Designing Templates for Eliciting Commonsense Knowledge from Pretrained Sequence-to-Sequence Models

Sheng-Chieh Lin*[1], Jheng-Hong Yang*[1], Rodrigo Nogueira[2],
Ming-Feng Tsai[1,3], Chuan-Ju Wang[1] and Jimmy Lin[2]

[1]Research Center for Information Technology Innovation, Academia Sinica
[2]David R. Cheriton School of Computer Science, University of Waterloo
[3]Department of Computer Science, National Chenchi University

Abstract

While internalized "implicit knowledge" in pretrained transformers has led to fruitful progress in many natural language understanding tasks, how to most effectively elicit such knowledge remains an open question. Based on the text-to-text transfer transformer (T5) model, this work explores a template-based approach to extract implicit knowledge for commonsense reasoning on multiple-choice (MC) question answering tasks. Experiments on three representative MC datasets show the surprisingly good performance of our simple template, coupled with a logit normalization technique for disambiguation. Furthermore, we verify that our proposed template can be easily extended to other MC tasks with contexts such as supporting facts in open-book question answering settings. Starting from the MC task, this work initiates further research to find generic natural language templates that can effectively leverage stored knowledge in pretrained models.

1 Introduction

Scaling and unifying deep neural models via building task-agnostic architectures and pretraining objectives have brought about significant progress in performing various downstream natural language understanding tasks (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Lewis et al., 2019; Raffel et al., 2020). According to recent findings, impressive performance can be attributed to a sort of internalized implicit "knowledge base" contained in pretrained models (Petroni et al., 2019; Jiang et al., 2019). Since such knowledge is drawn from unstructured and unlabeled text via the design of neural architectures and pretraining objectives, it is non-trivial to identify an effective way to explicitly elicit such stored knowledge for each NLP task.

Challenging NLP tasks that require commonsense reasoning abilities (or background knowledge) could benefit from implicit knowledge captured in pretrained models. Various research lines have explored techniques that effectively leverage reasoning abilities based on BERT (Devlin et al., 2019). For example, Sakaguchi et al. (2019) use a classification head capped on pretrained BERT variants to tackle commonsense reasoning. Zhu et al. (2020) propose to perturb token embeddings to improve model performance and robustness. Alternatively, many researchers recast the multiple-choice (MC) problem setting, which is frequently used in verbal reasoning benchmarks, as a ranking problem regarding statement possibilities (Li et al., 2019; Pirtoaca et al., 2019). Tamborrino et al. (2020) further present a hybrid method that combines full-text inputs, pretrained token representations, and learning-to-rank objectives. Finally, Khashabi et al. (2020) propose to unify and fine-tune pretrained models on many QA tasks jointly.

As an alternative to designing task-specific architectures on top of pretrained transformers, we propose a new approach to downstream NLP tasks based on task-specific templates exploiting existing pretrained sequence-to-sequence models. We aim to identify a text-to-text approach that effectively elicits implicit knowledge embedded in the pretrained text-to-text transfer transformer

*Contributed equally.

Proceedings of the 28th International Conference on Computational Linguistics, pages 3449–3453
Barcelona, Spain (Online), December 8-13, 2020

Source	Target
hypothesis: *home* is smaller. **premise:** He never comes to my home, but I always go to his house because the	**entailment**
hypothesis: *house* is smaller. **premise:** He never comes to my home, but I always go to his house because the	**contradiction**

Table 1: Decomposing WinoGrande problems into training instances for T5.

(T5) (Raffel et al., 2020) for commonsense reasoning. Specifically, we find that T5 performs surprisingly well by framing multiple-choice (MC) QA problems as ranking multiple text sequences associated with two predefined (and also pretrained) output tokens. Empirically, we demonstrate that without fine-tuning, the pretrained sequence-to-sequence model achieves better-than-random performance on commonsense reasoning tasks. With fine-tuning, our approach yields state-of-the-art performance (at the time it was proposed) on three MC datasets: WinoGrande (Sakaguchi et al., 2019), OpenBookQA (Mihaylov et al., 2018), and ARC-Easy (Clark et al., 2018).

2 Approach

An approach based on natural language templates enables various options in formulating a verbal reasoning task as a text-to-text problem with T5. Our proposed novel formulation builds on T5's MNLI template, composed of a hypothesis and a premise.[1] Consider a concrete example in WinoGrande:

> He never comes to my home, but I always go to his house because the _ is smaller.
> Option1: home; Option2: house

In this case, the correct replacement for _ is Option1. We decompose the above problem into two source–target training examples, where _ is replaced with each option and annotated with the correct answer as the target token, as shown in Table 1. In addition, we reformulate each example into a commonsense reasoning "template" with two statements: hypothesis (from _ to the end of the original problem statement) and premise (the remaining part of the original problem statement). Note that the bold and colored fonts are for clarity only; those tokens are not marked in any way in the model input.

At inference (test) time, we also decompose the test problem into two inputs, where each input is formulated in exactly the same manner shown in Table 1, with either one of the answer options. We then feed each into T5 to predict a target token. In this scenario, there are four possible outcomes: (1) one produces "entailment" and the other "contradiction", (2) one produces "entailment" or "contradiction" and the other some other token, (3) both produce some other tokens, and (4) both produce the same token, either "entailment" or "contradiction."

Ideally, T5 would produce contrastive tokens for each input pair, as in case (1), which allows us to unambiguously select the final answer. However, the model might produce the same tokens for each input, or even tokens not in the predefined set, as in cases (2) to (4). To deal with these cases, we apply a softmax over the logits of the pair of predefined target tokens, similar to Nogueira et al. (2020). From this, we compute the probabilities of the predefined target tokens (in the case of Table 1, "entailment" and "contradiction"). Then, we compare the probabilities of both input instances, and in cases (2) to (4), we select the instance that has a higher probability as the correct answer.

3 Experiments

In our experiments, we fine-tune T5-3B on Google Colab's TPU v2 with a batch size of 16, a learning rate of 2×10^{-4}, save model checkpoints every 5000 steps and choose the checkpoints

[1] See appendix D.3 in (Raffel et al., 2020) for more details.

	Condition		Training size					
Condition	Target token	Logit	Zero-Shot	XS	S	M	L	XL
#1	entailment/contradiction		0.506	0.657	0.693	0.757	0.809	0.840
#2		✓	0.608	0.718	0.740	0.788	0.837	0.854
#3	true/false		0.477	0.676	0.697	0.760	0.823	0.852
#4		✓	0.566	0.723	0.752	0.800	0.843	0.865
Our leaderboard submission (test set)			-	0.683	0.705	0.776	0.824	0.846

Table 2: Results on WinoGrande, measured by the accuracy of models trained on different dataset sizes. Condition #2 is our leaderboard submission.

with the best performance on the development set. At inference, we use greedy decoding and select for evaluation the model checkpoint that achieves the highest score on the development set. Note that we do not experiment with T5-11B due to limited computational resources. We report accuracy as our scoring metric for evaluating model performance.

3.1 WinoGrande

Experimental results on WinoGrande are reported in Table 2 for different training dataset sizes; Conditions #1–#4 report development set results. Note that we fine-tune the model for each training dataset size separately. A ✓under the "logit" column indicates that we apply the softmax over the target tokens as described above. Without this technique, given the original two-choice question, if T5 outputs the same tokens for the two processed inputs, we simply assign Option1 as the answer. The table also reports "zero-shot" performance, i.e., performing inference on the development set without any model fine-tuning. Condition #2 represents our submission to the official leaderboard, which performs consistently well on the held-out test set.

We see that the logit trick improves performance, which is consistent with the observations in Nogueira et al. (2020). Applying the technique in the zero-shot setting yields performance that is clearly better than random. Our general problem setup allows us to choose different target tokens: In addition to selecting "entailment" vs. "contradiction" as the target, we also try the contrastive pair "true" vs. "false". The choice of the target token appears to have an impact on the performance, which is also consistent with the findings in Nogueira et al. (2020).

Looking at the WinoGrande leaderboard, our submission represents the state of the art at the time of submission. As of November 2020, two other entries achieve better performance, both based on T5-11B (an even larger model) and exploiting multi-task learning. Clearly, both are inspired by and built on our work.

3.2 OpenBookQA and ARC-Easy

We also investigate how T5's reasoning capability improves multiple-choice QA. Specifically, we conduct experiments on two QA datasets: OpenBookQA (Mihaylov et al., 2018) and ARC (Clark et al., 2018). Consider one example from ARC-Easy:

> **Question:** A green plant absorbs light. A frog eats flies. These are examples of how organisms
> **Choice A:** obtain energy; **B:** escape predators; **C:** produce offspring; **D:** excrete waste
> **Context:** organism that obtains energy by eating both plants and animals.

This task requires commonsense reasoning capabilities beyond the typical challenges expected for QA, such as the understanding of paraphrases and coreference resolution.[2] That is, to answer the above question, models need to know that frogs and flies are animals and light is a form of energy in the context.

[2]The tasks provide predetermined corpora containing scientific facts relevant to the questions. Although retrieval methods may affect the results, this is beyond the scope of our discussion.

Condition	Dataset	
	OpenBookQA	ARC-Easy
w/o contexts	0.768	0.808
w/ contexts	0.834	0.872
Our submission (test set)	0.832	0.891

Table 3: Results on OpenBookQA and ARC-Easy, measured by accuracy. We conduct the experiments with true/false target tokens and the logit trick, corresponding to condition #4 in Table 2.

Applying the template in Table 1, for each question–choice pair, we take the question text as its hypothesis, replace each choice text in _, and put the context text as its premise; the target token is annotated by the ground truth choice. That is to say, if "obtain energy" is the correct choice, we annotate the corresponding target token with the correct token from our predefined set ("true" in this case; "false", otherwise). Table 3 compares the model performance with and without contexts (i.e., removing the premise part). We observe from Table 3 that there exists a significant accuracy improvement (around 8% relative) with contexts in both datasets, which demonstrates T5's ability to incorporate explicit knowledge contained in the contexts. Based on the official leaderboard, our technique represents the state of the art at the time of submission.

4 Conclusions

Collectively, the success of large pretrained neural models, both encoder-only BERT-like architectures as well as encoder–decoder architectures such as T5, raises interesting questions for the pursuit of commonsense reasoning abilities. Researchers have discovered that previous models perform well on benchmark datasets because they detect incidental biases in the dataset that have nothing to do with the task; in contrast, the WinoGrande dataset has devoted considerable effort to reducing such biases, which may allow models to (inadvertently) "cheat" (for example, using simple statistical associations). While it is certainly true that datasets over-estimate the reasoning capabilities of modern models (Sakaguchi et al., 2019), there are alternative and complementary explanations as well.

It has been a fundamental assumption of the research community that commonsense reasoning is difficult because it comprises tacit rather than explicit knowledge (Winograd, 1972). That is, commonsense knowledge—like water is wet and that a tuba is usually too big to fit in a backpack—is not written down anywhere (unlike, say, factual knowledge, which can be modeled in a knowledge graph). As a result—the reasoning goes—data-driven techniques (even neural models) will be of limited use due to the paucity of *relevant* corpora.

Yet, previous encoder-only architectures (Devlin et al., 2019; Liu et al., 2019) that exploit a language modeling objective (that is, relying *only* on explicit textual knowledge) can clearly make headway in language reasoning tasks, and we can further improve upon these approaches with a encoder–decoder model. This leaves us with two possible explanations: despite careful controls, the WinoGrande challenge and other datasets *still* contain incidental biases that these more sophisticated pretrained models can exploit, or that we are genuinely making at least *some progress* in commonsense reasoning. The latter, in particular, challenges the notion that commonsense knowledge is (mostly) tacit. Perhaps it is the case that in a humongous corpus of natural language text, someone really has written about trying to stuff a tuba in a backpack?

Acknowledgments

This research was supported in part by the Canada First Research Excellence Fund, the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the Waterloo–Huawei Joint Innovation Lab.

References

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv:1803.05457*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pages 4171–4186.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2019. How can we know what language models know? *arXiv:1911.12543*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing format boundaries with a single QA system. *arXiv:2005.00700*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv:1910.13461*.

Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. 2019. Learning to rank for plausible plausibility. In *Proc. ACL*, pages 4818–4823.

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. *arXiv:1907.11692*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proc. EMNLP*.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv:2003.06713*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proc. EMNLP-IJCNLP*, pages 2463–2473.

George Sebastian Pirtoaca, Traian Rebedea, and Stefan Ruseti. 2019. Answering questions by learning to rank - learning to rank by answering questions. In *Proc. EMNLP-IJCNLP*, pages 2531–2540.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WinoGrande: An adversarial Winograd schema challenge at scale. *arXiv:1907.10641*.

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pre-training is (almost) all you need: An application to commonsense reasoning. In *Proc. ACL*, pages 3878–3887.

Terry Winograd. 1972. Understanding natural language. *Cogn. Psychol.*, 3(1):1 – 191.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv:1906.08237*.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. FreeLB: Enhanced adversarial training for natural language understanding. In *Proc. ICLR*.