Knowledge-Augmented Language Models for Cause-Effect Relation Classification

Anonymous ACL submission

Abstract

Previous studies have shown the efficacy of knowledge augmentation methods in pretrained language models. However, these methods behave differently across domains and downstream tasks. In this work, we investigate the augmentation of pretrained language models with knowledge graph data in the cause-effect relation classification and commonsense causal reasoning tasks. After automatically verbalizing triples in ATOMIC$^{20}_{20}$, a wide coverage commonsense reasoning knowledge graph, we continually pretrain BERT and evaluate the resulting model on cause-effect pair classification and answering commonsense causal reasoning questions. Our results show that a continually pretrained language model augmented with commonsense reasoning knowledge outperforms our baselines on two commonsense causal reasoning benchmarks, COPA and BCOPA-CE, and a Temporal and Causal Reasoning (TCR) dataset, without additional improvement in model architecture 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 understanding. Early methods 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 (PLMs) 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. Two examples are CausalBERT (Li et al., 2020) for guided generation of Cause and Effect and the model introduced by Guan et al. (2020) for commonsense story generation.

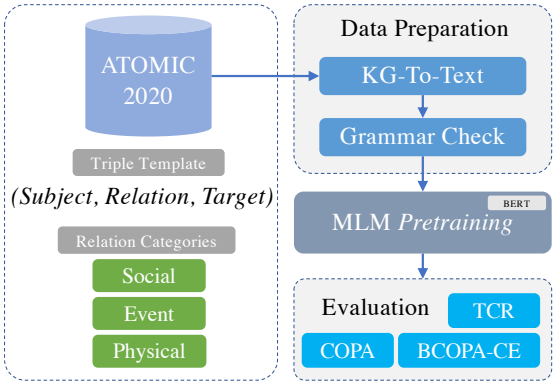

Figure 1: Overview of our proposed framework for continual pretraining PLM augmenting them with commonsense reasoning knowledge.

Motivated by the success of continual pretraining of 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 and *cause-effect* relation classification. It is worth highlighting that even though there are studies to show the efficacy of knowledge injection with continual pretraining for commonsense reasoning (Guan et al., 2020), performance of these techniques is very dependent on the domain and downstream tasks (Gururangan et al., 2020). And, to the best of our knowledge, there are limited studies on the effect of commonsense knowledge injection with knowledge graph data on cause-effect relation classification (Dalal

et al., 2021). Our contributions are as follows:

- We study performance of PLMs augmented with knowledge graph data in the less investigated cause-effect relation classification task.
- We demonstrate that a simple masked language modeling framework using automatically verbalized knowledge graph triples, without any further model improvement (e.g., new architecture or loss function) or quality enhanced data for fine-tuning, can significantly boost the performance in cause-effect pair classification.
- We publicly release our knowledge graph verbalization codes and models that are continually pretrained on cloud TPUs.

2 Method

The overview of our method is shown in Figure 1. We first convert triples in ATOMIC$^{20}_{20}$ (Hwang et al., 2021) knowledge graph to natural language texts. Then we continually pretrain BERT using Masked Language Modeling (MLM) and evaluate performance of the resulting model on different benchmarks. 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. 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.

Filtering triples: We remove all duplicates 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. This preprocessing step yields in 782,848 triples with 121,681, 177,706, and 483,461 from event, physical, and social categories, respectively.

Converting Triples: Each relation in ATOMIC$^{20}_{20}$ is associated with a human-readable template. For example, *xEffect*'s and *HasPrerequisite*'s templates

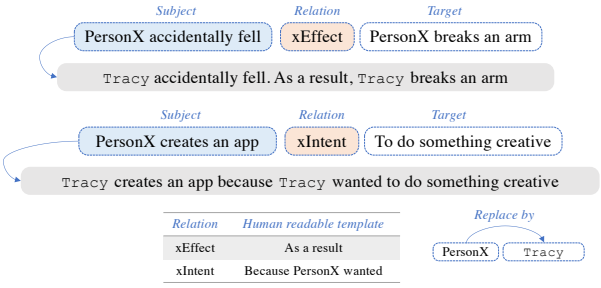

Figure 2: Examples of converting two triples in ATOMIC$^{20}_{20}$ to natural language text using human readable templates. Following Sap et al. (2019), we replace *PersonX* with a name.

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 2.

Checking Grammar: When we convert triples to natural language text, ideally we want to have grammatically correct sentences. Human readable templates provided by ATOMIC$^{20}_{20}$ are not necessarily rendered in a way to form error-free sentences when concatenated with subject and target in a triple. To address this issue, we use an open-source grammar and spell checker, LanguageTool,[1] to double-check our converted triples to ensure they do not contain obvious grammatical mistakes or spelling errors. 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). We leave investigating the effects of using other Data-To-Text and grammar-checking methods to future research.

3 Experiments

3.1 Benchmarks

We chose multiple benchmarks of commonsense causal reasoning and cause-effect relation classification to ensure we thoroughly test the effects of our newly trained models. These benchmarks include: 1) Temporal and Causal Reasoning (TCR) dataset (Ning et al., 2018), a benchmark for joint reasoning of temporal and causal relations; 2)

[1] https://tinyurl.com/yc77k3fb

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 3) BCOPA-CE (Han and Wang, 2021), a new benchmark inspired by COPA, that contains unbiased token distributions which makes it a more challenging benchmark. For COPA-related experiments, since COPA does not have a training set, we use COPA's development set for fine-tuning our models and testing them on COPA's test set (COPA-test) and BCOPA-CE. For TCR, we fine-tune and evaluate our models on train and test splits, respectively. In all experiments, we report the average performance of models using four different random seeds.

3.2 Models and Baseline

We use *bert-large-cased* pre-trained model in all experiments as our baseline. For COPA and BCOPA-CE, we convert all instances to a SWAG-formatted data (Zellers et al., 2018) and use Huggingface's *BertForMultipleChoice* –a BERT model with a multiple-choice classification head on top. And for TCR, we convert every instance by adding special tokens to input sequences as event boundaries and use the R-BERT [2] model (Wu and He, 2019). We chose R-BERT for our relation classification since it not only leverages the pretrained embeddings but also transfers information of target entities (e.g., events in a relation) through model's architecture and incorporates encodings of the targets entities. Examples of COPA and TCR are shown in Figure 3. BCOPA-CE has the same format as COPA.

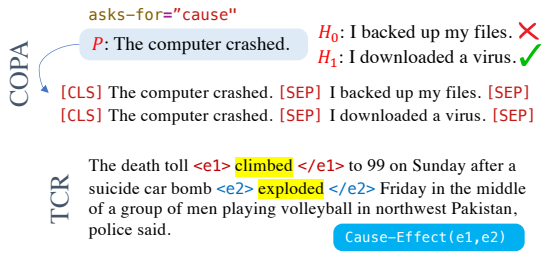

Figure 3: COPA and TCR examples. The COPA instance is converted to Multiple Choice format.

3.3 Continual Pretraining

As mentioned earlier, we use (MLM) to continually pretrain our PLM, *BERT-large-cased* (Devlin et al., 2018). We follow the same procedure as BERT to create the input data to our pretraining (e.g. number of tokens to mask in input examples). We run the pretraining by default for 10 epochs on a Google Colab TPU v2 using *PyTorch/XLA* package with a maximum sequence length of 30 and batch size of 128 and save the checkpoints at every 500 steps. [3] To avoid overfitting, we use early stopping with the patience of 3 on evaluation loss. We select the best model based on the lowest evaluation loss at the end of training. [4]

4 Results and Discussion

Results of our experiments on TCR are shown in Table 1. As can be seen, our model significantly outperforms both our baseline and the joint inference framework by Ning et al. (2018) formulated as an integer linear programming (ILP) problem.

Model	Acc (%)
Joint system (Ning et al., 2018)	77.3
BERT-large (baseline) ❊	75.0
ATOMIC-BERT-large$_{MLM}$ ❊	**91.0**

Table 1: TCR Accuracy results. ❊ Our models

Results of experiments on COPA-test are shown in Table 2. 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-centric 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 %5 which shows the effect of augmented pretrained language model with commonsense reasoning knowledge.

We further 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 that do not contain superficial cues. Results are shown in Table 3. As can be seen, our ATOMIC-BERT model significantly out-

[2] We use the following implementation of R-BERT: https://github.com/monologg/R-BERT

[3] %99.99 of ATOMIC$^{20}_{20}$ instances have 30 tokens or less.

[4] We use Huggingface's *BertForMaskedLM* implementation. All codes and trained models will be publicly available.

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-SocialIQA (Sap et al., 2019)*	80.1
BERT-large (baseline) ❊	75.9
ATOMIC-BERT-large$_{MLM}$ ❊	
- Event, Physical, Social	74.3
- Event only	80.9
Google T5-11B (Raffel et al., 2019)	94.8
DeBERTa-1.5B (He et al., 2020)	96.8

Table 2: COPA-test Accuracy results. ❊ Our models. * For a fair comparison, we report BERT-SocialIQA's average performance.

performs both the baseline and former models on Hard and Easy questions.

Model	Easy ↑	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 ❊	**89.1**	**75.9**

Table 3: COPA-test Accuracy results on Easy and Hard question subsets. ❊ Our models.

It is worth mentioning three points here. First, 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) and it shows how smaller models can be competitive and benefit from continual pretraining. 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. Third, we achieved a better performance than BERT-SocialIQA (Sap et al., 2019) while we did not use crowdsourcing or any *manual* re-writing/correction, which is expensive, for verbalizing KG triples to create our pretraining data.

4.1 BCOPA-CE: Prompt vs. No Prompt

Results of experiments on BCOPA-CE are shown in Table 4. As expected based on the results also

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 4: BCOPA-CE Accuracy results. ❊ Our models. * Base model in *b-l-** is BERT-large.

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 encoding input sequences, we decided to see whether adding question type as a prompt 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, the new 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 5) in which adding prompt did not help with performance improvement.

Train / Test	✗ Prompt	✓ Prompt
✗ Prompt	**80.9**	76.4
✓ Prompt	75.5	77.9

Table 5: COPA-test Accuracy ablation study results for prompt vs. no prompt.

5 Conclusion

We introduced a simple framework for augmenting PLMs with commonsense knowledge created by automatically verbalizing ATOMIC$^{20}_{20}$. Our results show that commonsense knowledge-augmented PLMs outperform the original PLMs on cause-effect pair classification and 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 our 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*.

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.

Dhairya Dalal, Mihael Arcan, and Paul Buitelaar. 2021. Enhancing multiple-choice question answering with causal knowledge. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 70–80.

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.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

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*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364.

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.