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\textsuperscript{2020}, 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

2.1 KG-To-Text Conversion

We convert triples in ATOMIC\textsuperscript{2020} (Hwang et al., 2021) knowledge graph to natural language texts to use them as input in our continual pretraining. Samples in ATOMIC\textsuperscript{2020} 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.](image)

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-parameters of our model. We then evaluate the model on the dev set and save the best model. We also flag all the corrected/modified samples.

\(^1\)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.
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\textsuperscript{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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMI (Roemmele et al., 2011)</td>
<td>58.8</td>
</tr>
<tr>
<td>b-l-reg (Han and Wang, 2021)</td>
<td>71.1</td>
</tr>
<tr>
<td>Google T5-base (Raffel et al., 2019)</td>
<td>71.2</td>
</tr>
<tr>
<td>BERT-large (Kavumba et al., 2019)</td>
<td>76.5</td>
</tr>
<tr>
<td>CausalBERT (Li et al., 2020)</td>
<td>78.6</td>
</tr>
<tr>
<td>BERT-large (baseline) ✪</td>
<td>75.1</td>
</tr>
<tr>
<td>ATOMIC-BERT-large_MLM ✪</td>
<td></td>
</tr>
<tr>
<td>- Event, Physical, Social</td>
<td>74.3</td>
</tr>
<tr>
<td>- Event only</td>
<td>79.2</td>
</tr>
<tr>
<td>Google T5-11B (Raffel et al., 2019)</td>
<td>94.8</td>
</tr>
<tr>
<td>DeBERTa-1.5B (He et al., 2020)</td>
<td>96.8</td>
</tr>
</tbody>
</table>

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.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Easy</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Han and Wang, 2021)</td>
<td>-</td>
<td>69.7</td>
</tr>
<tr>
<td>(Kavumba et al., 2019)</td>
<td>83.9</td>
<td>71.9</td>
</tr>
<tr>
<td>BERT-large (baseline) ♦</td>
<td>84.1</td>
<td>69.7</td>
</tr>
<tr>
<td>ATOMIC-BERT-large ♦</td>
<td>88.3</td>
<td>73.5</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>b-l-aug (Han and Wang, 2021)</td>
<td>51.1</td>
</tr>
<tr>
<td>b-l-reg (Han and Wang, 2021)</td>
<td>64.1</td>
</tr>
<tr>
<td>BERT-large (baseline) ♦</td>
<td>55.8</td>
</tr>
<tr>
<td>ATOMIC-BERT-large ♦</td>
<td>54.1</td>
</tr>
<tr>
<td>- Event, Physical, Social</td>
<td>58.1</td>
</tr>
<tr>
<td>- Event only</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>COPA-test</th>
<th>Train</th>
<th>Prompt</th>
<th>✓ Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓ Prompt</td>
<td>79.2</td>
<td>76.4</td>
</tr>
<tr>
<td></td>
<td>✓ Prompt</td>
<td>75.5</td>
<td>77.9</td>
</tr>
</tbody>
</table>

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


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


