Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models

Bill Yuchen Lin **Seyeon Lee** **Rahul Khanna** **Xiang Ren**
{yuchen.lin,seyeonle,rahulkha,xiangren}@usc.edu
Department of Computer Science,
University of Southern California

Abstract

Recent works show that pre-trained language models (PTLMs), such as BERT, possess certain commonsense and factual knowledge. They suggest that it is promising to use PTLMs as "neural knowledge bases" via predicting masked words. Surprisingly, we find that this may not work for *numerical commonsense knowledge* (e.g., a bird usually has *two* legs). In this paper, we investigate whether and to what extent we can induce numerical commonsense knowledge from PTLMs as well as the robustness of this process. To study this, we introduce a novel probing task with a diagnostic dataset, NUMERSENSE[1], containing 13.6k masked-word-prediction probes (10.5k for fine-tuning and 3.1k for testing). Our analysis reveals that: (1) BERT and its stronger variant RoBERTa perform poorly on the diagnostic dataset prior to any fine-tuning; (2) fine-tuning with distant supervision brings some improvement; (3) the best supervised model still performs poorly as compared to human performance (54.06% vs. 96.3% in accuracy).

1 Introduction

Pre-trained language models (PTLMs), such as BERT (Devlin et al., 2019), have yielded state-of-the-art performance on many natural language processing tasks. Given PTLMs' cited ability to create general, yet useful text representations, an investigation of their ability to encode commonsense knowledge into representations is warranted—commonsense knowledge is often required to have a full understanding of language.

Recently there have been a few recent works that do investigate the inquiry of whether PTLMs possess *commonsense knowledge* (Petroni et al., 2019; Davison et al., 2019; Bouraoui et al., 2020). Overall, these prior studies suggest that PTLMs are creating text representations that often have commonsense knowledge encoded in them. We, however, find it surprising that when posed with a similar reasoning-based masked-word-prediction task, PTLMs perform poorly in recalling the required *numerical commonsense knowledge* (see Figure 1).

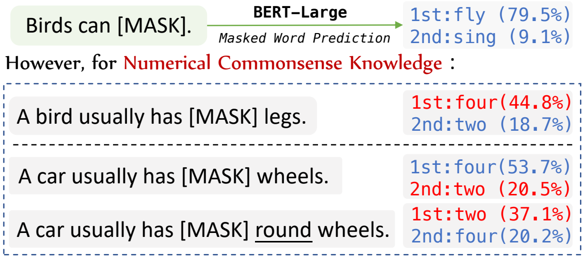

Figure 1: **Top:** PTLMs often cannot solve masked language modeling tasks needing ***numerical commonsense knowledge***, hence our title. **Bottom:** Even when PTLMs seemingly succeed, they fail to stay consistent under small perturbations.

Therefore, in this paper, our goal is to study whether PTLMs capture numerical commonsense knowledge, i.e., commonsense knowledge that provides an understanding of the numeric relation between entities. We propose measuring this capability via a masked-word-prediction based probing task, where, the ranking of numeric words by what the model believes most probably fills the mask would expose the capabilities of PTLMs to capture *numeric commonsense knowledge*. For example, the masked position in the sentence "*A bird usually has* [MASK] *legs*." is best filled by the number "two" when considering only numerical words.

Around this concept, we built a carefully crafted dataset, NUMERSENSE, of 3,145 probes that covers questions from 8 different categories such as everyday objects, biology, geometry, etc. In our initial experiments, we find PTLMs to be brittle against adversarial attacks. As shown in the bottom section of Figure 1, BERT initially correctly predicts the masked word to be "four", but it changes its top result to "two" in the slightly perturbed second

[1] https://inklab.usc.edu/NumerSense/

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6862–6868,
November 16–20, 2020. ©2020 Association for Computational Linguistics

sentence (a simple insertion of the word ‘round’). Thus, we intentionally included adversarial examples in the probes to test the robustness.

We evaluate PTLMs in two settings (Section 3): (1) a zero-shot setting, meaning no probes from our dataset were used to fine-tune the models before evaluation; (2) a distant supervision setting, where models were fine-tuned on examples from related commonsense reasoning datasets before being evaluated on ours. Our findings reveal that PTLMs are still much worse than humans on the task, although fine-tuning with distant supervision can help. We also provide some cursory analysis on why PTLMs perhaps perform so poorly, pointing to interesting future research. We also hope our work can benefit future works in: 1) improving PTLMs' abilities to faithfully capture (numerical) commonsense, 2) populating numerical facts in current commonsense knowledge bases, and 3) open-domain QA —“Q: *How many legs do ants have?” “A: Six!*”

2 The NumerSense Probing Task

We introduce our numerical commonsense reasoning probing task, as well as the creation process of the namesake dataset, NumerSense. Then, we provide a breakdown of what types of knowledge are covered by the probes and finally include additional high-quality distant supervision to test if fine-tuning can improve performance.

2.1 Task Formulation

We essentially probe PTLMs with the distribution of words a PTLM thinks could fill the masked position, by ranking their softmax scores (greatest to least). If the ranking demonstrates numerical commonsense knowledge—the highest ranked *number word* (e.g., “one”, “two”, and so on) is the correct answer—then that probe is successfully completed by the PTLM. The masked position in each probe is chosen such that a number word is an extremely probable way of filling in the blank.

2.2 Probing Data Collection

To build a suitable dataset for the proposed probing task, we make use of an existing corpus consisting of commonsense assertions, named *Open Mind Common Sense* (OMCS) (Singh et al., 2002). We first extracted the sentences from OMCS that had at least one of the following 12 *number words*: {“no”[2], “zero”, “one”, “two”, ..., “ten” }.

Category	Example
Objects(35.2%)	A bicycle has *two* tires.
Biology(13.5%)	Ants have *six* legs.
Geometry(11.7%)	A cube has *six* faces.
Unit(6.3%)	There are *seven* days in a week.
Math(7.3%)	I will be *ten* next year, as I am *nine* now.
Physics(5.7%)	Water will freeze at *zero* degrees centigrade.
Geography(2.9%)	The world contains *seven* continents.
Misc.(17.5%)	There are *no* princes in the United States.

Table 1: NumerSense examples of each category.

However, as to be expected, there were many noisy statements which were either 1) incorrect, 2) containing typos, or 3) having no numerical commonsense logic. We thus manually and pragmatically refined these sentences and did two rounds of vetting by different graduate students, from which we only kept the statements that were accepted by all annotators. After this strict filtration process, we ended up **1,131** cleaned statements for probing.

We did an initial test and observed that PTLMs can be brittle under a simple perturbation of inserting an adjective near the masked number word. Thus, in order to study the robustness of models in our proposed task, we also added adversarial examples to our dataset by adding adjectives before the noun involved in the numerical reasoning in each probe. The candidate adjectives are generated by querying relevant triples (e.g. <wheel, HasProperty, round> for the example in Fig. 1) in the commonsense knowledge graph, ConceptNet (Speer et al., 2017), and further selected or modified by human annotators to assure adversarial examples are still valid and natural. We finally have **3,145** testing probes for NumerSense as the diagnostic dataset.

We also manually annotated the category label for each instance so that we can better understand the covered topics and their percentage. We found 8 types of numerical commonsense knowledge ranging from tangible everyday objects (e.g., car, guitar, and table) to geometry (e.g., cube). Table 1 lists some concrete examples of each category.

2.3 Supervision for Fine-Tuning PTLMs

One may wonder if fine-tuning towards this task could improve the performance. In order to an-

[2] We include “no”, as there exists statements involving numerical commonsense knowledge, where “no” is used in place of zero, “There are **no** princes in the United States.”

	Core Probes			**+ Adversarial Examples**		
Models	**hit@1**	hit@2	hit@3	**hit@1**	hit@2	hit@3
GPT-2	29.86	50.88	67.49	24.73	44.21	62.30
BERT-Base	31.98	55.92	70.58	25.24	48.66	64.81
RoBERTa-Base	36.04	60.42	72.08	28.39	51.91	67.29
BERT-Large	37.63	62.01	76.77	27.18	52.89	70.22
RoBERTa-Large	**45.85**	66.70	80.04	**35.66**	58.52	74.44
Ft. BERT-L.	50.00	66.34	74.91	43.58	62.27	72.92
Ft. RoBERTa-L.	**54.06**	69.61	79.15	**47.52**	66.43	76.76
Human Bound	$89.7^{(\alpha)}$ / $96.3^{(\beta)}$			$88.3\ ^{(\alpha)}$ / $93.7\ ^{(\beta)}$		

Table 2: Results (%) of PTLMs on NumerSense. 'Ft.' stands for 'Fine-tuned.' The human performance is shown by closed testing (α='no external information') / open testing (β='Wikipedia is allowed').

swer this question, we further collected training sentences from the GenericsKB corpus (Bhakthavatsalam et al., 2020). The sentences in GenericsKB are generic commonsense statements that are extracted from Simple Wikipedia, Common Crawl within educational domains, ARC corpus, etc.

We collected these sentences by first obtaining a list of frequent nouns from various caption corpora such as MSCOCO (Lin et al., 2014) and VATEX (Wang et al., 2019). Then, we selected collected sentences contained at least one number word of interest and finally go through the same human annotator verification process as the test data. We ended up collecting **10,492** sentences for fine-tuning and believe these sentences, if used properly, can improve PTLMs' ability to recall the numerical commonsense knowledge.

2.4 Statistics of NumerSense

We show the distribution of the truth number words in the test data in Fig. 2. The average length of the sentence in training data is 11.1 and it is 8.9 in test data.

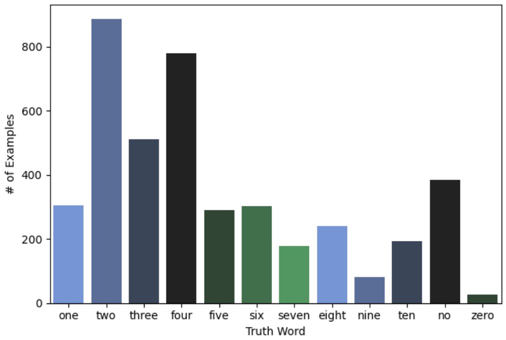

Figure 2: Truth number distribution of the test set.

3 Empirical Analysis

We introduce the set-up of the experiments and then present results from different PTLMs in both a zero-shot setting and a distantly supervised fine-tuned one. We will also provide some analysis on the robustness and biases in the various models, and finally a study of the performance of a state-of-the-art open-domain question-answering model.

3.1 Experiment Set-up

We run our experiments in two settings, *zero-shot inference* and additional supervision via fine-tuning. In the first setting, we probe PTLMs without any modifications, specifically we use BERT and RoBERTa with pre-trained masked-word-prediction heads.

In our second setting, we use our collected additional supervision dataset (Sec. 2.3) and mask the *number words* in each sentence. We then proceed to fine tune the models above on these masked sentences, before evaluating them on NumerSense.

3.2 Evaluation Metric and Human Bound

A masked-word-prediction head (either fine-tuned or not) produces a probability distribution over its whole vocabulary via a softax layer. As mentioned in Sec. 2.1, NumerSense is the task of using this probability distribution to rank all number words, and evaluating this ranking. To evaluate, we use hit@1/2/3 accuracy, which calculates the percentage of predictions where the correct number word is ranked in the top k number words.[3]

To estimate human performance on the task, we sampled 300 examples and asked two groups of three people to fill in the masked word, where one group had access to external information (**open-book** test) from the Web such as Wikipedia and the other did not (**closed-book** test). We take the majority label as the final human label.

3.3 Experimental results

We show our experimental results in Table 2. The first four lines are results from PTLMs in the zero-shot inference setting. We see that size matters, as there is a clear performance gain when the model sizes increases. Also, RoBERTa's results are consistently better than BERT's, which is probably because RoBERTa uses a larger training corpora

[3] We also report the performance of GPT-2 by iteratively filling the masked word and rank with their perplexity.

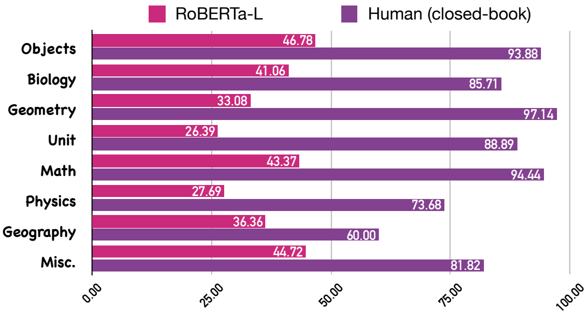

Figure 3: Performance of RoBERTa-Large V.S. human performance (closed-book tests) on different categories of numerical commonsense knowledge.

and focuses more on masked language modeling in its pre-training stage.

We see that our fine-tuning efforts do help improve model performance: "37.63 → 50.00" for BERT-large and "45.85 → 54.06" for RoBERTa-large. However, both are still far from the human's closed-book evaluation. Figure 3 shows PTLMs performance is poor across all categories within the core set of NUMERSENSE.

Comparing the performance of a PTLM on the "Core Probes" set (#=1,131) versus the "+ Adversarial Examples" set (#=3,145), we can measure their robustness. We found all models incur a significant performance drop when being evaluated on the adversarial set. This suggests that PTLMs (even when fine-tuned) can be brittle towards adversarial attacks, and future direction in pre-training language models should consider more structured inductive biases such as dependencies and semantic roles when learning contextual representations.

4 Case Studies

Object bias. Recall the example "a bird usually has `[MASK]` legs," which BERT-Large predicts to be "four". Does BERT-Large always predict "four" as long as the adjacent word after the [MASK] is 'legs'? To investigate if the bias exists, we show some case studies in Table 3. As 1,000 different randomly generated words fill the '[x]'s we see that both BERT and RoBERTa have a bias towards a certain answer, evidenced by the existence of a dominant answer in the softmax distribution. However, it seems that RoBERTa's (Liu et al., 2019) modified pre-training strategy helps it have less bias. We argue that future studies should further control the bias in masked language modeling.

Attention distribution. Following the prior probing work (Clark et al., 2019) on the relationship between attention weights and syntactic structures, we plot the attention distribution of the sentence "A bird usually has *two* legs." with respect to the word 'two' in Figure 4. We find that the root word 'has' enjoys the maximum attention at in the first few and middle layers, while the word 'two' gets the maximum attention to itself in the end. The important words for querying the numerical commonsense, namely 'birds' and 'legs', always have low attention weights. This suggests that the BERT (and RoBERTa) may inherently lose the relationship between subject/object and number words.

5 Open-Domain 'How-Many' Questions

The examples in the NUMERSENSE can be also seen as open-domain questions targeting 'how-many' commonsense—"how many legs does a fly usually have?" Answering these open-domain numerical commonsense questions is a practical downstream application of models that are successful in the NUMERSENSE. Thus, as a side note, we also report the performance of the state-of-the-art open-domain QA model (Asai et al., 2020).

We use the model that is trained on the Natural Question (NQ) dataset (Kwiatkowski et al., 2019), where we replace the '[MASK]'s in our examples with 'how many', so that our probes are in a similar format to NQ examples. For example "a fly usually has [MASK] legs" is converted to "*how many legs* a fly usually has?"[4] The accuracy of the state-of-the-art model is only **15.4%**, which is even lower than using BERT-base without fine-tuning. This indicates that improving performance on NUMERSENSE can help improve the performance on answering open-domain "how-many" questions.

6 Related Work

Probing Tasks for PTLMs. Prior work in probing language models have primarily focused on analysis of linguistic phenomena. Clark et al. (2019) investigated the relationship between BERT's attention weights and syntactic structures, while such as dependency (e.g. direct objects, noun modifiers), coreference, and sentence segmentation. Tenney et al. (2019) was able to display where certain types of linguistic information is captured within BERT—they in fact find the layers in a PTLM represent the steps of a classical NLP pipeline: POS

[4] We also manually test some queries such as "*how many legs* does a fly usually have?", which have similar results.

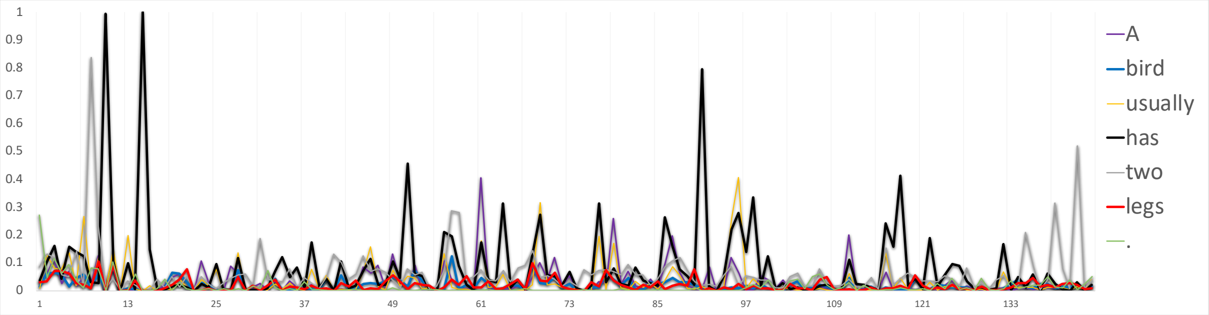

Figure 4: The attention distribution of the sentence “A bird usually has two legs.” on RoBERTa-base. We plot the attention weights (y) between each word and the number word ‘two’ at different position (x), e.g., $x = 13$ means (Layer 2, Head 1).

Template:	a [x] usually has [MASK] legs.
BERT-L	four: 39.3%, two: 18.3%, three: 10.1%
RoBERTa-L	four: 20.8%, two: 9.0%, three: 8.1%
Template:	most [x] have [MASK] wheels.
BERT-L	four: 25.3%, two: 14.1%, three: 5.1%
RoBERTa-L	four: 9.2%, two: 7.8%, three: 4.6%
Template:	all [x] have [MASK] sides.
BERT-L	two: 28.3%, three: 12.9%, four: 12.9%
RoBERTa-L	two: 16.6%, no: 2.9%, three: 2.3%

Table 3: The average Softmax of top 3 predictions in templates where ‘[x]’ is filled with 1k random words.

tagging, parsing, NER, semantic roles, and coreference. This line of work has indeed helped us understand the ability of PTLMs to capture *linguistic knowledge* via self-supervised learning from unlabeled data. We are interested in the numerical commonsense knowledge of PTLMs.

Probing Commonsense Knowledge. Besides the works that we have discussed in Section 1, Zhou et al. (2020) and Talmor et al. (2019a) also proposed to probe the commonsense knowledge of pre-trained language models, following the prior work by Trinh and Le (2018a and 2018b). They both utilized various existing language understanding datasets targeting commonsense knowledge to test if PTLMs can capture certain commonsense knowledge. Lin et al. (2019a) also show that PTLMs can retrieve paths from ConceptNet that aid in interpreting the decision made by the PTLMs on the CommonsenseQA dataset (Talmor et al., 2019b). Lin et al. (2019b) probe the commonsense knowledge in pre-trained language generation models via a constrained text generation task. However, they do not consider numerical commonsense knowledge, which is relatively under-explored area.

Numerical Commonsense Knowledge. Forbes and Choi (2017) and Goel et al. (2019) studied commonsense comparisons between two physical objects (e.g., a house is usually bigger than a person) in pre-trained word embeddings. Elazar et al. (2019) and Yamane et al. (2020) propose to induce the commonsense distribution of quantitative attributes (e.g., mass, length, and currency) of objects. Their goal is to extract or crowd-source such numerical attributes, and then obtain distributions that reflect commonsense knowledge. NumerSense, however, mainly focuses on exact numerical commonsense facts (e.g., a bird has *two* legs) instead of a range of values (e.g., a tiger weighs *around 120kg*), and have a larger number of arguments besides physical attributes.

Encoding Numerics for Computation. Wallace et al. (2019) probe PTLMs in terms of the ability to represent numeracy tokens by a regression task (e.g., “71” → 71.0), and also find that BERT is not good at encoding numerical tokens. Some works focus on incorporate algebra computation ability in PTLMs (Zou and Lu, 2019; Geva et al., 2020), thus making them able to answer math reasoning tasks such as MAWPS (Koncel-Kedziorski et al., 2016) and DROP (Dua et al., 2019). Note that these models and tasks are not targeting numerical commonsense knowledge but mainly the numerical-related computation within text.

7 Conclusion

We present a probing task, NumerSense, to induce numerical commonsense knowledge from pre-trained language models. We collect a new diagnostic dataset carefully verified by human annotators, which covers 8 different topics. Powerful pre-trained models such as BERT and RoBERTa perform surprisingly poorly, even after fine-tuning with high-quality distant supervision. We hope our findings and probing dataset will provide a basis for improving pre-trained masked language models’ *numerical* and other concrete types of commonsense knowledge.

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract No. 2019-19051600007, the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research, the Defense Advanced Research Projects Agency with award W911NF-19-20271, and NSF SMA 18-29268. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. We would like to thank all the collaborators in USC INK research lab for their constructive feedback on the work.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements. *ArXiv*, abs/2005.00660.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from bert. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. 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. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. 2019. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3973–3983, Florence, Italy. Association for Computational Linguistics.

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)*, pages 266–276, Vancouver, Canada. Association for Computational Linguistics.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. *ArXiv*, abs/2004.04487.

Pranav Goel, Shi Feng, and Jordan Boyd-Graber. 2019. 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. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019a. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Bill Yuchen Lin, Ming Shen, Yu Xing, Pei Zhou, and Xiang Ren. 2019b. CommonGen: A constrained text generation dataset towards generative commonsense reasoning. *ArXiv*, abs/1911.03705.

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 740–755. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar S. Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 1223–1237. Springer.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019a. olmpics - on what language model pre-training captures. *ArXiv*, abs/1912.13283.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019b. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Trieu H. Trinh and Quoc V. Le. 2018a. Do language models have common sense.

Trieu H. Trinh and Quoc V. Le. 2018b. A simple method for commonsense reasoning. *ArXiv*, abs/1806.02847.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. 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 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*.

Hiroaki Yamane, Chin-Yew Lin, and Tatsuya Harada. 2020. Measuring numerical common sense: Is a word embedding approach effective?

Xuhui Zhou, Y. Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Yanyan Zou and Wei Lu. 2019. Text2Math: End-to-end parsing text into math expressions. 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 5327–5337, Hong Kong, China. Association for Computational Linguistics.