 266–276.

- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. [Breaking NLI systems with sentences that require simple lexical inferences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–5. Association for Computational Linguistics.
- Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 241–7.
- Alicia Krebs, Alessandro Lenci, and Denis Paperno. 2018. Semeval-2018 task 10: Capturing discriminative attributes. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 732–740.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–55. Springer.
- Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 76–85.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. [A corpus and cloze evaluation for deeper understanding of commonsense stories](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–49, San Diego, California. Association for Computational Linguistics.
- Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. [SemEval-2018 task 11: Machine comprehension using commonsense knowledge](#). In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 747–757, New Orleans, Louisiana. Association for Computational Linguistics.
- Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. [A decomposable attention model for natural language inference](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–55, Austin, Texas. Association for Computational Linguistics.
- Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword. *Linguistic Data Consortium*.
- Ellie Pavlick and Chris Callison-Burch. 2016. Most babies are little and most problems are huge: Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2164–2173.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–43.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–37. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. [Improving language understanding by generative pre-training](#). Preprint.
- Marek Rei, Daniela Gerz, and Ivan Vulic. 2018. Scoring lexical entailment with a supervised directional similarity network. *arXiv preprint arXiv:1805.09355*.
- Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. [Social Bias in Elicited Natural Language Inferences](#). In *The 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Workshop on Ethics in NLP*.
- Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. [The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.
- Robert Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–51.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.
- Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. In *Proceedings of the 2018 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), volume 2, pages 303–308.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. [SWAG: A large-scale adversarial dataset for grounded commonsense inference](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. [Ordinal common-sense inference](#). *Transactions of the Association for Computational Linguistics*, 5:379–395.