Commonsense Knowledge-Augmented Pretrained Language Models for Causal Reasoning Classification

Anonymous ACL submission

Abstract

Commonsense knowledge can be leveraged for identifying causal relations in text. In this work, we convert triples in ATOMIC$^{20}_{20}$, a wide coverage commonsense reasoning knowledge graph, to natural language text and continually pretrain a BERT pretrained language model. We evaluate the resulting model on answering commonsense reasoning questions. Our results show that a continually pretrained language model augmented with commonsense reasoning knowledge outperforms our baseline on two commonsense causal reasoning benchmarks, COPA and BCOPA-CE, without additional improvement on the base model or using quality-enhanced data for fine-tuning.

1 Introduction

Automatic extraction and classification of causal relations in text has been an important yet challenging task in natural language processing and understanding. Early methods back in the 80s and 90s (Joskowicz et al., 1989; Kaplan and Berry-Rogghe, 1991; Garcia et al., 1997; Khoo et al., 1998) mainly relied on defining hand-crafted rules to find cause-effect relations. Starting 2000, machine learning tools were utilized in building causal relation extraction models (Girju, 2003; Chang and Choi, 2004, 2006; Blanco et al., 2008; Do et al., 2011; Hashimoto et al., 2012; Hidey and McKeown, 2016). Word-embeddings and pretrained language models have also been leveraged in training models for understanding causality in language in recent years (Dunietz et al., 2018; Pennington et al., 2014; Dasgupta et al., 2018; Gao et al., 2019).

Investigating the true capability of pretrained language models in understanding causality in text is still an open question. More recently, Knowledge Graphs (KGs) have been used in combination with pretrained language models to address commonsense reasoning. CausalBERT (Li et al., 2020) for guided generation of Cause and Effect or the model introduced by Guan et al. (2020) for commonsense story generation are two examples.

Motivated by the success of Continual pretraining of already Pre-trained Language Models (PLMs) for downstream tasks (Gururangan et al., 2020), we explore the impact of common sense knowledge injection as a form of continual pretraining for causal reasoning. We hypothesize that continual pretraining of LMs using commonsense knowledge should improve performance on commonsense reasoning and causality identification. Moreover, models with a significantly fewer number of parameters (BERT) compared to large PLMs such as DeBERTa (He et al., 2020), Google T5 (Raffel et al., 2019), or GPT-3 (Brown et al., 2020) can benefit from such a continual pretraining.

2 Method

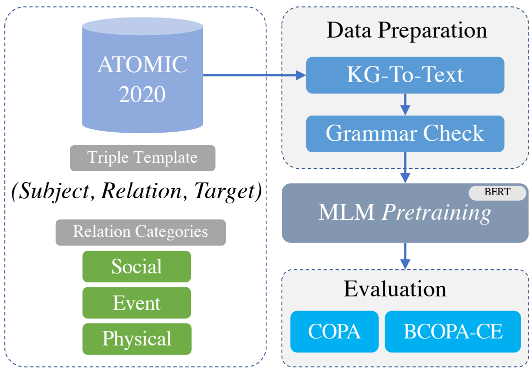

Figure 1: Overview of our proposed framework to continually pretrain language models to augment them with commonsense reasoning knowledge.

2.1 KG-To-Text Conversion

We convert triples in ATOMIC$^{20}_{20}$ (Hwang et al., 2021) knowledge graph to natural language texts to use them as input in our continual pretraining. Samples in ATOMIC$^{20}_{20}$ are stored as triples in form of *(head/subject, relation, tail/target)* in three splits including train, development, and test. We only use triples from the *train* split in our pretraining. ATOMIC$^{20}_{20}$ has 23 relation types that are classified into three categorical types including commonsense relations of social interactions, physical-entity commonsense relations, and event-centric commonsense relations. In the rest of the paper, we refer to these three categories as social, physical, and event, respectively.

Before converting the triples, we also take some preprocessing steps to filter out some triples in ATOMIC$^{20}_{20}$ that we think may not suit our goal here. In particular, we remove all duplicates[1] and ignore all triples in which the target value is *none*. Moreover, we ignore all triples that include a blank. Since in masked language modeling we need to know the gold value of masked tokens, a triple that already has a blank (masked token/word) in it may not help our pretraining. For instance, in the triple: `[PersonX affords another ___, xAttr, useful]` it is hard to know why or understand what it means for a person to be useful without knowing what they afforded. The preprocessing step resulted in 782,848 triples with 121,681, 177,706, and 483,461 from event, physical, and social categories, respectively. Distribution of these relations is shown in Figure 2.

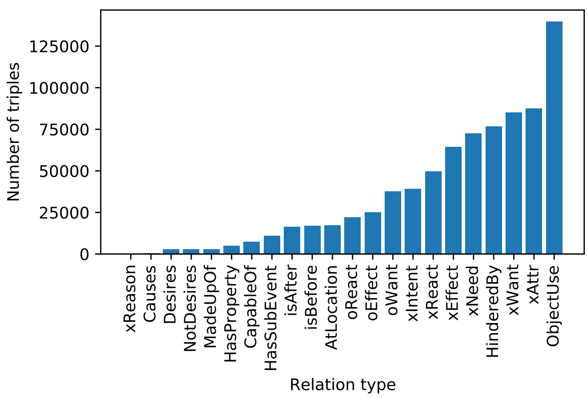

Figure 2: Number of relation types from ATOMIC$^{20}_{20}$ used in our pretraining.

Converting Triples: Each relation in ATOMIC$^{20}_{20}$ is associated with a human-readable template. For example, *xEffect*'s and *HasPrerequisite*'s templates are *as a result, PersonX will* and *to do this, one requires*, respectively. We use these templates to convert triples in ATOMIC$^{20}_{20}$ to sentences in natural language by concatenating the subject, relation template, and target. Examples of converting triples to text are shown in Figure 3.

2.2 Checking Grammar

When we convert triples to natural language text, ideally we want to have grammatically correct sentences. For example, after concatenating relation type and target in a tuple of knowledge graph, we may have a sentence such as: *As a result, PersonX wants leave* which is grammatically incorrect since there is a *to* missing after *wants*. To address this issue, we use an open-source grammar and spell checker, LanguageTool,[2] to double-check our converted triples to ensure they do not contain obvious grammatical mistakes. Similar approaches that include deterministic grammatical transformations were also previously used to convert KG triples to coherent sentences (Davison et al., 2019). It is worth pointing out that the Data-To-Text generation (KG verbalization) for itself is a separate task and there have been efforts to address this task (Agarwal et al., 2021). Investigating other Data-To-Text and grammar checking methods to see whether they improve the quality of generated text from KG can be considered as one next step.

The grammar checking process resulted in modifying total of 151,783 samples (%19 of all samples).[3]

2.3 Continual Pretraining

We use Masked Language Modeling (MLM)[4] to continually pretrain our PLM, *BERT-large-cased* (Devlin et al., 2018). We follow the same procedure as BERT to create our pretraining samples (e.g. number of tokens to mask in input examples). We run the pretraining by default for 15 epochs on a Google Colab TPU v2 with block size (maximum sequence length) of 32 and batch size of 32 and save the checkpoints at every 5000 steps. To avoid overfitting, we stop the pretraining when the pretrained model shows no improvement in terms of *training loss* after one epoch.

3 Experiments

In our experiments, we first run a 10-fold cross-validation on the training set for tuning the hyper-

[1]There are 68,626, 7,410, and 8,473 duplicate triples in train, development, and test sets of ATOMIC$^{20}_{20}$, respectively. These duplicate triples are redundant and indicate multiple annotators for some head/relation pairs.

[2]https://languagetool.org/

[3]We make the converted samples and conversion codes publicly available. We have also flagged all the corrected/modified samples.

[4]*BertForMaskedLM* implementation from the Huggingface's transformers. We will share our pretrained models publicly on Huggingface's model hub.

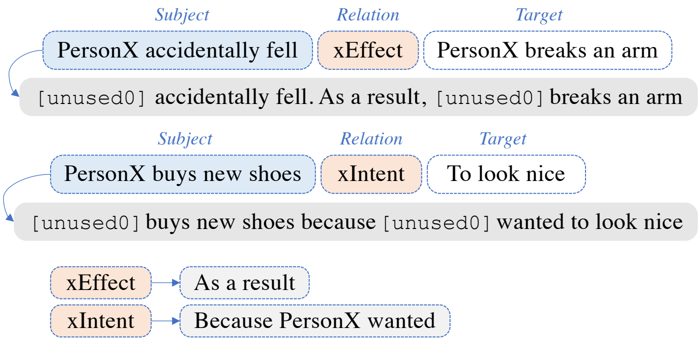

Figure 3: Examples of converting two triples in ATOMIC$^{20}_{20}$ in form of (Subject, Relation, Target) to natural language text using human readable templates. *PersonX* is replaced by *[unused0]* token from BERT's vocabulary to avoid an out-of-vocabulary issue.

parameters. Then, using the best hyperparameter tuning trial, we fine-tune our models with four different random seeds using the entire training set, evaluate the fine-tuned models on the test set, and report the average performance.

3.1 Benchmarks

We chose two benchmarks of commonsense causal questions: 1) the Choice Of Plausible Alternatives (COPA) (Roemmele et al., 2011) dataset which is a widely used and notable benchmark (Rogers et al., 2021) for commonsense causal reasoning. And, 2) BCOPA-CE (Han and Wang, 2021), a new benchmark inspired by COPA, that contains unbiased token distributions which makes it a more challenging benchmark to distinguish cause and effect in causal reasoning. Since COPA does not have a training set, we use COPA's development set (COPA-dev) in all experiments for fine-tuning our models and test the fine-tuned models on COPA's test set (COPA-test) and BCOPA-CE.

Baseline: we use the original *bert-large-cased* pretrained model in all experiments as our baseline. We use the Huggingface's MultipleChoice head on top of BERT and convert COPA and BCOPA-CE samples to a SWAG-formatted data (Zellers et al., 2018) suitable as input for our task. An example of converting a sample in COPA is shown in Figure 4 (Example A).

4 Results and Discussion

Results of our experiments on COPA-test are shown in Table 1. We initially observed that a continually pretrained model using all three types of relations has a lower performance than our baseline. By taking a closer look at each relation type, we decided to train another model, this time only using the *event* relations. The reason is that event relations in ATOMIC$^{20}_{20}$ specifically contain commonsense knowledge about event interaction for understating likely causal relations between events in the world (Hwang et al., 2021). In addition, event relations have a relatively longer context (# of tokens) than the average of all three relation types combined which means more context for a model to learn from. Our new pretrained model outperformed the baseline by %4.1 which shows the effect of augmented pretrained language model with commonsense reasoning knowledge.

Model	Acc (%)
PMI (Roemmele et al., 2011)	58.8
b-l-*reg* (Han and Wang, 2021)	71.1
Google T5-base (Raffel et al., 2019)	71.2
BERT-large (Kavumba et al., 2019)	76.5
CausalBERT (Li et al., 2020)	78.6
BERT-large (baseline) ❇	75.1
ATOMIC-BERT-large$_{MLM}$ ❇	
- Event, Physical, Social	74.3
- Event only	79.2
Google T5-11B (Raffel et al., 2019)	94.8
DeBERTa-1.5B (He et al., 2020)	**96.8**

Table 1: COPA-test Accuracy results. Our Models are marked by ❇. **b-l-* is a BERT-large model.

We also ran another experiment on the *Easy* and *Hard* question splits in COPA-test separated by Kavumba et al. (2019) to see how our best model performs on harder questions in COPA-test that do not contain superficial cues. Results are shown in Table 2. As can be seen, our ATOMIC-BERT model outperforms both the baseline and former models on Hard and Easy questions.

It is worth mentioning two points here. First,

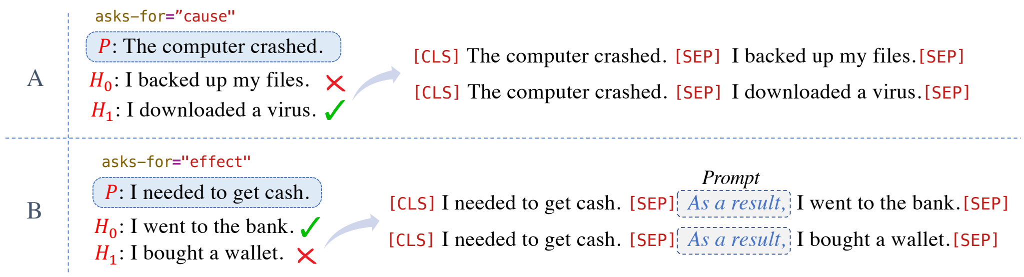

Figure 4: Examples of converting COPA samples to MultipleChoice format with and without adding prompt to the second sentence. For samples with `asks-for="cause"`, we add *It is because* as prompt.

Model	COPA-test Easy ↑	COPA-test Hard ↑
(Han and Wang, 2021)	-	69.7
(Kavumba et al., 2019)	83.9	71.9
BERT-large (baseline) ❇	84.1	69.7
ATOMIC-BERT-large ❇	**88.3**	**73.5**

Table 2: COPA-test Accuracy results on Easy and Hard question subsets. Models marked by ❇ are our models.

our model, BERT-large, has a significantly lower number of parameters than state-of-the-art models, Google T5-11B (∼32x) and DeBERTa-1.5B (∼4x). Second, we have not yet applied any model improvement methods such as using a margin-based loss introduced by Li et al. (2019) and used in CausalBERT (Li et al., 2020), an extra regularization loss proposed by Han and Wang (2021), or fine-tuning with quality-enhanced training data, BCOPA, introduced by Kavumba et al. (2019). As a result, there is still great room to improve current models that can be a proper next step and follow up on our work.

Model	**Acc (%)**
b-l-*aug* (Han and Wang, 2021)	51.1
b-l-*reg* (Han and Wang, 2021)	**64.1**
BERT-large (baseline) ❇	55.8
ATOMIC-BERT-large$_{MLM}$ ❇	
- Event, Physical, Social	54.1
- Event only	58.1

Table 3: BCOPA-CE Accuracy results. Models marked by ❇ are our models. **b-l-* is a BERT-large model.

4.1 BCOPA-CE: Prompt vs. No Prompt

Results of experiments on BCOPA-CE are shown in Table 3. As expected based on the results also reported by Han and Wang (2021), we initially observed that our models are performing nearly as random baseline. Since we do not use the type of question when we encode input sequences, we decided to see whether adding question type as prompt shown in Figure 4 (Example B) to input sequences will improve the performance. We added `It is because` and `As a result,` as prompt for `asks-for="cause"` and `asks-for="effect"`, respectively. Interestingly, results illustrate that our model outperforms the baseline and Han and Wang (2021)'s *b-l-aug* model that is fine-tuned with the same data as ours, when question types are added as prompts to input sequences of correct and incorrect answers in the test set. We also ran a similar experiment on COPA-test (Table 4) in which adding prompt did not help with performance improvement.

Train	COPA-test ✗ Prompt	COPA-test ✓ Prompt
✗ Prompt	**79.2**	76.4
✓ Prompt	75.5	77.9

Table 4: COPA-test Accuracy ablation study results for prompt vs. no prompt.

5 Conclusion

In this work, we introduced a framework for augmenting PLMs with commonsense knowledge. Our results show that commonsense knowledge-augmented PLMs outperform the original PLMs on answering commonsense causal reasoning questions. As the next step, it would be interesting to see how the previously proposed model improvement methods or using unbiased fine-tuning datasets can potentially enhance the performance of current knowledge-augmented models.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565.

Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *Lrec*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Du-Seong Chang and Key-Sun Choi. 2004. Causal relation extraction using cue phrase and lexical pair probabilities. In *International Conference on Natural Language Processing*, pages 61–70. Springer.

Du-Seong Chang and Key-Sun Choi. 2006. Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. *Information processing & management*, 42(3):662–678.

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 306–316.

Joe Davison, Joshua Feldman, and Alexander M 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

Jesse Dunietz, Jaime G Carbonell, and Lori Levin. 2018. Deepcx: A transition-based approach for shallow semantic parsing with complex constructional triggers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1691–1701.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. 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 1808–1817.

Daniela Garcia et al. 1997. Coatis, an nlp system to locate expressions of actions connected by causality links. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 347–352. Springer.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 76–83. Association for Computational Linguistics.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Mingyue Han and Yinglin Wang. 2021. Doing good or doing right? exploring the weakness of commonsense causal reasoning models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 151–157, Online. Association for Computational Linguistics.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 619–630. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

Leo Joskowicz, T Ksiezyck, and Ralph Grishman. 1989. Deep domain models for discourse analysis. In *[1989] Proceedings. The Annual AI Systems in Government Conference*, pages 195–200. IEEE.

Randy M Kaplan and Genevieve Berry-Rogghe. 1991. Knowledge-based acquisition of causal relationships in text. *Knowledge Acquisition*, 3(3):317–337.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When choosing plausible alternatives, clever hans can be clever. *EMNLP 2019*, page 33.

Christopher SG Khoo, Jaklin Kornfilt, Robert N Oddy, and Sung Hyon Myaeng. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, 13(4):177–186.

Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. 2019. Learning to rank for plausible plausibility. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4818–4823.

Zhongyang Li, Xiao Ding, Ting Liu, J Edward Hu, and Benjamin Van Durme. 2020. Guided generation of cause and effect. IJCAI.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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 preprint arXiv:1910.10683*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *arXiv preprint arXiv:2107.12708*.

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.