Enhancing Multiple-Choice Question Answering with Causal Knowledge

Dhairya Dalal[1] and **Mihael Arcan**[2] and **Paul Buitelaar**[1,2]
[1]SFI Centre for Research and Training in Artificial Intelligence
[2]Insight SFI Research Centre for Data Analytics
Data Science Institute, National University of Ireland Galway
d.dalal1@nuigalway.ie,
{mihael.arcan,paul.buitelaar}@nuigalway.ie

Abstract

The task of causal question answering aims to reason about causes and effects over a provided real or hypothetical premise. Recent approaches have converged on using transformer-based language models to solve question answering tasks. However, pretrained language models often struggle when external knowledge is not present in the premise or when additional context is required to answer the question. To the best of our knowledge, no prior work has explored the efficacy of augmenting pretrained language models with external causal knowledge for multiple-choice causal question answering. In this paper, we present novel strategies for the representation of causal knowledge. Our empirical results demonstrate the efficacy of augmenting pretrained models with external causal knowledge. We show improved performance on the COPA (Choice of Plausible Alternatives) and WIQA (What If Reasoning Over Procedural Text) benchmark tasks. On the WIQA benchmark, our approach is competitive with the state-of-the-art and exceeds it within the evaluation subcategories of In-Paragraph and Out-of-Paragraph perturbations.

1 Introduction

Recent model-based approaches for question answering tasks have primarily focused on finetuning pretrained transformer-based language models, such as BERT (Devlin et al.) and RoBERTa (Liu et al., 2019c), on task-specific datasets. These language models have been found to contain transferable linguistic knowledge (Liu et al., 2019a) and general knowledge (Petroni et al., 2019) that are effective for most downstream natural language processing (NLP) tasks. For more complex tasks, such as causal reasoning, pretrained language models are often limited as they lack the specific external background knowledge required to effectively reason about causality.

Events

1. Pressure pushes up from inside the volcano.
2. Lava comes out of the volcano.
3. Ash clouds and rocks also come out of some volcanos.
4. The eruption lasts for a long time for some eruptions.
5. The things that come out of the volcano cause disturbances in the environment.
6. The volcano loses the built up pressure.
7. The lava and other debris stop coming out of the volcano.

Question: Suppose ***MORE*** ash clouds forming happens, how will it affect disturbances in the environment.
A. More B. Less C. No Effect

Figure 1: Example question from WIQA. The question poses an perturbation for Event 3 and asks what the implication is on Event 5.

The term causal knowledge has a long history rooted in philosophy, psychology, and many other academic disciplines (Goldman, 1967). In this paper, we will refer to causal facts and causal knowledge interchangeably. Broadly, causal knowledge captures relational knowledge between concepts, which can be useful for reasoning about causality. Causal facts are generally extracted from natural language descriptions. For example, the statement *Global warming is caused primarily by human activities such as coal-burning power plants* would yield the causal fact *factories cause global warming*. These causal facts can also be described explicitly in a knowledge base or expressed formally as triples with an explicit cause-effect relation. For ex-

Proceedings of Deep Learning Inside Out (DeeLIO):
The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 70–80
Online, June 10, 2021. ©2021 Association for Computational Linguistics

ample, the causal fact *factories cause global warming* would be expressed as the triple `(factory, cause-effect, global warming)`. As causal facts are generated from descriptions, the veracity of these facts can be questionable. Ascertaining the verisimilitude of causal knowledge is an open problem and out-of-scope for our experiments. In this paper, we explore if causal knowledge is useful for question answering and present strategies on how to enhance a pretrained language model with causal knowledge.

There is limited work on incorporating external causal knowledge to improve question answering and no prior work on using causal knowledge to improve multiple-choice question answering. The task of causal question answering aims to reason about cause and effects over a provided real or hypothetical premise. Specifically, we explore the multiple-choice formulation of this task in the context of the COPA (*Choice of Plausible Alternatives*) (Gordon et al., 2012b) and WIQA (What If Reasoning over Procedural Text) (Tandon et al., 2019) benchmark tasks. COPA and WIQA are both challenging causal reasoning tasks.

WIQA requires reasoning on hypothetical perturbations to procedural descriptions of events. Consider the example in Figure 1. To answer the hypothetical question about the downstream effect of an increase of *ash and cloud on the environment*, the model must be able to causally link **Event 3** (about *ash clouds*) to **Event 5** (*erupted materials disturb the environment*). If provided a causal fact such as (ash clouds, cause-effect, environmental disturbances), the model could make the causal association and logical leap that the magnitude of the effect is more.

COPA is another multiple-choice causal reasoning task. COPA requires external commonsense causal knowledge to answer questions about the causes and effects for a provided premise. Consider the following example from COPA:

- Premise: Air pollution in the city worsened. What was the CAUSE of this?
- Alternative 1: Factories increased their production.
- Alternative 2: Factories shut down.

Lexically, there is limited information in the premise and alternatives that the model can exploit to answer the question. To successfully answer this question, the model requires both background knowledge about factories and the ability to make causal leaps about the impact of factories on the environment. Causal facts can succinctly capture that knowledge. Consider the following claimed causal fact triples from CauseNet (Heindorf et al., 2020):

- (factory, cause-effect, pollution)
- (factory, cause-effect, air pollution)
- (production, cause-effect, pollution)

If the model was provided these facts apriori, it could reason that factories cause air pollution and the increase of production would worsen the air quality.

This paper presents empirical findings on the efficacy of augmenting pretrained models with causal facts extracted to improve multiple-choice causal question answering. Our contributions can be summarized as follows:

- We present a general method for selecting relevant causal facts from CauseNet for a provided multiple-choice question.
- We present two novel strategies for representing external causal knowledge as embeddings for downstream question answering.
- We present a novel end-to-end neural architecture that augments RoBERTa with external causal knowledge for multiple-choice question answering.

Our experiments demonstrate that augmenting pretrained models with external causal knowledge improves results over the baseline on the COPA and WIQA benchmark tasks. For the WIQA benchmark, we present findings that show causal knowledge improves RoBERTa's performance to nearly match the current state-of-the-art (SOTA) and improve upon the SOTA in specific sub-categories such as in-paragraph and out-of-paragraph reasoning.

2 Related Work

Enhancing language models with external knowledge (in the form of a knowledge graph or knowledge base) remains an open problem. Several promising strategies have emerged for injecting knowledge into large language models as part of the pretraining process. Peters et al. (2019) present the Knowledge Attention and Recontextualization

(KAR) layer which can be inserted into a neural language model architecture and used to train knowledge enhanced contextual embeddings. Liu et al. (2019b) introduce the K-BERT model which learns knowledge enabled representations from sentence trees that consist of inputs augmented with knowledge triples. Sun et al. (2020) introduce the CoLAKE model which jointly learns language and knowledge representations through pretraining on word-knowledge (WK) graphs. To the best of our knowledge, there is no prior work on enhancing language models specifically with causal knowledge.

Next we provide a summary of the question answering tasks which require causal reasoning. The task of binary causal question answering poses questions of cause and effect as yes/no questions (i.e. *Could X cause Y?*). Hassanzadeh et al. evaluate the application of cause-effect pairs extracted from Gigawords corpus for binary question answering. Kayesh et al. (2020) extends this work to automatically learn the yes/no threshold using word embeddings from BERT, RoBERTa, and other transformer-based models. Sharp et al. and Xie and Mu (2019) consider the task of answer reranking for open-ended causal questions. Both papers are evaluated on a set of causal question extracted from the Yahoo! Answers corpus which follows the patterns *What causes ...* and *What is the result of ...*. Sharp et al. present three distributional similarity models to model the contextual relationship between cause and effect phrases. Xie and Mu (2019) extend Sharp et al. by proposing methods for building causal embeddings from cause-effect phrase pairs by transferring causal relationships from the phrase-pair level to word-pair level. Our `CausalSkipgram` model for representing causal knowledge expands upon the adapted Skipgram model presented by Sharp et al..

Finally, we summarize the current approaches to causal knowledge extraction and knowledge graph population. Causal relation extraction aims to identify cause and effect phrases in various texts. The extracted cause/effect phrases can be used to populate causal knowledge bases. Recent approaches frame causal relation extraction as a structured sequence classification problem. Dasgupta et al. propose a LSTM architecture that uses word-level embeddings to predict cause and effect tags within a sentence. Li et al. (2021) present SCITE, a BiLSTM-CRF model which uses pretrained Flair embeddings and multi-headed self-attention to extract causal phrases. To date, there are few publicly available causal knowledge bases. CauseNet (Heindorf et al., 2020) is currently the largest publicly available knowledge graph of claimed causal facts. CauseNet consists of about 12 million concepts and 11.5 million relations extracted from Wikipedia and ClueWeb12 [1]. ConceptNet (Speer et al., 2017), a public knowledge graph, consists of 36 relations and includes a *causes* relation. The ATOMIC (Sap et al., 2019) knowledge base consists of 877k textual descriptions of inferential knowledge organized around event prompts and agent-centric activities. ATOMIC describes the social and commonsense knowledge of these events along nine if-then relations which describe the event's causes and effects on other agents/participants. COMET (Bosselut et al., 2019) is a language model adaptation framework that is trained on ATOMIC and ConceptNet to generate novel commonsense facts and construct robust commonsense knowledge bases. This paper uses CauseNet as its primary source for causal knowledge as it contains a broad and deep set of causal facts (including descriptions of physical processes relevant to WIQA).

3 Data

In this section, we describe the datasets used for causal knowledge extraction and our benchmark evaluation. We use CauseNet as the primary source of causal knowledge for our experiments. COPA and WIQA are the benchmark datasets used to evaluate causal knowledge on downstream multiple-choice question answering problems that require causal reasoning.

3.1 CauseNet

CauseNet consists of millions of concepts and causal relations extracted from ClueWeb12 and Wikipedia. ClueWeb12 is comprised of 733,019,372 English web pages crawled between February and March 2012 (Heindorf et al., 2020). Linguistic rules are used to generate candidate sentences that contain causal relations and a BiLSTM-CRF model is used to extract cause and effect concepts from the candidate sentences. Due to the unsupervised methodology used to populate CauseNet, the relations are presented as claimed causal relations. There are two versions of CauseNet, CauseNet-Full and CauseNet-Precision.

[1] https://lemurproject.org/clueweb12/

CauseNet-Precision is a subset of CauseNet-Full where all concepts are manually evaluated and selected to ensure high precision. CauseNet-Full consists of 11,609,890 relations and 12,186,195 concepts.

3.2 COPA (The Choice of Plausible Alternatives)

COPA was first introduced as a SemEval 2012 shared task (Gordon et al., 2012a). 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 successfully answer questions as there is limited lexical overlap between the premise and alternatives. The COPA dataset consists of 1,000 questions, broken into 500 development and 500 test questions.

Recent pretrained models such as BERT and RoBERTa have seen improved performance on the COPA dataset. However, Kavumba et al. (2019) found that these models exploited superficial cues such as the token frequency in the correct answers. To mitigate this effect, Kavumba et al. expanded the development set to include mirror instances to balance the lexical distribution between correct and incorrect answers. For each set of alternatives, the mirror instance introduces a new premise, where the previous correct alternative is now incorrect. 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.

4 Methodology

In this section, we present our methodologies for causal fact selection and causal representation. Causal facts are extracted from CauseNet using token-based retrieval heuristics. We also present three strategies for representing causal knowledge. The first strategy is input augmentation, where extracted causal facts are converted to causal statements and appended to the plain text input. The second and third strategies involve generating causal embeddings using distributed similarity and knowledge graph embedding approaches.

4.1 Causal Fact Selection

Selecting relevant causal facts for a provided input is an unresolved challenge. We extracted causal facts from CauseNet using a set of retrieval heuristics. Given the large number of concepts and causal relations ($\sim$11.5 million relations and $\sim$12 million concepts), it is computationally expensive to consider all facts during model training. To narrow down the scope of relevant facts, we consider only the question text in WIQA and the premise description in COPA.

First, we extract a list of tokens T from the input question/premise. T consists of unique words as well as unique noun phrases. Each word in the noun phrase is lower-cased and lemmatized. The normalized noun phrase is then converted to a single token by replacing spaces with underscores. Next, we generate a list of potential causal fact candidates. Since we do not know a priori which tokens correspond to causes and effects, we apply a strict filter to ensure that selected causal effects have lexical overlap with the input text. The causal fact table is queried to return all candidate facts where both c and e exist as tokens in T. The causal facts are ranked by frequency and the top five ranked candidates are selected as the final set of relevant causal facts for the input question.

4.2 Causal Knowledge Representation

4.2.1 Distributed Causal Embeddings

In this section, we present our method for modelling causality using a distributional similarity model. `CausalSkipgram` is similar to `cEmbed` presented by Sharp et al.. As mentioned in Section 2, Sharp et al. first proposed adapting the skip-gram word embedding approach (Mikolov et al., 2013) to model causal pairs. Two embeddings are learned for cause and effect concepts respectively. The effect embeddings serve as a context for the cause concepts and the cause embeddings in turn are used as a context for the effect concepts. Sharp et al. consider the cause and effect vectors separately.

`CausalSkipgram` differs from `cEmbed` in three ways. To learn word-level embeddings, `cEmbed` decomposes multi-word phrases and generates word pairs such that each word in the causal phrase is matched with each word in the effect phrase. In contrast, multi-word concepts are converted to a single token during the normalization process for `CausalSkipgram`. Thus, CausalSkipgram learns embeddings for each single token representation of cause and effect concepts. Second, we use Negative Sampling loss (Mikolov et al., 2013) to train CausalSkipgram. Finally, Sharp et al. consider the cause and effect vectors as separate features for their question answering application. Instead, we generate a single representation for each causal tuple by mean pooling the cause and effect vectors.

4.2.2 Causal Knowledge Graph Embeddings

In this section, we present CausalKGE, which represents causal knowledge as a knowledge graph embedding. We adapt the TransE model presented by Bordes et al. (Bordes et al., 2013). Given a relational triple (consisting of head h, relation r, and tail t), TransE represents entities and relations in a lower-dimensional space such that $h + r \approx t$. TransE treats knowledge graph embeddings as a link prediction problem where the goal is identify what the relation is given two nodes in the graph. TransE treats relations as translations in the embeddings space where adding a the relation vector to the head to should results in a vector that close to the tail vector representation. 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.

5 Experimental Settings

In this section, we describe how we trained our causal representations and the experimental settings for augmenting RoBERTa with causal knowledge for downstream question answering.

5.1 Causal Representation

5.1.1 CausalSkipgram

CausalSkipgram generates 256 dimensional embeddings. It takes as input a cause/effect tuple and predicts if the pair is a valid causal fact. We generate five negative examples per causal tuple by randomly matching cause and effect tokens. The samples are filtered to ensure that the generated negative sample does not exist as a valid causal fact. A dataset is generated by first combining the known causal tuples with the negative samples. The dataset is then randomly split into a train, validation, and test set following a standard 70-10-20 split ratio.

The CausalSkipgram model is trained for 100 epochs using a batch size of 256 and negative sampling loss (Mikolov et al., 2013). We use the sparse Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001 and cosine annealing to learn the learning rate. To extract an embedding for a causal tuple, we extract the hidden cause and effect concept embeddings that comprise the CausalSkipgram model. The causal tuple is then represented by a 256-dimensional vector that is generated by mean pooling the cause and effect vectors that comprise the tuple.

5.1.2 CausalKGE

CausalKGE produces 100 dimensional TransE embeddings. To train our knowledge graph embedding, we generate a dataset with negative samples following the same process as Section 5.1.1. The key difference is that our dataset consists of causal triples instead of causal tuples. We use the MKB (Sourty et al., 2020) library to train the 100-dimensional TransE embeddings for 25 epochs using the following hyperparameters: gamma value of 6, batch size of 32, negative sample of 5 examples per input. The model is trained to minimize the adversarial loss using the Adam optimizer with a learning rate of 0.001.

5.2 Causality Enhanced RoBERTa for Multiple-Choice Question Answering

In this section, we describe the model architectures and experimental settings for finetuning on the COPA and WIQA tasks.

5.2.1 Baseline

Our baseline multiple-choice question answering model is RoBERTa with a linear head for sequence classification. We use the base RoBERTa implementation and pretrained weights provided by the Huggingface library (Wolf et al.). Two separate baseline models are trained with respect to the COPA and WIQA task definitions.

The input for COPA consists of a premise p, two alternatives a_1, a_2 and a question q, which are all a sequence of tokens. The expected output is a binary value corresponding to either alternative 1 or 2. We format the text input to the RoBERTa models using the convention below, where the separator token is denoted as **<sep>**:

```
<sep>premise<sep>question<sep>
```

Model	COPA Test	COPA-Balanced Hard
RoBERTa baseline	53.00	58.39
+ *CausalSkipgram*	57.80	58.38
+ *CausalKGE*	*59.20 (+6.2%/+11.69%)*	62.25
+ *InputAugmentation*	59.00	*62.29 (+3.9%/+6%)*
Deberta Ensemble - SOTA (He et al., 2020)	**98.40**	N/A

Table 1: Accuracy on the COPA test set and COPA-BALANCED Hard set. CausalKGE improves accuracy over the RoBERTa baseline by 6.2% (absolute) and 11.69% (relative). On the COPA-BALANCED Hard, InputAgumentation improves accuracy by 3.9% (absolute) and 6% (relative) over the baseline.

```
alternative 1<sep>
alternative 2<sep>
```

WIQA entries are similarly formatted and consist of a procedural text P, which comprises of a list of events $e_1...e_n$, question q, and answer options $[a_1, a_2, a_3]$. The expected output is the softmax distribution over $[a_1, a_2, a_3]$. The procedural text is flattened into a single string which denote as below context. The WIQA input is formatted as follows:

```
<sep>context<sep>question<sep>
more<sep>less<sep>no effect<sep>
```

The inputs are then encoded using the default byte-pair encoder and passed to the base RoBERTa model. Next the pooled input representation H_1, which consists of the 768 last layer hidden-state representation of the first token of the sequence, is passed to a linear projection classification head. To encourage generalization, dropout with a probability of 0.5 is applied to the classification head as well.

This model is trained to minimize the cross-entropy loss using the AdamW optimizer (Loshchilov and Hutter, 2017) and a learning rate scheduler. We use a learning rate of 0.001 and 500 warmup steps with a weight decay of 0.01 for the scheduler. For both WIQA and COPA we use a batch size of 24 and enable 16-bit floating precision for training. The model is trained for 10 epochs on the WIQA dataset and 50 epochs on the COPA dataset (we use a higher number of epochs as COPA has fewer than 1,000 training examples). We select the checkpoint with the highest validation accuracy and use those weights for evaluation on the provided test sets.

5.2.2 Input Augmentation

The most direct way to incorporate causal information is to append them to the end of text input which we term as `InputAugmentation` method. Relevant causal tuples are converted into causal statements which follow the pattern *C causes E*. Multi-word concepts in the tuples which represented as single tokens are separated back out. For example, the tuple `(human_activity, climate_change)` would be converted into the statement *Human activity causes climate change*.

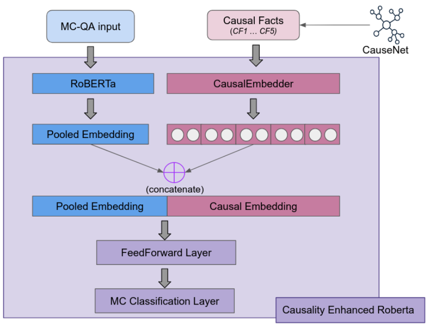

Figure 2: Architecture of Causality Enhanced RoBERTa. The architecture takes as input the multiple-choice question input and relevant causal facts selected from CauseNet.

Inputs for both COPA and WIQA follow the input formatting described in section 5.2.1 with the additional causal facts appended to the input. For example inputs for COPA are formatted using the following convention and RoBERTa specific separator token denoted as **<sep>** below.

```
<sep>premise<sep>alternative 1<sep>
alternative 2<sep>
causal statements 1...5<sep>.
```

The augmented inputs are passed into the base RoBERTa model as presented in section 5.2.1 and trained using the same experimental settings.

5.2.3 Causality Enhanced RoBERTa

To incorporate causal embeddings with RoBERTa, we propose a modified neural architecture (Figure 2). This architecture is used for both `CausalSkipgram` and `CausalKGE`, with the primary difference being the size of the causal em-

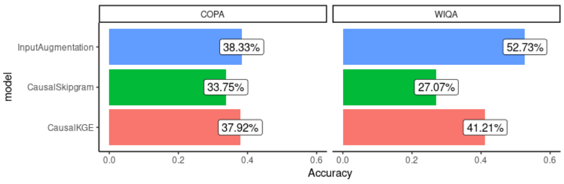

Figure 3: Performance over difficult questions that RoBERTa baseline answered incorrectly.

beddings. The first layer is the causal enhanced input layer which combines the pooled embedding output of RoBERTa with the external causal embeddings. For inputs that have extracted causal facts, a causal embedding vector is generated by concatenating and flattening all the causal embeddings. We extracted up to five causal facts per input. As a result, the combined `CausalSkipgram` embedding input is 1,280 and 500 dimensions for `CausalKGE`. Zero-valued vectors are used if causal facts are missing. The RoBERTa pooled output is then concatenated with causal embeddings. This input is further passed into a FeedForward Network (FFN) with a hidden layer and classifier. The first layer of the FFN has a hidden dimension of 512 and we apply dropout with a probability of 0.5 and ReLU (Agarap, 2018) activation to it. The second layer is the output layer with a softmax activation.

WIQA provides data that has already been split into train, validation, and test sets. We use COPA-Balanced instead of COPA. The balanced set includes mirror instances that make it more difficult for RoBERTa to exploit superficial lexical cues present in the correct answers. We randomly split the COPA-Balanced train set into a train and validation set using an 85 - 15 split.

This model is trained to minimize the cross-entropy loss using the AdamW optimizer and a learning rate scheduler. We use a learning rate of 0.001 and 500 warmup steps with a weight decay of 0.01 for the scheduler. For both WIQA and COPA, we use a batch size of 24 and enable 16-bit floating precision for training. The model is trained for 10 epochs on the WIQA dataset and 50 epochs on the COPA dataset (we use a higher number of epochs as COPA has fewer than 1,000 training examples). We select the checkpoint with the highest validation accuracy and use those weights for evaluation on the provided test sets.

6 Results

In this section, we present the results of our experiments. We find that the inclusion of causal facts improves the performance on both the COPA and WIQA datasets. Additionally, on the WIQA dataset, we observed the Augmented Input method nearly matches the SOTA in overall accuracy and exceeds the SOTA in two of the three subcategories of perturbations.

6.1 COPA Results

We present results on the COPA test set and the COPA-Balanced Hard subset in Table 1. The current state-of-the-art on COPA is DeBERTa-Large, which consists of 3.5 billion parameters. DeBERTa (He et al., 2020) modifies the BERT architecture using the disentangled attention mechanism and an enhanced mask decoder used to predict masked tokens during pretraining. While we are unable to match the performance of DeBERTa, we provide the SOTA as a fair reference for the current benchmark leader. Additionally, our augmentation methodology is not unique to RoBERTa and could be used to augment any language model with external causal information.

We were able to extract causal information from CauseNet for 32% of the questions in the test set, with an average of one causal tuple per question. About 36% of the questions with causal information had two or more extracted causal tuples.

Through the inclusion of external causal information, all three methods outperform the RoBERTa baseline. The `CausalKGE` and `Input Augmentation` have similar performance, improving accuracy by 11.69% and 6% relatively over the RoBERTa baseline on the COPA test set and COPA-Balanced Hard set. In Figure 3, we further evaluate all three methods on the subset of questions that the baseline model was unable to answer. On average, all three methods can answer 36% of questions correctly that the baseline missed, with the Input Augmentation method performing the best.

6.2 WIQA Results

Table 2 provides the results for our experiments on the WIQA dataset. The current SOTA for WIQA is the QUARTET model presented by Rajagopal et al. (Rajagopal et al., 2020). QUARTET modifies the WIQA task to include an explanation structure which identifies the supporting events from

Model	Overall	In-Para.	Out-of-Para.	No Effect
Bert-Baseline (Tandon et al., 2019)	73.80	**79.68**	56.10	89.38
QUARTET - SOTA (Tandon et al., 2019)	**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

Table 2: Accuracy of causal augmentation methods on the WIQA dataset. InputAugmentation has the best overall accuracy amongst the augmentation methods. Additionally, it achieves higher accuracy in the In-Paragrah (+3.3%) and Out-of-Paragraph (+2%) sub-categories over the current state-of-the-art QUARTET.

the procedural description that best explain the proposed perturbation. The supporting events come from the explanations influence graph which were 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. Our approach nearly matches the overall accuracy of QUARTET while outperforming 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 (+3.3%/+4.5%) and Out-of-Paragraph (+2%/+3%) categories. We do, however, see a -3% 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.

In Figure 3, we also present the results of the augmentation methods on the questions the baseline RoBERTa model was unable to answer. We find the `InputAugmentation` method can answer 52.73% of the difficult questions that the baseline failed to answer.

7 Conclusion

This paper considers the challenge of enhancing pretrained language with causal knowledge to solve multiple-choice causal question answering problems which require causal reasoning. Specifically, we evaluate our methods on the COPA and WIQA benchmark datasets. We present methods of selecting knowledge from CauseNet and three strategies for representing causal knowledge (`InputAugmentation`, `CausalSkipgram`, and `CausalKGE`). We evaluated the efficacy of enhancing RoBERTa with causal knowledge multiple-choice question answering tasks. We provide results that show improved performance over the RoBERTa baseline on both the COPA and WIQA benchmark tasks. RoBERTa with `CausalKGE` provides a 6.2%/11.69% improvement in accuracy over the baseline. RoBERTa with `Input Augmentation` posts a 3.9%/6% improvement on the COPA-Balanced Hard dataset. We also observed that on average the inclusion of causal knowledge allows RoBERTa to answer 36% of the questions the baseline was unable to answer. On WIQA, our approach is competitive with the SOTA and exceeds SOTA within specific evaluation subcategories. RoBERTa with `InputAugmentation` improves accuracy on the in-paragraph and out-of-paragraph perturbations by (+3.3%/+4.5%) and (+2%/+3%) respectively. On average, the inclusion of causal knowledge allows RoBERTa to answer 40% of the questions that the baseline was unable to answer on the WIQA test set.

Our work demonstrates that causal knowledge is valuable for causal reasoning tasks and that there are many opportunities for future work. Further work can explore improving recall on causal fact selection from CauseNet and more sophisticated techniques to reduce the selection of irrelevant facts. On the language modeling side, future work can explore generalizing the entity-based methods which inject knowledge into the pretraining process to consider explicit causal knowledge. Additionally, further work can evaluate causal knowledge in other reasoning benchmarks such as ROPES and COSMOSQA as well as other causal reasoning tasks.

8 Broader Impact

This paper focused narrowly on the efficacy of causal knowledge for multiple-choice question answering. To the best of our knowledge there are limited societal implications of this research. Broadly improvements to question answering systems have commercial value for information retrieval and other knowledge management commercial use cases. Causal reasoning is one of the outstanding challenges of AI research. We imagine that improvements to causal reasoning can have broader impacts on real-world applications. Models with causal reasoning capacities have the potential to impact applications ranging from medical drug discovery and stock market trading to scientific knowledge mining. There is also a growing interest in the regulatory space for causal systems that can conduct counterfactual reasoning around the allocation of resources to protected groups and audit policy decisions made by automated systems.

9 Acknowledgements

This work was supported by Science Foundation Ireland under grants SFI/18/CRT/6223 (Centre for Research Training in Artificial Intelligence) and SFI/12/RC/2289_P2 (Insight), co-funded by the European Regional Development Fund.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *ArXiv*, abs/1803.08375.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. *CoRR*, abs/1906.05317.

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL 2019 Proceedings: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics.

Alvin I. Goldman. 1967. A causal theory of knowing. *The Journal of Philosophy*, 64(12):357–372.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012a. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, Montréal, Canada. Association for Computational Linguistics.

Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012b. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *SemEval@NAACL-HLT*. Association for Computational Linguistics.

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 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.

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

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *CIKM*. ACM.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When choosing plausible alternatives, clever hans can be clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*.

H. Kayesh, M. Saiful Islam, J. Wang, S. Anirban, A. S. M. Kayes, and P. Watters. 2020. Answering binary causal questions: A transfer learning based approach. In *2020 International Joint Conference on Neural Networks (IJCNN)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2021. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing*, 423.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *NAACL-HLT*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019b. K-BERT: enabling language representation with knowledge graph. *CoRR*, abs/1909.07606.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Tomas Mikolov, Kai Chen, G. S. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. *CoRR*, abs/1909.04164.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Dheeraj Rajagopal, Niket Tandon, Peter Clark, Bhavana Dalvi, and Eduard Hovy. 2020. 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*. Association for Computational Linguistics.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. Creating causal embeddings for question answering with minimal supervision. In *ACL 2016 Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Raphaël Sourty, Jose G. Moreno, François-Paul Servant, and Lynda Tamine-Lechani. 2020. Knowledge base embedding by cooperative knowledge distillation. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI Press.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. Colake: Contextualized language and knowledge embedding.

Niket Tandon, B. D. Mishra, Keisuke Sakaguchi, Antoine Bosselut, and P. Clark. 2019. Wiqa: A dataset for "what if..." reasoning over procedural text. In *EMNLP/IJCNLP*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.

Zhipeng Xie and Feiteng Mu. 2019. Distributed representation of words in cause and effect spaces. In *AAAI*.

A Appendix

A.1 CauseNet Processing Details

Our approach for fact selection is identical for COPA and WIQA. spaCy [2], a python based NLP library, is used for tokenization, lemmatization, and noun-phrase extraction.

CauseNet is formatted as a JSON file where each cause-effect entry consists of relevant concepts, source sentences, and associated linguistic pattern used for causal extraction. Causal facts need to be programmatically extracted and normalized. For simplicity, we define a causal fact as a tuple consisting of cause c and effect e, where $c, e \in Concepts$. $Concepts$ in CauseNet range from single word entities to multi-word expressions (e.g. rising sea levels). We normalize multi-word concepts by first lemmatizing all its constituent words and then joining them into a single token by replacing white spaces with underscores. So the causal concept "rising sea levels" would be normalized to token "rise_sea_level". After iterating through all the entries in CauseNet and normalizing the extracted facts, we store the cause and effect tokens in a two-column causal fact table.

[2] https://spacy.io/