Knowledge Augmented Language Models for Causal Question Answering ⋆

Dhairya Dalal[1][0000−0003−0279−234X]

SFI Centre for Research Training in Artificial Intelligence,
Data Science Institute, National University of Ireland - Galway
d.dalal1@nuigalway.ie

Abstract. The task of causal question answering broadly involves reasoning about causal relations and causality over a provided premise. Causal question answering can be expressed across a variety of tasks including commonsense question answering, procedural reasoning, reading comprehension, and abductive reasoning. Transformer-based pretrained language models have shown great promise across many natural language processing (NLP) applications. However, these models are reliant on distributional knowledge learned during the pretraining process and are limited in their causal reasoning capabilities. Causal knowledge, often represented as cause-effect triples in a knowledge graph, can be used to augment and improve the causal reasoning capabilities of language models. There is limited work exploring the efficacy of causal knowledge for question answering tasks. We consider the challenge of structuring causal knowledge in language models and developing a unified model that can solve a broad set of causal question answering tasks.

Keywords: causal reasoning · causal question answering · language models · causal knowledge graphs.

1 Problem Statement

Historically, research on causal reasoning in natural language processing (NLP) has primarily focused on causal relation identification and extraction. Recently, there has been an emerging interest in more complex applications of causal reasoning, especially around question answering and several new benchmark tasks were released. The task of causal question answering can be expressed across a variety of tasks. Broadly these tasks involve reasoning about causality and causal relations over a provided premise. Pretrained transformer-based language models such as BERT [4] and RoBERTa [13] have been found to be generally effective for these tasks. However, the distributional knowledge contained in these models is opaque and reliant on the quality and depth of scope of the pretraining corpus. Additionally, it is unclear the extent to which language models support causal reasoning. We hypothesize that causal knowledge can help language models represent causality and identify causal relations necessary for downstream question

⋆ With generous support from the Science Foundation Ireland Centre for Research Training in Artificial Intelligence

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3005/03paper.pdf

answering. Causal facts, often extracted from causal descriptions and expressed as cause-effect triples, can succinctly express causal knowledge. We consider the challenge of structuring causal knowledge in language models and developing a unified causal language model that can be effective across all causal reasoning tasks.

2 Importance

Causal reasoning has a long and rich history rooted in philosophy, psychology, and many other academic disciplines. Psychologists and philosophers have posited that causal reasoning is critical to our mental models of reality and that our knowledge is defined through identifying causal chains over our observations of the world [5]. In the context of natural language applications, causal reasoning can allow us to produce new knowledge from disparate observations and explore various hypotheses. For example, causal search engines in the clinical domain aim to identify causal factors that can help develop new drugs and diagnose rare medical conditions. Causal question answering systems can be used to better understand the causes of observed events and explore counterfactuals. Language models augmented with causal knowledge can be used to develop more transparent and explainable AI systems. At inference time these model could produce causal explanations of the predicted answer.

3 Related Work

There is limited historical work on causal question answering with external causal knowledge. Hassanzadeh et al. [7] and Kayesh et al. [11] consider the problem of binary question answering which poses questions about causes and effects as yes/no questions. Extracted cause-effect pairs are scored using a mixture of co-occurrence statistics and cosine-similarity scores over BERT embeddings. These scores are then evaluated against a threshold to answer yes/no for an input question. Sharp et al. [17] and Xie and Mu [21] consider the task of answer re-ranking for open-ended causal question answering. Both papers are evaluated on a set of causal questions extracted from the Yahoo! Answers corpus, which follow the patterns *What causes ...* and *What is the result of ....* [17]. Sharp et al. present three distributional similarity models (adapted Skipgram, monolingual alignment, and a convolutional neural network) to model the contextual relationship between cause and effect phrases. The answer choices are re-ranked based on the cosine similarity between extracted cause and effect vectors. Our `CausalSkipgram` model for representing causal knowledge expands upon the adapted Skipgram model presented by Sharp et al.

Next, we summarize the current causal question answering benchmark datasets. ROPES (reasoning over paragraph effects in situations) [12] is a reading comprehension dataset where the goal is to use causal relationships expressed in a background passage to answer questions about a hypothetical premise. CosmosQA [9] is a multiple-choice reading comprehension challenge where the aim is to answer

questions concerning likely causes or effects of events that require commonsense knowledge outside of the provided context. COPA (Choice of Plausible Alternatives) [6] is a multiple-choice question answering task where the goal is to identify which alternative is the likely cause or effect of a provided premise. WIQA (What if question answering) [19] is another multiple-choice task that aims to reason about the magnitude effects of perturbations to procedural descriptions of events. aNLI (Abductive Natural Language Inference) [1] is a multiple-choice task where the goal is to identify which of provided hypotheses best explains a provided context. Our preliminary work has primarily focused on the COPA and WIQA datasets as they allow the most direct evaluation of causal knowledge in the context of multiple-choice causal question answering.

Finally, we summarize sources of causal knowledge. CauseNet [8] is currently the largest publicly available knowledge graph of claimed causal facts. It contains over 11 million relations and 12 million concepts that were extracted from Wikipedia and ClueWeb [8]. ConceptNet [18], a public knowledge graph, consists of 36 relations and includes a *causes* relation. The ATOMIC knowledge [16] graph focuses on knowledge for commonsense inference. ATOMIC is organized around if-then relations that primarily describe the relations and interactions around human-centric activities. We use CauseNet as the primary source for causal knowledge in our experiments.

4 Research Questions

RQ1 **Does incorporating structured causal knowledge into language models improve performance on causal question answering tasks?**

Current model-based solutions have converged on fine-tuning pretrained language models on task-specific datasets. These approaches rely on the transferability of distributional knowledge learned during the pretraining process. To the best of our knowledge, there is no empirical research that demonstrates the efficacy of external causal knowledge in the context of causal question answering. Our work aims to establish those baselines.

RQ2 **What is the most effective way of representing causal knowledge in language models for causal question answering tasks?**

RQ2.1 How can language models be augmented with external knowledge from causal knowledge graphs for downstream causal question answering tasks?

RQ2.2 How can causal knowledge be injected into the language model during the pretraining process such that it is available as transferable distributional knowledge for downstream causal question answering tasks?

Augmenting language models with structured knowledge is an emerging area of research. Our research aims to provide a methodology for representing causal knowledge that used by language models in causal question answering tasks. We consider the strategies of knowledge augmentations and knowledge injection. The knowledge augmentation approach (*RQ2.1*) aims to train the language model to consider external knowledge provided as input features during prediction time. The knowledge injection approach (*RQ2.2*) aims to convert causal knowledge into structured distributional knowledge during pretraining process

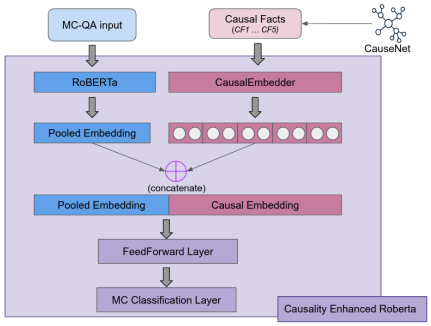

Fig. 1. Architecture of Causality Enhanced RoBERTa. The end-to-end architecture takes as input the multiple-choice question input and relevant causal facts selected from CauseNet.

to produce causality-aware language models (CALM) which would ideally support any downstream causal question answering task.

RQ3 **How do we evaluate the causal reasoning capabilities of language models in the context of question answering?**

To date, most research on causal reasoning applications in NLP focuses on task-specific model implementations. There is no comprehensive definition of causal question answering nor a unified way to evaluate a language model's cause reasoning capabilities. We hope to contextualize causal question answering as an extension of fundamental NLP problems and produce a unified benchmark (similar in spirit to GLUE (General Language Understanding Evaluation) benchmark [20]).

5 Preliminary Results

Our experiments explore the efficacy augmenting RoBERTa [13] with causal knowledge for multiple-choice question answering on the COPA and WIQA benchmark tasks. Causal facts are extracted from CauseNet [8] and selected based on the lexical overlap between the cause-effect concepts and question text. Additional details on knowledge selection and experiment results can be found in [3].

COPA consists of a premise and two alternatives. The task is to identify which alternative is most likely the cause or effect of the provided premise. Background commonsense causal knowledge is required to answer questions as there is limited lexical overlap between the premise and alternatives. WIQA consists of multiple-choice questions where the answer options (`more, less, and no effect`) describe the magnitude effect of a proposed perturbation to a procedural event. Each question has an associated procedural description consisting of a sequence of events. The question proposes a perturbation to a specific event and asks what impact that perturbation would have on another event in the procedural description.

Next, we present three strategies for representing causal knowledge to a language model. The most direct way to incorporate causal information is to append it to the end of the input text, which we call the `InputAugmentation` method. Relevant causal tuples are converted into causal statements which follow the pattern *C causes E*. `CausalSkipgram` adapts the skip-gram word embedding approach [14] to model causal pairs. The last method is `CausalKGE`, which represents causal knowledge as a knowledge graph embedding. We adapt the TransE model presented by Bordes et al. [2]. To model our causal tuples as a knowledge graph, we add the explicit relation "cause-effect" to each tuple. The modeling goal of TransE is thus to predict an effect E, given a cause C and "cause-effect" CR such that $C + CR \approx E$. A causal triple is represented by a single vector which is generated by mean pooling the head, tail, and relation vectors.

To incorporate causal embeddings with RoBERTa, we propose the `Causality Enhanced RoBERTa` neural architecture (Figure 1). This architecture is used with both the `CausalSkipgram` and `CausalKGE` embeddings. The first layer is the causality-enhanced input layer which combines the pooled embedding output of RoBERTa with the causal knowledge embeddings. For inputs where we were able to extract causal facts, the causal embedding vector is generated by concatenating and flattening all relevant causal embeddings. Up to five causal facts are selected per input. The RoBERTa pooled output is then concatenated with causal embeddings. This input is next passed into a FeedForward Network (FFN) with a hidden layer and classifier

Table 1. Accuracy on the COPA test set and COPA-BALANCED Hard set. CausalKGE improves accuracy over the RoBERTa baseline by +6.20pp on COPA Test set and +3.86pp on the COPA-Balanced Hard set.

Model	COPA Test	COPA-Balanced Hard
`RoBERTa baseline`	53.00	58.39
`+CausalSkipgram`	57.80	58.38
`+CausalKGE`	*59.20 (+6.20pp)*	62.25
`+InputAugmentation`	59.00	*62.29 (+3.86pp)*

Table 2. Accuracy of causal augmentation methods on the WIQA dataset. `InputAugmentation` achieves higher accuracy in the In-Paragrah (+3.3pp) and Out-of-Paragraph (+2.0pp) sub-categories over the current state-of-the-art QUARTET.

Model	Overall	In-Para.	Out-of-Para.	No Effect
Bert-Baseline [19]	73.80	**79.68**	56.10	89.38
QUARTET - SOTA [15]	**82.07**	73.49	65.65	**95.30**
RoBERTa baseline	67.00	64.0	42.10	92.50
+ CausalSkipgram	65.00	53.96	41.38	92.29
+ CausalKGE	74.00	71.70	55.17	93.78
+ InputAugmentation	80.00	**76.79**	**67.65**	92.43

6 Evaluation

Table 1 provides the results of our experiments on the COPA test set and the COPA-Balanced Hard set. Recent pretrained models such as BERT and RoBERTa have seen improved performance on the COPA dataset. However, Pride et al. [10] found that these models exploited superficial cues such as the token frequency in the correct answers. To mitigate this effect, Pride et al. expanded the development set to include mirror instances to balance the lexical distribution between correct and incorrect answers. This new dataset, called COPA-Balanced, also categorized the test set into easy and hard groups. The easy group consists of 190 questions where RoBERTa-Large and BERT-Large could answer correctly without the provided premise and the hard group is the remaining 310 questions. We use the COPA-Balanced development set for training and the hard category (which we will refer to as COPA-Balanced Hard) for evaluation. For the COPA test set, we were able to extract causal information from CauseNet for 32% of the questions. All three causal augmentation methods outperform the RoBERTa baseline. The `CausalKGE` and `Input Augmentation` have similar performance, improving accuracy on average by +6.0pp and +3.9pp over the RoBERTa baseline on the COPA test set and COPA-Balanced Hard set.

Table 2 provides the results for our experiments on the WIQA dataset. The current state-of-the-art for WIQA is the QUARTET model presented by Rajagopal et al. [15]. QUARTET modifies the WIQA task to include an explanation structure that identifies the supporting events from the procedural description that best explains the proposed perturbation. The supporting events come from the explanations influence graph which was selected by human annotators for each question in the WIQA dataset. QUARTET models the explanation task as a multi-task learning problem where the model must predict both the gold relevant supporting sentences and the associated impact of the perturbation for each supporting event. While our approach is -2pp less than the overall accuracy of QUARTET and we outperform QUARTET in the In-Paragraph and Out-of-Paragraph subcategories.

We were able to select causal information for 55% (1,661) of the questions in the test set, with an average of one causal tuple extracted per question. 37% of questions had two or more extracted causal tuples. The `CausalSkipgram` method was the least successful, performing worse than the RoBERTa baseline across all categories. The `CausalKGE` and `InputAugmentation` methods both improved accuracy upon the RoBERTa baseline in all categories. The `InputAugmentation` method was competitive with the QUARTET method and outperformed it in both the In-Paragraph (+3.3pp) and Out-of-Paragraph (+2.0pp) categories. We do, however, see a -3.0pp decrease in accuracy in the No Effect category. This is likely due to extraneous or irrelevant causal tuples being selected. Future work can explore improving the precision of the causal extraction process.

7 Discussion and Future Work

Our initial work validates (*RQ1*) by demonstrating the efficacy of causality enhanced language models on the COPA and WIQA question answering benchmarks. Further work will explore improving recall on causal fact selection from CauseNet and more sophisticated techniques to reduce the selection of irrelevant facts. We also plan to explore knowledge injections techniques described in *RQ2.2*. We are investigating adapting the masked-language modeling objective to predict masked causal concepts across sentence-level descriptions of causal events.

Broadly, our goal is to develop a unified causal knowledge-enhanced language model that can be effective across all causal reasoning tasks. To that extent we need to be able to define and measure the causal reasoning capabilities of the language model (*RQ3*). While COPA and WIQA are both multiple-choice question answering tasks, the causal reasoning requirements are distinct for each application. Yet we find our simple augmentation strategies are effective in both cases. This raises interesting questions about how language models are using causal knowledge and if our current tasks accurately represent causal reasoning. We hope to create a more meaningful definition of causal reasoning by exploring the causal knowledge needs of existing NLP tasks and develop new probing methods to better understand the causal reasoning capabilities of these language models. We hope this work is a stepping stone towards the more ambitious goals of general AI with causal reasoning capabilities. In order to make that jump, language models need to be able reason across semantic knowledge found on the web and in causal graphs.

8 Acknowledgments

This work has been funded with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223 and is supervised by Dr. Paul Buitelaar and Dr. Mihael Arcan.

References

1. Bhagavatula, C., Bras, R.L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S.W., Choi, Y.: Abductive commonsense reasoning. CoRR (2019)
2. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. vol. 26 (2013)
3. Dalal, D., Arcan, M., Buitelaar, P.: Enhancing multiple-choice question answering with causal knowledge. In: Proceedings of Deep Learning Inside Out (DeeLIO): The Second Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. Association for Computational Linguistics (2021)

4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL 2019 Proceedings: Human Language Technologies, Volume 1 (Long and Short Papers)
5. Goldman, A.I.: A causal theory of knowing. The Journal of Philosophy **64**(12), 357–372 (1967)
6. Gordon, A.S., Kozareva, Z., Roemmele, M.: Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In: SemEval@NAACL-HLT (2012)
7. Hassanzadeh, O., Bhattacharjya, D., Feblowitz, M., Srinivas, K., Perrone, M., Sohrabi, S., Katz, M.: Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization
8. Heindorf, S., Scholten, Y., Wachsmuth, H., Ngomo, A.C.N., Potthast, M.: Causenet: Towards a causality graph extracted from the web. In: CIKM (2020)
9. Huang, L., Bras, R.L., Bhagavatula, C., Choi, Y.: Cosmos QA: machine reading comprehension with contextual commonsense reasoning. CoRR (2019)
10. Kavumba, P., Inoue, N., Heinzerling, B., Singh, K., Reisert, P., Inui, K.: When choosing plausible alternatives, clever hans can be clever. In: Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing (2019)
11. Kayesh, H., Saiful Islam, M., Wang, J., Anirban, S., Kayes, A.S.M., Watters, P.: Answering binary causal questions: A transfer learning based approach. In: 2020 International Joint Conference on Neural Networks (IJCNN) (2020)
12. Lin, K., Tafjord, O., Clark, P., Gardner, M.: Reasoning over paragraph effects in situations. In: MRQA@EMNLP (2019)
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR (2019)
14. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (2013)
15. Rajagopal, D., Tandon, N., Clark, P., Dalvi, B., Hovy, E.: What-if I ask you to explain: Explaining the effects of perturbations in procedural text. In: Findings of the Association for Computational Linguistics: EMNLP 2020 (2020)
16. Sap, M., LeBras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N.A., Choi, Y.: Atomic: An atlas of machine commonsense for if-then reasoning (2019)
17. Sharp, R., Surdeanu, M., Jansen, P., Clark, P., Hammond, M.: Creating causal embeddings for question answering with minimal supervision. In: ACL 2016 Proceedings of Conference on Empirical Methods in Natural Language Processing
18. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI Press (2017)
19. Tandon, N., Mishra, B.D., Sakaguchi, K., Bosselut, A., Clark, P.: Wiqa: A dataset for "what if..." reasoning over procedural text. In: EMNLP/IJCNLP (2019)
20. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A multi-task benchmark and analysis platform for natural language understanding (2019), in the Proceedings of ICLR.
21. Xie, Z., Mu, F.: Distributed representation of words in cause and effect spaces. In: AAAI (2019)