Relevant CommonSense Subgraphs for "What if..." Procedural Reasoning

Chen Zheng
Michigan State University
zhengc12@msu.edu

Parisa Kordjamshidi
Michigan State University
kordjams@msu.edu

Abstract

We study the challenge of learning causal reasoning over procedural text to answer "What if..." questions when external commonsense knowledge is required. We propose a novel multi-hop graph reasoning model to 1) efficiently extract a commonsense subgraph with the most relevant information from a large knowledge graph; 2) predict the causal answer by reasoning over the representations obtained from the commonsense subgraph and the contextual interactions between the questions and context. We evaluate our model on WIQA benchmark and achieve state-of-the-art performance compared to the recent models.

1 Introduction

In recent years, large-scale pre-trained language models (LMs) have made a breakthrough progress and demonstrate a high performance in many NLP tasks, including procedural text reasoning (Tandon et al., 2019; Rajagopal et al., 2020). There is a large amount of knowledge that is stored implicitly in language models that help in solving various NLP tasks (Devlin et al., 2019b). When we reason over text, sometimes, the knowledge contained in a given text is sufficient to predict the answer, as it is shown in the question 1 of Figure 1. This knowledge is directly encoded and used by LMs models (Tandon et al., 2019). However, there are many cases in which the required knowledge is not included in the procedural text itself. For example, for the question 2 in Figure 1, the information about the "nutrient" on the seeds does not exist in the procedural text. Therefore, the external commonsense knowledge is required.

There are several existing resources that contain world knowledge and commonsense. Examples are knowledge graphs (KGs) like ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019). Looking back at the question 2, we observe that through providing the external knowledge triplets (nutrient, relatedto, soil) and (soil, relatedto, seed) derived from ConceptNet, we can build an explicit reasoning chain and choose an explainable answer.

Procedural Text:
1. A plant produces a seed.
2. The seed falls to the ground.
3. The seed is buried.
4. The seed germinates.
5. A plant grows.
6. The plant produces flowers.
7. The flowers produce more seeds

Questions and Answers:
1. suppose plants will produce more seeds happens, how will it affect less plants.
(A) More (B) **Less** (C) No effect
2. suppose the soil is rich in nutrients happens, how will it affect more seeds are produced.
(A) **More** (B) Less (C) No effect
3. suppose The sun comes out happens, how will it affect less plants.
(A) More (B) Less (C) **No effect**

Figure 1: WIQA contains procedural text, and different types of questions. The bold choices are the answers.

Two challenges exist in procedural text reasoning and using external KBs. The first challenge is effectively extracting the most relevant external information and reducing the noise from the KB. The second challenge is reasoning over the extracted knowledge. Several works enhance the QA model with commonsense knowledge (Lin et al., 2019; Lv et al., 2020). However, the noisy knowledge from KG will seriously mislead the QA model in predicting the answer. Moreover, using KBs is often investigated in the tasks that perform QA directly over KB itself, such as CommonsenseQA (Talmor et al., 2019), etc. There are less sophisticated techniques proposed for using external knowledge explicitly (i.e. not through training LMs) in reading comprehension for aiding QA over text. REMNet (Huang et al., 2021) is the only work that uses commonsense for WIQA and uses a memory network to extract the external triplets to solve the first challenge. However, this work has no reasoning process over the extracted knowledge and uses a simple multi-head attention operator to predict the answer. EIGEN (Madaan et al., 2020) constructs an influence graph to find the chain of reasoning given

Findings of the Association for Computational Linguistics: ACL 2022, pages 1927 - 1933

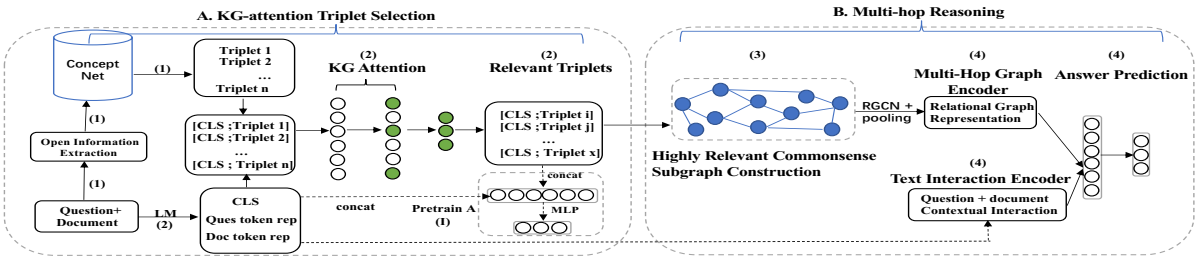

Figure 2: MRRG Model is composed of Candidate Triplet Extraction, KG Attention, Commonsense Subgraph Construction, Text encoder with contextual interaction, Graph Reasoning, and Answer prediction modules.

procedural text. However, EIGEN cannot deal with the challenge when the required knowledge is not in the given document.

To solve these two challenges, we propose a **M**ulti-hop **R**easoning network over **R**elevant CommonSense Sub**G**raphs (MRRG) for casual reasoning over procedural Text. Our motivation is to effectively and efficiently extract the most relevant information from a large KG to help procedural reasoning. First, we extract the entities, retrieve related external triplets from KG, and learn to extract the most relevant triplets to a given the procedure and question input by a novel KG attention mechanism. Then, we construct a commonsense subgraph based on the extracted KG triplets in a pipeline. We use the extracted subgraphs as a part of end-to-end QA model to help in filling the knowledge gaps in the procedure and performing multi-hop reasoning. The final model predicts the causal answer by reasoning over the contextual interaction representations over the question and the document and learning graph representations over the KB subgraphs. We evaluate our MRRG on the "what if" WIQA benchmark. MRRG model achieves SOTA and brings significant improvements compared to the existing baselines.

The contributions of our work are: **1)** We train a separate module that extracts the relevant parts of the KB given the procedure and question to avoid the noisy and inefficient usage of the information in large KBs. 2) We design an end-to-end model that uses the extracted QA-dependent KB as a subgraph to guide the reasoning over the procedural text to answer the questions. 3) Our MRRG achieves SOTA on the WIQA benchmark.

2 Model Description

2.1 Problem Formulation and Overview

Formally, the problem is to predict an answer a from a set of pre-defined answers given input question q, a document $\mathcal{C}$ which is composed of several sentences $\mathcal{C} = \{s_1, \ldots, s_n\}$, and a large knowledge graph KG.

Figure 2 shows the proposed architecture. (1) We extract the entities from question and context in preprocessing step and use them to retrieve the set of **candidate triples** from the ConceptNet. (2) We train the **KG Attention** module to extract the most relevant triplets given the procedure and question and reduce the noisy concepts from candidate triplets. (3) We augment the **commonsense subgraph** based on the relevant triplets. (4) We train a model that uses two components, the commonsense subgraph as a relational graph network and a text encoder including question and document to do **procedural reasoning**. Below, we describe the details of each module.

2.2 Candidate Triplet Extraction from KG

Given the input q and $\mathcal{C}$, we extract the contextual entities (concepts) by a open Information Extraction (OpenIE) model (Stanovsky et al., 2018). For each extracted entity t_{in}, we retrieve the relational triplets $t = (t_{in}, r, t_{out})$ from KG, where t_{out} is the concept taken from ConceptNet and r is a semantic relation type. We then apply a pre-trained Language Model, RoBERTa, to obtain the representation of each triplet: $E^t = f_{LM}([t_{in}, r, t_{out}]) \in \mathbb{R}^{3\times d}$, where f_{LM} denotes the language model operation and the triplets are given as a sequence of concepts and relations to the LM.

2.3 KG Attention

The KG attention module is shown in Figure 2-A and Figure 3. We concatenate q and $\mathcal{C}$ to form $Q = [[CLS]; q; [SEP]; \mathcal{C}]$, where [CLS] and [SEP] are special tokens in the LMs tokenizer process (Liu et al., 2019). We use RoBERTa to obtain the list of token representations $E_{[CLS]}$, E_q, and $E_{\mathcal{C}}$. $E_{[CLS]}$ is the summary representation of the question and paragraph, E_q is the list of the question tokens embeddings, and $E_{\mathcal{C}}$ is the list of the paragraph tokens embeddings output of Roberta.

Given triplet E^t that is generated based on the triplet extraction described in Section 2.2, we build a context-triplet pair $E^t_z = [E_{[CLS]}; E^t_{in}; E^t_r; E^t_{out}]$, where E^t_{in} is the representation of the head entity from text, E^t_{out} is the representation of the tail entity from KG, and E^t_r is the representation of the relation. Afterwards, we compute context-triplet pair attention and a softmax layer to output the **C**ontext-**T**riplet pairwise importance **S**core CTS. The process is computed as follows: $\text{CTS}_t = \frac{\exp\big(MLP(E^t_z)\big)}{\sum_{j=1}^{m}\exp(MLP(E^t_z))}$.

Then we choose the top-k relevant triplets with the top CTS scores and then use the relevant triplets to construct the subgraph. For each selected triplet, we obtain the triplet representation $E'^t = [E'^t_{in}, E^t_r, E'^t_{out}] \in \mathbb{R}^{3\times d}$, where $E'^t_{in} = f_{in}([CTS_t \cdot E^t_{in}; CTS_t \cdot E^t_r])$ and $E'^t_{out} = f_{out}([CTS_t \cdot E^t_{out}; CTS_t \cdot E^t_r])$. Notice that f_{in} and f_{out} are MLP layers, $[;]$ is the concatenation, and $[\cdot]$ is the scalar product.

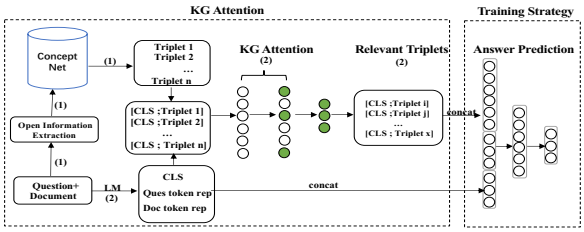

Figure 3: The architecture of training the KG Attention module.

2.4 Commonsense Subgraph Construction

We construct the subgraph G_s based on the relevant triplets from KG attention for each question and answer pair. We add more edges to the subgraph as follows: Two entities in the triplets will have an edge if a relation r in the KG exists between them. The assumption is that the augmented commonsense subgraph will contain the reasoning paths. We use E'^t_{in} and E'^t_{out} for the KG subgraph initial node representation $h^{(0)}$ which is used in RGCN formulation in Section 2.5.

2.5 Procedural Reasoning

Procedural Reasoning composes of two parts: Multi-Hop Graph Reasoning and Text Contextual Interaction Encoder.

(I) Multi-Hop Graph Reasoning: this is the Graph Reasoning part of Figure 2-B. Given the subgraph G_s, we use RGCN (Schlichtkrull et al., 2018) to learn the representations of the relational graph. RGCN learns graph representations by aggregating messages from its direct neighbors and relational semantic edges. The $(l+1)$-th layer node representation $h_i^{(l+1)}$ is updated based on the neighborhood node representations h_j^l from the l-layer multiplied by the relational matrices $W^{(l)}_{r_1}, \ldots, W^{(l)}_{r_{|R|}}$. The representation $h_i^{(l+1)}$ is computed as follows:

$$h_i^{(l+1)} = \sigma\Big(\sum_{r\in\mathcal{R}}\sum_{j\in N_i^r}\frac{1}{|N_i^r|}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}\Big),$$

where σ denotes a non-linear activation function, N_i^r represents a set that includes neighbor indices of node i under semantic relation r. Finally, we obtain the E_{G_s} after several hops of message passing.

(II) Text Contextual Interaction Encoder: We have obtained the contextual token representations $E_{[CLS]}$, E_q, and $E_{\mathcal{C}}$ in the KG attention module that described in Section 2.3. Followed by Seo et al., we utilize BiDAF style contextual interaction module to feed E_q and $E_{\mathcal{C}}$ to Context-to-Question Attention $E_{\mathcal{C}\to q} = softmax(sim(E_q^T, E_{\mathcal{C}}))E_q$ and Question-to-Context Attention $E_{q\to\mathcal{C}}$ to obtain the contextual interaction between question and context. Then we use LSTM to obtain the hidden state representations: $F_{q\to\mathcal{C}} = LSTM(E_{q\to\mathcal{C}})$, and $F_{\mathcal{C}\to q} = LSTM(E_{\mathcal{C}\to q})$.

2.6 Answer Prediction

We concatenate $E_{[CLS]}$, $F_{q\to\mathcal{C}}$, $F_{\mathcal{C}\to q}$, and the compact subgraph representation E'_{G_s} obtained from attentive pooling, and use it as the final representation: $F = [E_{[CLS]}; F_{q\to\mathcal{C}}; F_{\mathcal{C}\to q}; E'_{G_s}]$. Then we utilize a classifier MLP (F) to predict the answer. Our MRRG has two separate training modules used in a pipeline for triplet selection and procedural reasoning.

(I) Training KG Attention for Triplet Selection: Figure 3 and the left block of Figure 2 show the same triplet selection model. The architecture of Figure 2.B is taken and 3 extra MLP layers added to it for training as shown in Figure 3. The MLP is applied on the concatenation of the concatenation of $[E_{[CLS]}; E_q; E_{\mathcal{C}}; E'^t_1; \ldots; E'^t_k]$ to predict the answer. We use the cross-entropy as the loss function to train the model.

(II) Training End-to-End MRRG: After pretraining the KG attention, we keep the learned parameters and extract the most relevant concepts and construct the multi-relational commonsense subgraph G_s. We combine subgraph representation and text interaction representation as input

to train the answer prediction module by cross-entropy loss.

3 Experiments and Results

We implemented our MRRG framework using PyTorch [1]. We use a pre-trained RoBERTa (Liu et al., 2019) to encode the contextual information in the input. The maximum number of triplets is 50 and the maximum number of nodes in the graph is 100. Further details of hyper-parameters of the graph are shown in Table 3. The maximum number of words for the paragraph context is 256. For the graph construction module, we utilize *open Information Extraction* model (Stanovsky et al., 2018) from AllenNLP[2] to extract the entities. The maximum number of hops for the graph module is 3. The learning rate is $1e-5$. The model is optimized using Adam optimizer (Kingma and Ba, 2015).

3.1 Datasets

WIQA is a large dataset for "what if" causal reasoning. WIQA contains three types of questions: 1) the questions can be directly answered based on the text, called in-paragraph questions. 2) the questions require external knowledge to be answered, called out-of-paragraph questions, and 3) irrelevant causes and effects, called no-effect questions. WIQA contains 29808 training samples, 6894 development samples, 3993 test samples (test V1), and 3003 test samples (test V2).

3.2 Baseline Description

We briefly describe the most recent baselines that use the Transformer-based language model as the backbone. We separately fine-tune the BERT and RoBERTa as the first two baselines.

EIGEN (Madaan et al., 2020) is a baseline that builds an event influence graph based on a document and leverages LMs to create the chain of reasoning to predict the answer. However, EIGEN does not use any external knowledge to solve the problem.

Logic-Guided (Asai and Hajishirzi, 2020) is a baseline that combines neural networks and logic rules. Specifically, the Logic-Guided model uses logic rules including symmetry and transitivity rules to augment the training data. Moreover, the base language model uses the rules as a regularization term during training to impose the consistency between the answers of multiple questions.

RGN (Zheng and Kordjamshidi, 2021) is the recent SOTA baseline that utilizes a gating network (Zheng et al., 2020) to effectively filter out the key entities and relationships in the given document and learns the contextual representations to predict the answer. RGN does not consider the external knowledge for procedural reasoning challenges.

REM-Net (Huang et al., 2021) proposes a recursive erasure memory network to find out the causal evidence. Specifically, REM-Net refines the evidence by a recursive memory mechanism and then uses a generative model to predict the causal answer. REM-Net is the only work that uses external knowledge for WIQA. REM-Net uses the external knowledge by training an attention mechanism that considers the KG triplet representations for finding the answer. It does not explicitly select the most relevant triplets as we do, and the graph reasoning is not exploited for finding the chain of reasoning.

Models	in-para	out-of-para	no-effect	Test V1 Acc
Majority	45.46	49.47	55.0	30.66
Polarity	76.31	53.59	27.0	39.43
Adaboost (Freund and Schapire, 1995)	49.41	36.61	48.42	43.93
emphDecomp-Attn (Parikh et al., 2016)	56.31	48.56	73.42	59.48
BERT (no para) (Devlin et al., 2019a)	60.32	43.74	84.18	62.41
BERT (Tandon et al., 2019)	79.68	56.13	89.38	73.80
RoBERTa (Tandon et al., 2019)	74.55	61.29	89.47	74.77
EIGEN (Madaan et al., 2020)	73.58	64.04	90.84	76.92
REM-Net (Huang et al., 2021)	75.67	67.98	87.65	77.56
Logic-Guided (Asai and Hajishirzi, 2020)	-	-	-	78.50
RoBERTa+KG-attention Triplet Selection	72.21	64.60	89.13	75.22
***MRRG* (RoBERTa-base)**	**79.85**	**69.93**	**91.02**	**80.06**
Human	-	-	-	96.33

Table 1: Model Comparisons on WIQA test V1 dataset.

3.3 Results

Table 1 and Table 2 show the performance of MRRG on the WIQA task compared to other baselines on two different test sets V1 and V2. First, Both tables show that our proposed KG Attention triplet selection model outperforms the RoBERTa and has 3.3% improvement on the out-of-para category. Second, our MRRG achieves SOTA results compared to all baseline models. MRRG achieves the SOTA on both in-para, out-of-para, and no-effect questions in WIQA V1 and V2.

Models	in-para	out-of-para	no-effect	Test v2 Acc
Random	33.33	33.33	33.33	33.33
Majority	00.00	00.00	100.0	41.80
BERT	70.57	58.54	91.08	74.26
RoBERTa	70.69	60.20	91.11	75.34
REM-Net	70.94	63.22	91.24	76.29
REM-Net (RoBERTa-large)	76.23	69.13	92.35	80.09
QUARTET (RoBERTa-large) (Rajagopal et al., 2020)	74.49	65.65	95.30	82.07
RGN (Zheng and Kordjamshidi, 2021)	75.91	66.15	92.12	79.95
RoBERTa+KG Attention Triplet Selection	70.02	62.30	91.23	75.86
***MRRG* (RoBERTa-base)**	76.80	67.83	92.28	80.39
MRRG (RoBERTa-large)	**78.82**	**71.10**	93.53	**82.95**
Human	-	-	-	96.30

Table 2: Model Comparisons on WIQA test V2 dataset.

[1] Our code is available at https://github.com/HLR/MRRG.

[2] https://demo.allennlp.org/open-information-extraction.

Question and Document Content	RoBERTa	+Interaction	Incorporating Triplets	+KG Attention	+Graph
Question: suppose more fruit is produced happens, how will it affect MORE plants? Content: ["The seed germinates.", "The plant grows.", "The plant flowers.", "Produces fruit.", "The fruit releases seeds."] **Gold Answer: More**	x	√	(fruit, createdby, plant)	√	√
Question: suppose the soil is rich in nutrients happens, how will it affect more seeds are produced. Content: ["A plant produces a seed", "The seed falls to the ground", "The seed is buried", "The seed germinates", "A plant grows", "The plant produces flowers", "The flowers produce more seeds."] **Gold Answer: More**	x	x	(nutrient, relatedto, soil) (soil, relatedto, seed)	√	√
Question: suppose more land available happens, how will it affect less igneous rock forming. Content: ["Different kinds of rocks melt into magma", "Magma cools in the crust", "Magma goes to the surface and becomes lava", "Lava cools", "Cooled magma and lava become igneous rock."] **Gold Answer: Less**	x	x	(igneous rock, isa, rock) (land, relatedto, rock) (land, relatedto, surface) (surface, relatedto, igneous rock)	x	√

Model	# hop = 1	# hop = 2	# hop = 3
BERT	71.6%	62.5%	59.5%
RoBERTa	73.5%	63.9%	61.1%
EIGEN	78.8%	63.5%	68.3%
MRRG	**81.0%**	**72.3%**	**70.4%**

Figure 4: Left: Case study of the MRRG Framework. "+interaction" means adding the contextual interaction module. "KG ATTN" means adding the KG Attention Triplet Selection module. 'X' indicates the model failed to predict the correct answer and "✓" means the prediction was successful with the included module. Right: Comparing the results over different number of hops.

4 Analysis

4.1 Effects of Using External Knowledge

In the WIQA, all the baseline models achieve significantly lower accuracy in the out-of-para than in-para and no-effect categories. MRRG achieves SOTA in the out-of-para category because of using the highly relevant commonsense subgraphs and the combination of reasoning over text interaction and the graph reasoning modules. As is shown in table 2, the advantage of the MRRG model is reflected on out-of-para questions. MRRG improves 4.61% over REM-Net. Notice that REM-Net is the only model that utilizes external knowledge on WIQA. Figure 4 shows a case in which the "soil" and "nutrient" only appear in the question and do not exist in the text. The baseline models fail to answer this out-of-para question due to missing external knowledge. However, our model predicts the correct answer by explicitly incorporating the (nutrient, relatedto, soil), (soil, relatedto, seed) that connects the critical information between the question and document.

Ablation	Model	Dev Acc
Text only	RoBERTa-base	75.51%
Text only	+ contextual interaction	76.85%
Text only	KG Attention Triplet Selection	77.39%
Text+Graph	- semantic relation	78.31%
	GNN dim=50	79.18%
	GNN dim=100	80.30%
	GNN dim=200	79.88%

Table 3: Ablation and hyper-para. choices on WIQA. "GNN dim" is the dimension of graph representation.

4.2 Relational Reasoning and Multi-Hops

Both in-para and out-of-para question types require multiple hops of reasoning to find the answer in the WIQA. As shown in the right side of Figure 4, the MRRG model accuracy improved 2% for 1 hop, 8% for 2 hops, and 2% for 3 hops compared to EIGEN. MRRG made a sharp improvement in reasoning with multiple hops due to the relational graph reasoning and the effectiveness of the extracted commonsense subgraph. We study some cases to analyze the multi-hop reasoning and the reasoning chains. In the third case in Figure 4, the extracted relevant triplets (land, relatedto, surface), (surface, relatedto, igneous rock) construct a two-hop reasoning chain "land→surface→igneous rock" that helps MRRG to find the correct answer.

4.3 Ablation Study

Table 3 shows the ablation study results of MRRG using WIQA. Firstly, we remove the commonsense subgraph and graph network. The accuracy decreases 3.4% compared to MRRG. Second, we remove the contextual interaction module and the accuracy decreases 1.3%. In an additional experiment, we use the KG attention triplet selection module to directly predict the answer without the pipeline of constructing the subgraph and using the graph reasoning module. We show the result as KG Attention Triplet Selection in Table 3. The result shows that removing the triplet selection module decreases the accuracy by 1.8%. In the same table 3, we report results about the impact of including the relation types in the RGCN graph and the influence of changing the dimensionality of the node representations in the model.

5 Conclusion

We propose MRRG model for using external knowledge graph in reasoning over procedural text. Our model extracts a relevant subgraph for each question from the KG and uses that knowledge subgraph for answering the question. The extracted subgraph includes the reasoning path for answering the question and helps multi-hop reasoning to predict an explainable answer. We evaluate MRRG on the WIQA and achieve SOTA performance.

References

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5642–5650. Association for Computational Linguistics.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Y. Freund and R. Schapire. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT*.

Yinya Huang, Meng Fang, Xunlin Zhan, Qingxing Cao, Xiaodan Liang, and Liang Lin. 2021. Rem-net: Recursive erasure memory network for commonsense evidence refinement. In *AAAI*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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 2829–2839, Hong Kong, China. Association for Computational Linguistics.

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

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8449–8456.

Aman Madaan, Dheeraj Rajagopal, Yiming Yang, Abhilasha Ravichander, Eduard Hovy, and Shrimai Prabhumoye. 2020. Eigen: Event influence generation using pre-trained language models. *arXiv preprint arXiv:2010.11764*.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

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*, pages 3345–3355, Online. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *ICLR*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. 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 6076–6085, Hong Kong, China. Association for Computational Linguistics.

Chen Zheng, Quan Guo, and Parisa Kordjamshidi. 2020. Cross-modality relevance for reasoning on language and vision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7642–7651. Association for Computational Linguistics.

Chen Zheng and Parisa Kordjamshidi. 2021. Relational gating for "what if" reasoning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4015–4022. International Joint Conferences on Artificial Intelligence Organization. Main Track.