TR at SemEval-2020 Task 4: Exploring the Limits of Language-model-based Common Sense Validation

Don Teo
Center for AI and Cognitive Computing
Thomson Reuters
Toronto, Canada
don.teo@tr.com

Abstract

In this paper, we present our submission for subtask A of the Common Sense Validation and Explanation (ComVE) shared task. We examine the ability of large-scale pre-trained language models to distinguish commonsense from non-commonsense statements. We also explore the utility of external resources that aim to supplement the world knowledge inherent in such language models, including commonsense knowledge graph embedding models, word concreteness ratings, and text-to-image generation models. We find that such resources provide insignificant gains to the performance of fine-tuned language models. We also provide a qualitative analysis of the limitations of the language model fine-tuned to this task.

1 Introduction

The task of assimilating general world knowledge from textual data for the purpose of commonsense reasoning and inference has been a long-standing challenge in natural language understanding (Davis, 1990; Schubert, 2002). A general approach to this problem that has gained recent popularity is the use of neural-network-based language models (LM) (Bengio et al., 2003). Such models, when trained on massive amounts of diverse text corpora, have been found to capture implicitly a remarkable amount of commonsense knowledge (Trinh and Le, 2018). Moreover, recent advances in training deep contextualized word representations using language-model-type objectives over large text corpora have substantially improved the state-of-the-art performance on a wide variety of natural language understanding tasks (Peters et al., 2018; Radford, 2018; Devlin et al., 2019; Yang et al., 2019; Raffel et al., 2019), including an even greater ability to capture commonsense knowledge (Zhou et al., 2019b; Porada et al., 2019).

Despite this success, a significant gap remains between the performance of such pre-trained LMs and human-level performance on commonsense reasoning tasks. In fact, there is evidence suggesting that current models fail to learn factual knowledge effectively (Poerner et al., 2019) and have difficulty with a variety of basic reasoning abilities necessary for commonsense inference (Talmor et al., 2019a; Kassner and Schütze, 2019). This motivates us to explore supplementary resources aimed at augmenting the world knowledge inherent in pre-trained LMs.

In this paper, we explore the ability of a state-of-the-art pre-trained LM in tackling the ComVE subtask A (Wang et al., 2020a). In particular, given a pair of sentences, we evaluate the model's ability in identifying which sentence least agrees with common sense. We consider the model from both a language modeling and a supervised learning perspective. Moreover, we explore the utility of additional resources meant to augment the perceptual world knowledge of language models in solving this task. Finally, we provide a categorization of the types of sentences that our best performing model struggles against.

2 Related Work

A variety of datasets have been created to examine a system's general commonsense knowledge and inference capability through such tasks as anaphora resolution (Levesque et al., 2012; Sakaguchi et al.,

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 601–608
Barcelona, Spain (Online), December 12, 2020.

2019) and question answering (Talmor et al., 2019b; Huang et al., 2019; Zellers et al., 2018). Certain tasks have attempted to measure commonsense knowledge specifically pertaining to the physical (Bisk et al., 2020), temporal (Zhou et al., 2019a), and causal (Gordon et al., 2012) aspects of reasoning.

Recent approaches to tackling these challenges have focused on methods to inject external knowledge into pre-trained LMs. The methods vary by the choice of knowledge sources and the training objective (Zhang et al., 2019; Lauscher et al., 2019; Levine et al., 2019; Peters et al., 2019; Xiong et al., 2019; He et al., 2019; Wang et al., 2020b). Complementary to such approaches, methods have been developed to automatically expand the coverage of existing commonsense knowledge bases (CKB) (Li et al., 2016; Saito et al., 2018; Bosselut et al., 2019; Zou, 2020). Due to the immense number of relations that would be needed to capture the wide variety of commonsense knowledge, He et al. (2020) proposed a method to "conceptualize" explicit relations into broader, more abstract concepts. In our work, we examine the benefit of using external resources such as CKB-enhanced word embeddings as simple, additional features to a model, rather than attempting to infuse the knowledge within the pre-trained LM. We also consider more perception-based resources, such as word concreteness ratings and text-to-image generation models.

3 System Descriptions

3.1 Masked language model

As a baseline system, we use the pre-trained RoBERTa$_{LARGE}$ (Liu et al., 2019) model as a masked language model (MLM). For a given sentence s, we measure the probability of each token in s conditioned on all other tokens in the sentence. We then consider the average p_{avg} or minimum p_{min} token probability across the entire sentence. When evaluating the pair of sentences (s_1, s_2), the sentence with the higher probability is taken to be the commonsense statement. We denote this approach as MLM$_{avg}$ and MLM$_{min}$ when using p_{avg} and p_{min}, respectively, as the comparator.

3.2 Feature-based model

The second system we consider uses a set of features that attempts to augment the MLM. In addition to using the difference Δp_{avg} (Δp_{min}) of the average (minimum) token probability scores between the pair of sentences, we build additional features that target specific differences between the two sentences and that leverage external knowledge resources.

3.2.1 Token difference perplexity

We consider specifically the subset of tokens that differ between s_1 and s_2. For example, in the pair of sentences *(He poured orange juice on his cereal., He poured milk on his cereal.)*, we are interested primarily in the MLM probability for the words *milk* and *orange juice*. Since the difference can span multiple tokens in a given sentence, we consider the perplexity $P_{\{s_i\}\setminus\{s_j\}}$ of the tokens remaining in sentence s_i after taking the setwise token difference with sentence s_j, where $\{s_k\}$ is the set of tokens in s_k. We take as a feature the difference $\Delta P = P_{\{s_1\}\setminus\{s_2\}} - P_{\{s_2\}\setminus\{s_1\}}$ of the two perplexities.

3.2.2 Concrete word similarity

Concreteness – the degree to which a concept refers to a perceptible entity – has been one of the most extensively studied variables in the field of psycholinguistics. For example, it has been observed that abstract and concrete concepts are represented and processed differently in the brain (Crutch and Warrington, 2004) and exhibit differences in terms of recall/memory (Walker and Hulme, 1999; Allen and Hulme, 2006; Romani et al., 2008; Miller and Roodenrys, 2009) and word association capability (Groot, 1989). Concrete concepts are an important aspect in commonsense knowledge and can, a priori, be particularly challenging for a model trained purely with textual data to learn about and represent effectively.

There have been multiple efforts to collect concreteness ratings for a large vocabulary (Paivio et al., 1968; Wilson and Division, 1997). Most recently, Brysbaert et al. (2014) collected a dataset of concreteness ratings for about 40K words and two-word phrases. We use this collection and consider all word and phrases as concrete if they have an average concreteness rating of at least 4.0, which reduces the list to about 9200 words and phrases.

The ConceptNet Numberbatch[1] (Speer et al., 2017) word embedding set combines the commonsense knowledge captured by the ConceptNet knowledge graph with existing word embedding sets learned through distributional semantics, such as word2vec and GloVe. We leverage these embeddings in the following way. For a given sentence, we measure the average cosine distance d_{conc} over all pairs of concrete words within the sentence. The intuition is that a commonsense statement should, at a minimum, have concrete words that are more semantically related. Continuing with the example given in Section 3.2.1, the concrete words for the two sentences are *(He poured orange* **juice** *on his* **cereal***., He poured* **milk** *on his* **cereal***.)*. One expects the words *milk* and *cereal* to be closer in embedding space than *juice* and *cereal*, not merely because the latter two words appear more often together than the former, but also because of the ConceptNet relation *ReceivesAction(milk, eaten with cereal)*. We use as a feature the difference in d_{conc} values between the two sentences.

3.2.3 Text-to-image generation

In addition to purely text-based resources, we explore the utility of incorporating world knowledge and common sense through visual perception. Specifically, we consider a state-of-the-art text-to-image (TTI) generation model that has been trained over a large corpus of captioned images. Descriptions of real-world images are a form of text that may be under-represented in the corpora that have been used for LM pre-training (e.g. stories, news and encyclopedia articles). More importantly, images contain commonsense information about the physical world that is often not explicitly stated in text (e.g. the relative sizes of objects). The intuition is that a TTI model would find it more difficult to generate an image for sentences that defy the constraints of its world model compared to the sentences that obey common sense. For example, consider the the pair of sentences *(He picked up a cup of orange juice., He picked up a cup of an elephant.)*. A useful TTI model would be able to generate an image of a cup given the first sentence more easily than when given the second sentence. A recently proposed measure of generated image quality is the Semantic Object Accuracy (SOA) (Hinz et al., 2019). In essence, the objects that are mentioned in the sentence ought to be present in the image, as determined by an appropriate object detector. We therefore consider the following approach:

1. Given TTI model ϕ and sentence s, generate N images for the sentence: $\phi(s)_1, \phi(s)_2, ..., \phi(s)_N$
2. Compute the object-category-averaged SOA over the generated images

$$\text{SOA-C} = \frac{1}{C} \sum_{c \epsilon C} \frac{1}{N} \sum_{i=1}^{N} I_Y(\phi(s)_i, c) \tag{1}$$

   where C is the number of object classes present in the sentence, and I_Y is the indicator function for whether image $\phi(s)_i$ contains an object of class c as determined by the object detection model Y.

In the computation of SOA-C, we restrict C to the objects that are common to both sentences so that differences in the object class generation ability of the TTI model are factored out. The sentence with the higher value of SOA-C would be considered the more commonsense statement. We use as a feature the difference Δ_{SOA-C} in SOA-C scores between the two sentences.

There are, however, significant limitations to the above approach. Crucially, a TTI model is limited by its training corpus in the types of objects that it can reasonably generate. Currently, most TTI models are trained over the Microsoft Common Objects in Context (COCO) dataset (Lin et al., 2014), which covers a set of only 80 object categories. In the ComVE task, only about 10% of sentence pairs have a common COCO object in both sentences. Moreover, the generation capability of the model depends strongly on the object class. For example, cats and dogs can be generated with much better accuracy than cups and bottles. Nevertheless, we consider the usefulness of this feature for this small subset of sentence pairs.

In our experiments, we use the DM-GAN network pretrained over the COCO dataset as the TTI model ϕ (Zhu et al., 2019) and generate $N = 3$ images per sentence. For the object detection model Y, we use the YOLOv3 model trained over the same dataset (Redmon and Farhadi, 2018). Figure 1 shows two generated images for a sample sentence pair.

[1] https://github.com/commonsense/conceptnet-numberbatch

Figure 1: Sample generated images for *The cat ran away from the dog* (left) and *The house ran away from the dog* (right). In the left image, both a dog and a cat were identified by the object detector. The SOA-C scores for the two sentences were 1/3 and 0, respectively ($N = 3$, $C = 1$ (*dog*)).

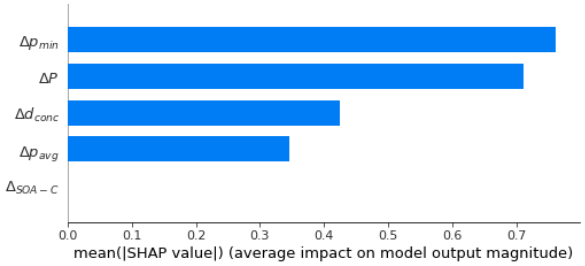

Figure 2: Feature importance for the GBDT model.

3.2.4 GBDT Model

The above features are combined with a Gradient Boosted Decision Tree (GBDT) trained using the XGBoost library (Chen and Guestrin, 2016). Figure 2 illustrates the relative feature importance of the model using SHAP values (Lundberg and Lee, 2017; Lundberg et al., 2020). The most important features are the difference in p_{min} scores and the difference in perplexity scores of differing tokens. As can be seen, the difference in SOA-C scores was an ineffective feature. Even within the 10% of sentence pairs that contain a common COCO object, the difference in SOA-C scores exhibited a negligible correlation (< 0.05 Pearson coefficient) with the prediction label. Thus, for this particular task, this feature had no impact to the final test set predictions of the model.

3.3 Fine-tuned LM

The final approach we consider is fine-tuning a pre-trained LM for this particular task. We use again the RoBERTa$_{LARGE}$ pre-trained model for this purpose. Each sentence pair (s_1, s_2) is fed directly into the model in the same manner as any other type of downstream sentence-pair task. The output of the special `[CLS]` input token is fed into a multilayer perceptron as a classification task. We also consider a second GBDT model (referred to below as the GBDT-RoBERTa model) that uses the fine-tuned RoBERTa model's classification score as an additional feature to the features described above.

4 Results and Discussion

Table 1 summarizes the results of the various approaches and compares with the performance of the top submission for the task. Overall, we find that the simple MLM-based approaches achieve a reasonable baseline performance on this task. The addition of targeted features in the GBDT model gives an improvement of several percentage points on top of the best baseline. The fine-tuned RoBERTa model shows a substantial performance gain over the GBDT model. Finally, the GBDT-RoBERTa model achieves a marginally higher score on the test set over the fine-tuned RoBERTa model alone.

Method	**Dev Acc.**	**Test Acc.**
MLM$_{avg}$	70.3	72.2
MLM$_{min}$	77.2	75.3
GBDT (official submission)	80.8	79.7
RoBERTa	**95.2**	92.7
GBDT-RoBERTa	95.0	**92.9**
Team *hit_itnlp* (1^{st} place submission)	-	97.0

Table 1: Accuracy on the development and test sets for each system considered.

To provide more insight into the errors the fine-tuned RoBERTa model makes, we categorize the incorrectly labeled sentence pairs depending on the key difference between the sentences. We consider the following broad types:

- Quantity: Statements that require an assessment of numeric values.
- Physical perception: Statements requiring perceptual knowledge (e.g. the relative size of objects) or general science knowledge (e.g. every person has a heart).
- Temporal perception: Statements dealing with the duration of events, temporal ordering/causation, or other time-related knowledge (e.g. when an event occurs).
- Definition: Statements that require knowledge of the definition of a particular word.
- Negation: One of the statements contains a negation.
- Data quality: Sentence pairs that were found to be grammatically awkward or where it was not immediately obvious which statement was the commonsense statement.

Table 2 gives a breakdown of the proportion of errors found within the development set[2]. We find that the bulk of the errors required some aspect of physical or temporal world knowledge.

We also compare several characteristics between the set of correct predictions $S_{\checkmark}$ and incorrect prediction $S_{\text{✗}}$ within the development set. In terms of the negation category, the proportion of sentence pairs for which at least one sentence contains a *not* lemma was substantially larger in $S_{\text{✗}}$ (12.8%) than in $S_{\checkmark}$ (2.6%), indicating some difficulty in interpreting negation statements. In terms of concept concreteness, we find that both sets show a similar distribution in the average number of unique concrete terms (2.4 in $S_{\checkmark}$ vs 2.1 in $S_{\text{✗}}$). This suggests that the model had no particular difficulty in representing commonsense knowledge involving concrete terms. Finally, in terms of the lexical similarity of the sentence pair, we find that incorrect predictions exhibited a slightly higher average number in the setwise symmetric difference of tokens ($\{s_1\} \ominus \{s_2\}$) between the two sentences than correct predictions (3.6 in $S_{\checkmark}$ vs 4.4 in $S_{\text{✗}}$).

Category	**Proportion**	**Example** (s_1, s_2)
Physical	38%	*(Babies are born naked., Babies are born with clothes on.)*
Temporal	19%	*(He lived without Sara for about a year., He lived without food for about a year.)*
Quantity	9%	*(You have three fingers on one hand., You have five fingers on one hand.)*
Negation	13%	*(Trees can sometimes live in saltwater., Trees can not live on the ground.)*
Definition	11%	*(Pork comes from cows., Pork comes from pigs.)*
Data Quality	10%	*(A bald man brushed his hair every day., A bald man washed his hair every day.)*

Table 2: Proportion of question categories in $S_{\text{✗}}$ for the RoBERTa fine-tuned model.

5 Conclusion and Future Work

In this paper, we evaluated the performance of a state-of-the-art pre-trained LM on the task of common sense validation. We also explored the usefulness of external resources meant to supplement the implicit commonsense knowledge of the LM. We found that a subset of these resources provide value to a system relying only LM probabilities, but give negligible improvement to an LM fine-tuned to the task. Further experiments are needed to evaluate the performance of pre-trained LMs on this task when they are explicitly adapted with external commonsense knowledge bases. Additionally, we leave as future work the study of TTI models with improved object coverage and generation quality that may eventually add value to the subspace of common sense validation that depend significantly on visual world knowledge.

[2]The categories are not mutually exclusive; we compute the overall proportion by evaluating the categories in the order or Data Quality, Negation, Definition, Quantity, Temporal, and Physical.

Acknowledgements

We would like to thank Shohreh Shaghaghian and John Hudzina for their comments on an earlier version of this work.

References

Richard Allen and Charles Hulme. 2006. Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language*, 55(1):64 – 88.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155, March.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *ArXiv*, abs/1911.11641.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904—911, September.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *ArXiv*, abs/1603.02754.

Sebastian J. Crutch and Elizabeth K. Warrington. 2004. Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3):615–627, 11.

Ernest Davis. 1990. *Representations of Commonsense Knowledge*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Annette Groot. 1989. Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15:824–845, 09.

Bin He, Di Zhou, JingHui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2019. Integrating graph contextualized knowledge into pre-trained language models. *ArXiv*, abs/1912.00147.

Mutian He, Yangqiu Song, Kun Xu, and Yu Dong. 2020. On the role of conceptualization in commonsense knowledge graph construction. *ArXiv*, abs/2003.03239.

Tobias Hinz, Stefan Heinrich, and Stefan Wermter. 2019. Semantic object accuracy for generative text-to-image synthesis. *ArXiv*, abs/1910.13321.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *EMNLP/IJCNLP*.

Nora Kassner and Hinrich Schütze. 2019. Negated lama: Birds cannot fly. *ArXiv*, abs/1911.03343.

Anne Lauscher, Ivan Vulic, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavas. 2019. Informing unsupervised pretraining with external linguistic knowledge. *ArXiv*, abs/1909.02339.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Yoav Levine, Barak Lenz, Or Dagan, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. Sensebert: Driving some sense into bert. *ArXiv*, abs/1908.05646.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *ACL*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

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

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Scott Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2, 01.

Leonie Miller and Steven Roodenrys. 2009. The interaction of word frequency and concreteness in immediate serial recall. *Memory cognition*, 37:850–65, 10.

Allan Paivio, John C. Yuille, and Stephen A. Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76 1:Suppl:1–25.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Matthew E. Peters, Mark Neumann, IV RobertLLogan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *EMNLP/IJCNLP*.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. Bert is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised qa. *ArXiv*, abs/1911.03681.

Ian Porada, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. Can a gorilla ride a camel? learning semantic plausibility from text. *ArXiv*, abs/1911.05689.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767.

Cristina Romani, Sheila McAlpine, and Randi Martin. 2008. Concreteness effects in different tasks: Implications for models of short-term memory. *Quarterly journal of experimental psychology (2006)*, 61:292–323, 03.

Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2018. Commonsense knowledge base completion and generation. In *CoNLL*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *ArXiv*, abs/1907.10641.

Lenhart K. Schubert. 2002. Can we derive general world knowledge from texts.

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 *NAACL-HLT*.

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

Ian Walker and Charles Hulme. 1999. Concrete words are easier to recall than abstract words : Evidence for a semantic contribution to short-term serial recall.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020a. SemEval-2020 task 4: Commonsense validation and explanation. In *Proceedings of The 14th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Cuihong Cao, Daxin Jiang, and Ming Zhou. 2020b. K-adapter: Infusing knowledge into pre-trained models with adapters. *ArXiv*, abs/2002.01808.

Michael Wilson and Informatics Division. 1997. Mrc psycholinguistic database: Machine usable dictionary, version 2.00. *Behav Res Methods*, 20, 06.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2019. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. *ArXiv*, abs/1912.09637.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *ACL*.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019a. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *EMNLP/IJCNLP*.

Xuhui Zhou, Y. Zhang, Leyang Cui, and Dandan Huang. 2019b. Evaluating commonsense in pre-trained language models. *ArXiv*, abs/1911.11931.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5795–5803.

Yanyan Zou. 2020. Mining commonsense facts from the physical world. *ArXiv*, abs/2002.03149.