Fact-Driven Logical Reasoning for Machine Reading Comprehension

Siru Ouyang[1*], **Zhuosheng Zhang**[2†], **Hai Zhao**[2,3,4†]
[1]Department of Computer Science, University of Illinois Urbana-Champaign
[2]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[3]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[4]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University
siruo2@illinois.edu, zhangzs@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

Recent years have witnessed an increasing interest in training machines with reasoning ability, which deeply relies on accurately and clearly presented clue forms. The clues are usually modeled as entity-aware knowledge in existing studies. However, those entity-aware clues are primarily focused on commonsense, making them insufficient for tasks that require knowledge of temporary facts or events, particularly in logical reasoning for reading comprehension. To address this challenge, we are motivated to cover both commonsense and temporary knowledge clues hierarchically. Specifically, we propose a general formalism of knowledge units by extracting backbone constituents of the sentence, such as the subject-verb-object formed "facts". We then construct a supergraph on top of the fact units, allowing for the benefit of sentence-level (relations among fact groups) and entity-level interactions (concepts or actions inside a fact). Experimental results on logical reasoning benchmarks and dialogue modeling datasets show that our approach improves the baselines substantially, and it is general across backbone models. Code is available at https://github.com/ozyyshr/FocalReasoner.

Introduction

Training machines to understand human languages is a long-standing goal of artificial intelligence (Hermann et al. 2015), with a wide range of application scenarios such as question-answering and dialogue systems. As a well-established task, machine reading comprehension (MRC) has attracted research interest for a long time. MRC challenges machines to answer questions based on a referenced passage. (Chen, Bolton, and Manning 2016a; Sachan and Xing 2016; Seo et al. 2017; Dhingra et al. 2017; Cui et al. 2017; Song et al. 2018; Hu et al. 2019; Zhang et al. 2020; Back et al. 2020; Zhang, Yang, and Zhao 2020). There has been remarkable progress in MRC, with human-parity benchmark results reported in examination-style MRC datasets like SQuAD (Rajpurkar, Jia, and Liang 2018) and RACE (Lai et al. 2017).

However, the extent to which systems have grasped the required knowledge and reading comprehension skills remains a concern (Sugawara et al. 2020). A recent trend is to decompose comprehension ability into a collection of skills, such as span extraction, numerical reasoning, commonsense reasoning, and logical reasoning. Among these tasks, logical reasoning MRC has shown to be particularly challenging. It requires the machine to examine, analyze and critically evaluate arguments based on the relationships of facts that occur in ordinary languages, instead of simple pattern matching as focused in earlier studies (Lai et al. 2021). ReClor (Yu et al. 2020) and LogiQA (Liu et al. 2020) are two representative datasets introduced to promote the development of logical reasoning in MRC, where their questions (Figure 1) are selected from standardized exams such as GMAT and LSAT[1].

Recent studies typically exploit a pre-trained language model (PLM) as a key encoder for effective contextualized representation. However, according to diagnostic tests (Ettinger 2020; Rogers, Kovaleva, and Rumshisky 2020), PLMs like BERT (Devlin et al. 2019), despite encoding syntactic and semantic information through large-scale pre-training, they tend to struggle with understanding role reversal and pragmatic inference, as well as role-based event knowledge. Therefore, studies show that PLMs perform poorly in logical reasoning MRC tasks (Yu et al. 2020; Liu et al. 2020), as the supervision required for these tasks is rarely available during pre-training.

Following the research trend of previous reasoning tasks such as commonsense reasoning, a natural interest is to model the entity-aware relationships (e.g., *isA* or *hasA* predicates) in the passages using graph networks (Yasunaga et al. 2021; Ren and Leskovec 2020; Huang et al. 2021; Krishna, Summers, and Wies 2020; Lv et al. 2020). However, for tasks like logical reasoning of text that involve deductive/inductive/abductive reasoning (Reichertz 2004), the text usually contains hypothetical conditions that cannot be represented solely by commonsense knowledge. Therefore, these methods may insufficiently capture necessary logical units for inducing answers, since they paay little attention to non-entity, non-commonsense clues (Zhong et al. 2021). Referring to example 2 in Figure 1, previous methods may only consider commonsense knowledge such as "*Earth is a planet*", without considering the temporary fact[2] like "*comet*

[*]Work done while at SJTU.
[†]Corresponding authors.

[1]https://en.wikipedia.org/wiki/Law_School_Admission_Test
[2]A fact can be seen as an observed event (Peterson 1997).

Question	*Passage*	*Answer*
Example 1 From this we know	Xiao Wang is taller than Xiao Li, Xiao Zhao is taller than Xiao Qian, Xiao Li is shorter than Xiao Sun, and Xiao Sun is shorter than Xiao Qian.	✓ A. Xiao Li is shorter than Xiao Zhao. B. Xiao Wang is taller than Xiao Zhao. C. Xiao Sun is shorter than Xiao Wang. D. Xiao Sun is taller than Xiao Zhao.
Example 2 Which one of the following statements, most seriously weakens the argument?	.... A large enough comet colliding with Earth could have caused a cloud of dust that enshrouded the planet and cooled the climate long enough to result in the dinosaurs' demise.	A. Many other animal species from same era did not become extinct at the same time the dinosaurs did. B. It cannot be determined from dinosaur skeletons whether the animals died from the effects of a dust cloud. C. The consequences for vegetation and animals of a comet colliding with Earth are not fully understood. ✓ D. Various species of animals from the same era and similar to them in habitat and physiology did not become extinct.

Figure 1: Two examples from LogiQA and ReClor respectively are illustrated. There are arguments and relations between arguments emphasized by different colors. Keywords in questions and key options are highlighted in purple and gray.

caused dust".

To mitigate the challenge, we are motivated to bridge the gap between commonsense and temporary knowledge. First, we propose a general formalism of knowledge units by extracting backbone constituents of sentences such as the subject-verb-object formed "facts"[3]. We then develop the FOCAL REASONER, a fact-driven logical reasoning model, which builds supergraphs on top of these fact units. This approach not only captures the entity-level relations inside a fact unit within supernodes, but also enhances information flow among fact units at the sentence level through holistic supergraph modeling.

Our model is evaluated on two challenging logical reasoning benchmarks including ReClor(Yu et al. 2020), LogiQA(Liu et al. 2020), and one dialogue reasoning dataset Mutual, for verifying the effectiveness and the generalizability across different domains and question formats. To sum up, our contributions are three folds:

(i) We propose a general formalism to support representing logic units using backbone constituents of the sentences, as fine-grained knowledge carriers for logical reasoning.

(ii) We design a hierarchical fact-driven approach to construct a supergraph on top of our newly defined fact units. It models both the sentence-level (relations among fact groups) and entity-level (concepts or actions inside a fact) interactions.

(iii) Empirical studies verify the general effectiveness of our method on logical reasoning for QA and dialogues, with dramatically superior results over baselines. Analysis shows that our method can uncover complex logical structures with supergraph modeling on fact units.

Related Work

From Machine Reading Comprehension to Reasoning

Recent years have witnessed massive research on Machine Reading Comprehension (MRC) whose goal is training machines to understand human languages, which has become one of the most important areas of NLP (Chen, Bolton, and Manning 2016a; Sachan and Xing 2016; Seo et al. 2017; Dhingra et al. 2017; Cui et al. 2017; Song et al. 2018; Hu et al. 2019; Zhang et al. 2020; Back et al. 2020; Zhang, Yang, and Zhao 2020). Despite the success of MRC models on various datasets such as CNN/Daily Mail (Hermann et al. 2015), SQuAD (Rajpurkar, Jia, and Liang 2018), RACE (Lai et al. 2017) and so on, researchers began to rethink what extent does the problem been solved. Nowadays, there is massive research into the reasoning ability of machines. According to (Kaushik and Lipton 2018; Zhou et al. 2020; Chen, Bolton, and Manning 2016b), reasoning abilities can be broadly categorized into (i) commonsense reasoning (Davis and Marcus 2015; Bhagavatula et al. 2019; Talmor et al. 2019; Huang et al. 2019); (ii) numerical reasoning (Dua et al. 2019); (iii) multi-hop reasoning (Yang et al. 2018) and (iv) logical reasoning (Yu et al. 2020; Liu et al. 2020), among which logical reasoning is essential in human intelligence but has merely been delved into. Natural Language Inference (NLI) (Bowman et al. 2015; Williams, Nangia, and Bowman 2018; Nie et al. 2020) is a task closely related to logical reasoning. However, it has two obvious drawbacks in measuring logical reasoning abilities. One is that it only has three logical types which are *entailment, contradiction* and *neutral*. The other is its limitation on sentence-level reasoning. Hence, it is important for comprehensive and deeper logical reasoning abilities.

Logical Reasoning of Text

Neural and symbolic approaches have been explored in logical reasoning of text (Garcez et al. 2015, 2022; Ren and Leskovec 2020). Compared with neural methods, symbolic ones such as (Wang et al. 2021) heavily rely on data-specific patterns that are pre- and manually defined. It also suffers from error propagation and unscalable searching spaces. Our method is more related to the neural research line.

As shown in Table 1, our work mainly differs from previous work in knowledge format and modeling. Huang et al. (2021) uses discourse relations and designs the discourse-aware graph network to help logical reasoning. HGN (Chen, Zhang, and Zhao 2022) further leverages key phrases to build both inter-sentence and intra-sentence interactions in

[3]The definition follows Nakashole and Mitchell (2014). For example, those units may reflect the facts of *who did what to whom*, or *who is what*.

Model	Knowledge Format	Modeling Method	Entity-level	Sentence-level
LReasoner (Wang et al. 2021)	entities	manual rules/executor	✓	✗
DAGN (Huang et al. 2021)	EDUs	graph	✗	✓
HGN (Chen, Zhang, and Zhao 2022)	EDUs and phrases	graph	✗	✓
MERIt (Jiao et al. 2022)	entities	graph	✓	✗
Ours	fact units	supergraph	✓	✓

Table 1: Comparison between our approach FOCAL REASONER and previous methods on different aspects.

the context. The most recent work MERIt (Jiao et al. 2022) builds meta paths among logical variables (consists of entities and phrases) to model the logical relations. For knowledge formats, our proposed *fact units* can better represent both commonsense knowledge and temporary knowledge existing in the context. Previous methods either use elementary discourse units (EDUs) or phrases for temporary knowledge only or named entities to represent "isA"/"hasA" for commonsense knowledge only. For modeling methods, FOCAL REASONER is able to jointly model sentence-level and entity-level interactions via supergraphs, whereas previous methods capture either entity-level or sentence-level information with simple graph networks.

Methodology

This section presents our fact-driven approach, FOCAL REASONER. The overall architecture is shown in Figure 2. FOCAL REASONER consists of three stages. Firstly, it extracts fact units from raw texts via syntactic processing and constructs a supergraph. Then, it performs reasoning over the supergraph along with a logical fact regularization. Finally, it aggregates the learned representation to decode the right answer.

Fact Unit Extraction and Supergraph Construction

Fact Unit Extraction. The first step is to fetch triplets that constitute a fact unit. To keep the framework generic, we use a fairly simple fact unit extractor based on syntactic relations. Given a context consisting of multiple sentences, we first conduct dependency parsing on each sentence using off-the-shelf tools like SpaCy (Honnibal and Montani 2017). After that, we extract the subject, the predicate, and the object tokens to get the *"Argument-Predicate-Argument"* fact units corresponding to each sentence in the context.

Supergraph Construction. With the obtained fact units, we construct a super graph as shown in Figure 3. Concretely, the fact units are organized in the form of Levi graph (Levi 1942), which turns arguments and predicates all into nodes. An original fact unit is in the form of $F = (V, E, R)$, where V is the set of the arguments, E is the set of edges connected between arguments, and R is the relations of each edge which are predicates here. The corresponding Levi graph is denoted as $F_l = (V_L, E_L, R_L)$ where $V_L = V \cup R$, which makes the originally directly connected arguments be intermediately connected via relations. E_L is the edges connected between V_L. As for R_L, previous works such as (Marcheggiani and Titov 2017; Beck, Haffari, and Cohn 2018) designed three types of edges $R_L = \{\text{default}, \text{reverse}, \text{self}\}$ to enhance information flow. Here in our settings, we extend it into five types: *default-in, default-out, reverse-in, reverse-out, self*, corresponding to the directions of edges towards the predicates.

We construct the supergraph by making connections between fact units F_l. In particular, we take three strategies according to global information, identical concept, and co-reference information:

(i) We add a node V_g initialized with the question-option representation and connect it to all the fact unit nodes. The edge type is set as *aggregate* for better interaction.

(ii) There can be identical mentions in different sentences, resulting in repeated nodes in fact units. We connect nodes corresponding to the same non-pronoun arguments by edges with edge type *same*.

(iii) We conduct co-reference resolution on context using an off-to-shelf model[4] in order to identify arguments in fact units that refer to the same one. We add edges with type *coref* between them. The final supergraph is denoted as $S = (F_l \cup V_g, E_g)$ where E_g is the set of edges added with the previous three strategies.

Reasoning Process

Graph Reasoning. A natural way to model the supergraph is via Relational Graph Convolution Networks (Schlichtkrull et al. 2018). For a multiple-choice logical reasoning problem that consists of a context (C), a question (Q) and an option (O), we first concatenate C, Q, and O to form the input sequence. Then, the input sequence is fed to a pre-trained language model to obtain the encoded representations. We initialize the nodes with averaged hidden states of its tokens because our triplets extraction performs at the word level. For edges, we use a one-hot embedding layer to encode the relations.

Based on the relational graph convolutional network and given the initial representation h_i^0 for every node v_i, the feed-forward or the message-passing process with information control can be written as

$$h_i^{(l+1)} = \text{ReLU}\Big(\sum_{r \in R_L} \sum_{v_j \in \mathcal{N}_r(v_i)} g_q^{(l)} \frac{1}{c_{i,r}} w_r^{(l)} h_j^{(l)}\Big), \quad (1)$$

where $\mathcal{N}_r(v_i)$ denotes the neighbors of node v_i under relation r and $c_{i,r}$ is the number of those nodes. $w_r^{(l)}$ is the learnable parameters of layer l. $g_q^{(l)}$ is a gated value between 0 and 1. Through the graph encoder $F_G(.)$, we then

[4] https://github.com/huggingface/neuralcoref.

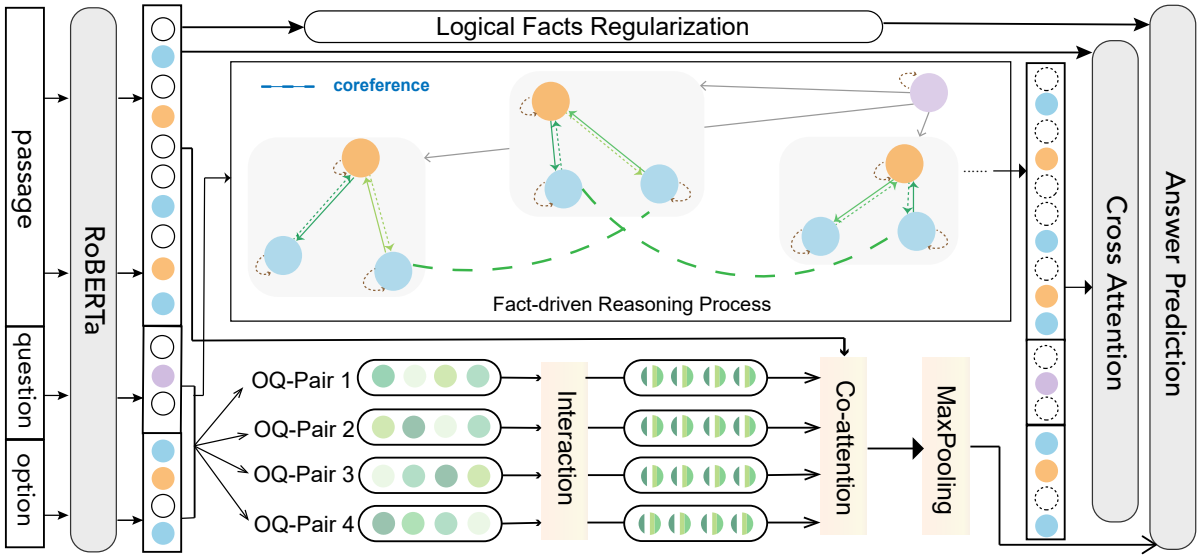

Figure 2: The framework or our model. For supergraph reasoning, in each iteration, each node selectively receives the message from the neighboring nodes to update its representation. The dashed circle means zero vector.

obtain the hidden representations of nodes in fact units as $\{h_0^F, ...h_m^F\} = F_G(\{v_{L,0}, ...v_{L,m}\}, E_L)$ where h_i^F is the node representation inside the fact unit. They are then concatenated as the representation for supernode as h_0^S. Therefore, we have $\{h_0, ...h_m\} = F_G(\{h_0^S, ...h_m^S\}, E_g)$

For node features on the supergraph, it is fused via the attention and gating mechanisms with the original representations of the context encoder H^C. We apply the attention mechanism to append the supergraph representation to the original one $\tilde{H} = \text{Attn}(H^c, K_f, V_f)$, where $\{K_f, V_f\}$ are packed from the learned representations of the supergraph, i.e., $\{h_0, ...h_m\}$, and $Attn$ is multi-head self-attention. We compute $\lambda \in [0, 1]$ to weigh the expected importance of supergraph representation of each source word $\lambda_1 = \sigma(W_\lambda \tilde{H} + U_\lambda H^C)$, where W_λ and U_λ are learnable parameters. H^C and $\tilde{H}$ are then fused for an effective representation $H = H^C + \lambda \tilde{H} \in \mathbb{R}^{4\times d}$.

Interaction. For the application to the concerned QA tasks that require reasoning, options have their inherent logical relations, which can be leveraged to aid answer prediction. Inspired by (Ran et al. 2019), we use an attention-based mechanism to gather option correlation information.

Specifically for an option O_i, the information it gets by interaction with option O_j is calculated as $O_i^{(j)} = [O_i^q - \tilde{O}_i^j; O_i^q \circ \tilde{O}_i^j]$, where O_i^q is the representation of the concatenation for the i-th option and question after the context encoder; $\tilde{O}_i^j = O_i^q Attn(O_i^q, O_j^q; v)$. Then the option-wise information is gathered to fuse the option correlation information

$$\hat{O}_i = \tanh(W_c[O_i^q; \{O_i^{(j)}\}_{i\neq j}] + b_c) \quad (2)$$

where $\mathbf{W}_c \in \mathbb{R}^{d\times 7d}$ and $b_c \in \mathbb{R}^d$.

For answer prediction, we seek to minimize the cross entropy loss by $\mathcal{L}_{ans} = -\log softmax(W_z C + b_z)_l$, where C is the combined representations of $\hat{O}$ and H.

Logical Fact Regularization. Since the subject, verb, and object in a fact should be closely related with some explicit relationships, we design a logical fact regularization technique to make the logical facts more of factual correctness. Inspired by (Bordes et al. 2013), the embedding of the tail argument should be close to the embedding of the head argument plus a relation-related vector in the hidden representation space, i.e., $v_{subject} + v_{predicate} \rightarrow v_{object}$. Specifically, given the hidden states of the sequence h_i from the Transformer encoder. Regularization is defined as

$$L_{lfr} = \sum_{k=1}^{m}(1 - \cos(h_{sub_k} + h_{pred_k}, h_{obj_k})), \quad (3)$$

where m is the total number of logical fact triplets as well as the option and k indicates the k-th fact triplet.

Training Objective

During training, the overall loss for answer prediction is $\mathcal{L} = \alpha\mathcal{L}_{ans} + \beta\mathcal{L}_{lfr}$, where α and β are two parameters. In our implementation, we set $\alpha = 1.0$ and $\beta = 0.5$.

Experiments

Experimental Setup

We conducted the experiments on three datasets. Two for specialized logical reasoning ability testing: ReClor (Yu et al. 2020) and LogiQA (Liu et al. 2020) and one for logical reasoning in dialogues: MuTual (Cui et al. 2020). We take RoBERTa-large (Liu et al. 2019) and DeBERTa-xlarge (He et al. 2020) as our backbone models for convenient comparison with previous works. We also compare our model with previous baseline models as listed in Table 1.

The model is end-to-end trained and updated by Adam (Kingma and Ba 2015) optimizer with an overall learning

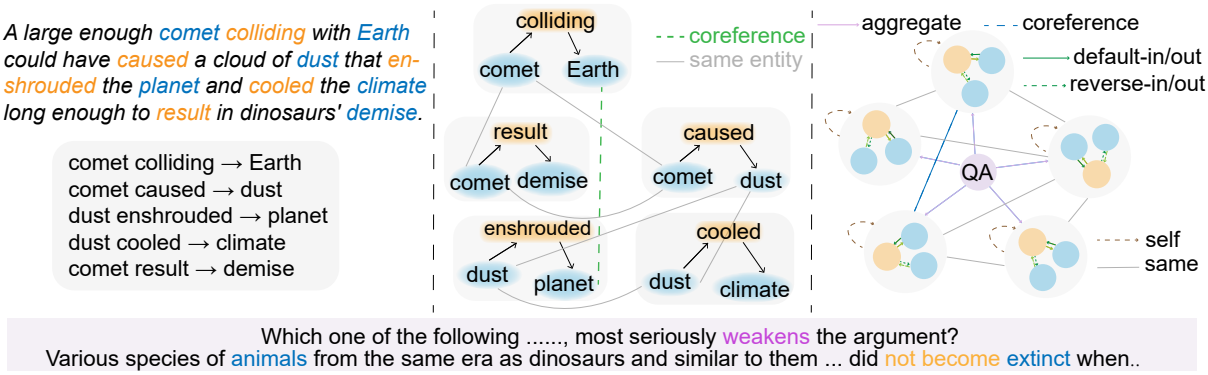

Figure 3: The process of constructing the fact chain and its corresponding Levi graph form of an example in Figure 1. Entities and relations are illustrated in their corresponding color.

Model	ReClor				LogiQA	
	Dev	Test	Test-E	Test-H	Dev	Test
Human Performance*	-	63.0	57.1	67.2	-	86.0
RoBERTa*	62.6	55.6	75.5	40.0	35.0	35.3
DAGN*	65.8	58.3	75.9	44.5	36.9	39.3
w/o data augmentation	65.2	58.2	76.1	44.1	35.5	38.7
LReasoner†	66.2	62.4	81.4	47.5	38.1	40.6
w/o data augmentation	65.2	58.3	**78.6**	42.3	-	-
MERIt†	66.8	59.6	78.1	45.2	40.0	40.3
w/o data augmentation	63.0	57.9	-	-	-	-
HGN†	66.4	58.7	77.7	43.8	40.1	39.9
FOCAL REASONER	**66.8** (↑4.2)	**58.9**(↑3.3)	77.1(↑1.6)	**44.6** (↑4.6)	**41.0** (↑6.0)	**40.3** (↑5.0)
DeBERTa*	74.4	68.9	83.4	57.5	44.4	41.5
LReasoner†	74.6	71.8	83.4	62.7	45.8	43.3
HGN†	76.0	72.3	84.5	62.7	44.9	44.2
MERIt†	78.0	73.1	86.2	**64.4**	-	-
FOCAL REASONER	**78.6** (↑4.2)	**73.3** (↑4.4)	**86.4**(↑3.0)	63.0 (↑5.5)	**47.3** (↑2.9)	**45.8** (↑4.3)

Table 2: Experimental results of our model compared with baseline models on ReClor and LogiQA dataset. Segment-1: Human performance; Segment-2: RoBERTa-based models; Segment-3: DeBERTA-based models. Test-E and Test-H denote Test-Easy and Test-Hard respectively. The results in bold are the best performance except for the human performance. * indicates that the results are taken from Yu et al. (2020) and Liu et al. (2020). Results with † are taken from their corresponding papers. Note that we are mainly comparing with previous literature without data augmentation (DA), as we hope to concentrate on our research problem on model architecture and logic relation discovery, instead of using additional tricks to bother the attention.

rate of 8e-6 for ReClor and LogiQA, and 4e-6 for MuTual. The weight decay is 0.01. We set the warm-up proportion during training to 0.1. Graph encoders are implemented using DGL, an open-source lib of python. The layer number of the graph encoder is 2 for ReClor and 3 for LogiQA. The maximum sequence length is 256 for LogiQA and MuTual, and 384 for ReClor. The model is trained for 10 epochs with a total batch size of 16 and an overall dropout rate of 0.1 on 4 NVIDIA Tesla V100 GPUs, which takes around 2 hours for ReClor and 4 hours for LogiQA.

Results

Tables 2 and 3 show the results on ReClor, LogiQA, and MuTual, respectively. All the best results are shown in bold. From the results, we have the following observations:

(i) Based on our implemented baseline model RoBERTa (basically consistent with public results), we observe dramatic improvements on both of the logical reasoning benchmarks, e.g., on ReClor test set, FOCAL REASONER achieves an absolute improvement of $+4.2\%$ on dev set and $+3.3\%$ on the test set. FOCAL REASONER also outperforms the prior best system LReasoner[5], reaching 77.05% on the EASY subset, and 44.64% on the HARD subset. The superiority of the HARD subset indicates that our method is better at solving more complex questions that rely on reasoning over complex logical clues. The performance suggests that FOCAL REASONER makes better use of logical structure inherent in the given context to perform reasoning than existing methods. Additionally, FOCAL REASONER achieves consistent improvement on DeBERTa backbone, which in-

[5]The test results are from the official leaderboard https://eval.ai/web/challenges/challenge-page/503/leaderboard/1347.

Model	Dev Set			Test Set		
	R_4@1	R_4@2	MRR	R_4@1	R_4@2	MRR
RoBERTa*	69.5	87.8	82.4	71.3	89.2	83.6
RoBERTa-MC*	69.3	88.7	82.5	68.6	88.7	82.2
FOCAL REASONER	**73.4** (↑4.1)	**90.3** (↑1.6)	**84.9** (↑2.4)	**72.7** (↑4.1)	**91.0** (↑2.3)	**84.6** (↑2.4)

Table 3: Experimental results of our model compared with baseline on MuTual dataset. * indicates that the results are taken from (Cui et al. 2020). For a fair comparison with our method, we also report the multi-choice method (RoBERTa-MC) in addition to the default Individual scoring method (RoBERTa).

Model	S	W	I	CMP	ER	P	D	R	IF	MS
RoBERTa	61.7	47.8	39.1	63.9	58.3	50.8	50.0	56.3	61.5	56.7
DAGN	63.8	46.0	39.1	69.4	57.1	53.9	46.7	62.5	62.4	56.7
FOCAL REASONER	72.3	66.4	47.8	91.7	76.2	76.9	66.7	68.8	73.5	86.7

Table 4: Accuracy on the dev set of ReClor on several representative question types. *S: Strengthen, W: Weaken, I: Implication, CMP: Conclusion/Main Point, ER: Explain or Resolve, D: Dispute, R: Role, IF: Identify a Flaw, MS: Match Structures*. All results are reported on the same PLM RoBERTa.

Model	Accuracy
FOCAL REASONER	66.8±0.13
Supergraph Reasoning	
w/o global edge	64.6±0.32
w/o co-reference edges	64.8±0.24
w/o logical fact regularization	64.2±0.12
w/o edge type	63.7±0.19
Interactions	
- interactions	65.5±0.52

Table 5: Ablation results on the ReClor dev set.

Number	ReClor		LogiQA	
	Train	Dev	Train	Dev
Fact Unit Argument	14,895	1,665	20,676	1,981
Named Entity	9,495	984	12,439	1,515

Table 6: Statistics for fact unit entities and named entities.

dicates that our method is effective on stronger baselines.

(ii) Table 4 specifies the accuracy of our model on the dev set of ReClor of different question types. Results show that our model can perform well on most of the question types, especially "Strengthen" and "Weaken", which generally involve statements such as "which of the following weakens the conclusion?" of negative semantics. This means that our model can well interpret the question type from the question statement and make the correct choice corresponding to the question, especially those with negation implications. Note that here our definition of "negation statement" is broader than traditional logic literature, which often contains words such as "not" and "never".

(iii) Our model outperforms the previous baseline models on ReClor without data augmentation (DA) and even performs on par with those with DA. On LogiQA dataset, FOCAL REASONER obtains the best performance even taking DA into consideration. Given LogiQA is a more abstractive dataset than ReClor, we may infer that fact units could indeed better capture the logical relations inside the context, which leads to broader coverage of the knowledge. Combining with hierarchical modeling methods, we can further improve the performance.

(iv) On the dialogue reasoning dataset MuTual, our model achieves substantial improvements compared with the RoBERTa-base LM.[6] Focal Reasoner is able to generalize to a different domain of datasets beyond the logical reasoning inherent in texts. This verifies our model's generalizability on other downstream reasoning task settings.

(v) For the model complexity, our method basically keeps as simple as previous models like DAGN. Our model only has 414M parameters compared with 355M in the baseline RoBERTa, and 400M in DAGN which also employs GNN. This showcases effectiveness and simplicity.

Analysis

Ablation Study

To dive into the effectiveness of different components in FOCAL REASONER, we conduct analysis by taking RoBERTa as the backbone of the ReClor dev set in Tables 5.

Supergraph Reasoning The first key component is supergraph reasoning. We ablate the global atom which is initialized with the representation of concatenation of the question and each option. and erase all the edges connected with it. The results suggest that the global atom indeed betters message propagation, leveraging performance from 64.6%

[6]Since there are no official results on RoBERTa-large LM, we use RoBERTa-base LM instead for consistency.

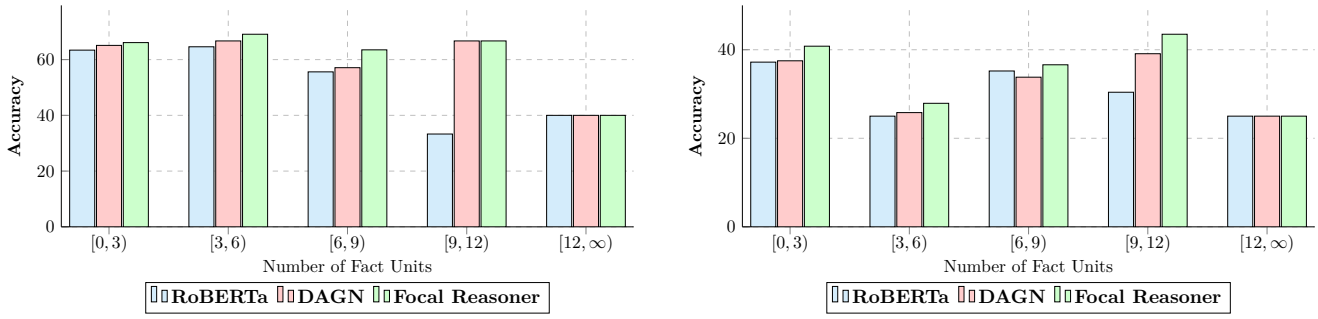

Figure 4: Accuracy of models on a number of fact units on dev set of ReClor (left) and LogiQA (right).

Model	Accuracy
FOCAL REASONER	66.8±0.13
w/ named entity only	62.8±0.26
w/ semantic role only	62.2±0.32

Table 7: Comparison about replacing fact units with commonsense knowledge such as named entities or semantic role labeling on the dev set of ReClor.

Dataset	$[0, 3)$	$[3, 6)$	$[6, 9)$	$[9, 12)$	$[12, \infty)$
ReClor	37.2%	48.6%	12.6%	0.6%	1.2%
LogiQA	47.5%	37.5%	10.9%	3.5%	0.6%

Table 8: Distribution of fact unit number on dev set.

to 66.8%. We also find that replacing the initial QA pair representation of the global atom with only question representation hurts the performance. In addition, without the logical fact regularization, the performance drops from 66.8% to 64.2%, indicating its usefulness. For edge analysis, when (i) all edges are regarded as a single type rather than the originally designed 8 types in total and (ii) co-reference edges are removed, the accuracy drops to 63.7% and 64.8%, respectively. It is proved that in our supergraph, edges link the fact units in reasonable manners, which properly uncovers the logical structures.

Interactions We further experimented with the query-option-interactions to see how it affects the performance. The results suggest that the features learned from the interaction process enhance the model. Considering that the logical relations between different options are a strong indicator of the right answer, this means that the model learns from a comparative reasoning strategy.

Comparison with Alternative Fact Units

Apart from our syntactically constructed fact units, there are two other ways in different granularities for construction. We replace fact units with named entities that are used in previous works like (Chen, Lin, and Durrett 2019). The statistics of fact units and named entities of ReClor and LogiQA are stated in Table 6, from which we can infer that there are indeed more fact units than named entities. Thus using fact units can better incorporate the logical information within the context. When replacing all the fact units with named entities and leaving the model architecture unchanged, we can see from Table 7 that it significantly decreases the performance. We also explore using semantic role labeling (SRL) in a similar way as in (Zhong et al. 2020). SRL, leveraging much more complex information as well as computation complexity, fails to achieve performance as good as our original fact unit.

Influence of Scale of Fact Units

To inspect the effects of the number of fact units, we split the original dev set of ReClor and LogiQA into 5 subsets. The statistics of the fact unit distribution on the datasets are shown in Table 8. The numbers of fact units for most contexts in ReClor and LogiQA are in $[3, 6)$ and $[0, 3)$, respectively. Comparing the accuracies of RoBERTa-large baseline, prior SOTA LReasoner and our proposed FOCAL REASONER in Figure 4, our model outperforms baseline models on all the divided subsets, which demonstrates the effectiveness and robustness of our proposed method. Specifically, for ReClor, the performance of FOCAL REASONER becomes more evident when the number of fact units locates in $[6, 9)$, while for LogiQA, FOCAL REASONER works better when the number of fact units locates in $[0, 3)$ and $[9, 12)$. The reason may lie in the difference in style of the two datasets. However, all the models including ours struggle when the number of fact units is above certain thresholds, i.e., the logical structure is more complicated, calling for better mechanisms to handle in the future.

Conclusions

In this work, we propose extracting a general form called "fact unit" to cover both commonsense and temporary knowledge units for logical reasoning. Our proposed FOCAL REASONER not only better uncovers the logical structures within the context but also better captures the logical interactions between context and options. Experimental results verify the effectiveness of our method.

Acknowledgements

This paper was partially supported by the Joint Research Project of Yangtze River Delta Science and Technology Innovation Community (No. 2022CSJGG1400).

References

Back, S.; Chinthakindi, S. C.; Kedia, A.; Lee, H.; and Choo, J. 2020. NeurQuRI: Neural Question Requirement Inspector for Answerability Prediction in Machine Reading Comprehension. In *ICLR*.

Beck, D.; Haffari, G.; and Cohn, T. 2018. Graph-to-Sequence Learning using Gated Graph Neural Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 273–283. Melbourne, Australia: Association for Computational Linguistics.

Bhagavatula, C.; Le Bras, R.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; Yih, W.-t.; and Choi, Y. 2019. Abductive Commonsense Reasoning. In *ICLR*.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-relational Data. In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., *NeurIPS*, volume 26.

Bowman, S.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*, 632–642.

Chen, D.; Bolton, J.; and Manning, C. D. 2016a. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *ACL*, 2358–2367.

Chen, D.; Bolton, J.; and Manning, C. D. 2016b. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2358–2367. Berlin, Germany: Association for Computational Linguistics.

Chen, J.; Lin, S.-T.; and Durrett, G. 2019. Multi-hop Question Answering via Reasoning Chains. *ArXiv*, abs/1910.02610.

Chen, J.; Zhang, Z.; and Zhao, H. 2022. Modeling Hierarchical Reasoning Chains by Linking Discourse Units and Key Phrases for Reading Comprehension. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1467–1479. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Cui, L.; Wu, Y.; Liu, S.; Zhang, Y.; and Zhou, M. 2020. MuTual: A Dataset for Multi-Turn Dialogue Reasoning. In *ACL*.

Cui, Y.; Chen, Z.; Wei, S.; Wang, S.; Liu, T.; and Hu, G. 2017. Attention-over-Attention Neural Networks for Reading Comprehension. In *ACL*, 593–602.

Davis, E.; and Marcus, G. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9): 92–103.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dhingra, B.; Liu, H.; Yang, Z.; Cohen, W.; and Salakhutdinov, R. 2017. Gated-Attention Readers for Text Comprehension. In *ACL*, 1832–1846.

Dua, D.; Wang, Y.; Dasigi, P.; Stanovsky, G.; Singh, S.; and Gardner, M. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *NAACL-HLT*, 2368–2378.

Ettinger, A. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *TACL*, 8: 34–48.

Garcez, A. d.; Bader, S.; Bowman, H.; Lamb, L. C.; de Penning, L.; Illuminoo, B.; Poon, H.; and Gerson Zaverucha, C. 2022. Neural-symbolic learning and reasoning: A survey and interpretation. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342: 1.

Garcez, A. d.; Besold, T. R.; De Raedt, L.; Földiak, P.; Hitzler, P.; Icard, T.; Kühnberger, K.-U.; Lamb, L. C.; Miikkulainen, R.; and Silver, D. L. 2015. Neural-symbolic learning and reasoning: contributions and challenges. In *2015 AAAI Spring Symposium Series*.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hermann, K. M.; Kočiskỳ, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *NeurIPS*, 1693–1701.

Honnibal, M.; and Montani, I. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hu, M.; Wei, F.; Peng, Y.; Huang, Z.; Yang, N.; and Li, D. 2019. Read+ verify: Machine reading comprehension with unanswerable questions. In *AAAI*, volume 33, 6529–6537.

Huang, L.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *EMNLP-IJCNLP*, 2391–2401.

Huang, Y.; Fang, M.; Cao, Y.; Wang, L.; and Liang, X. 2021. DAGN: Discourse-Aware Graph Network for Logical Reasoning. In *NAACL*.

Jiao, F.; Guo, Y.; Song, X.; and Nie, L. 2022. MERIt: Meta-Path Guided Contrastive Learning for Logical Reasoning. *arXiv preprint arXiv:2203.00357*.

Kaushik, D.; and Lipton, Z. C. 2018. How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 5010–5015. Brussels, Belgium: Association for Computational Linguistics.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *ICLR*.

Krishna, S.; Summers, A. J.; and Wies, T. 2020. Local reasoning for global graph properties. In *European Symposium on Programming*, 308–335. Springer, Cham.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 785–794. Copenhagen, Denmark: Association for Computational Linguistics.

Lai, Y.; Zhang, C.; Feng, Y.; Huang, Q.; and Zhao, D. 2021. Why Machine Reading Comprehension Models Learn Shortcuts? In *Findings of ACL-IJCNLP 2021*, 989–1002.

Levi, F. W. 1942. *Finite geometrical systems: six public lectues delivered in February, 1940, at the University of Calcutta*.

Liu, J.; Cui, L.; Liu, H.; Huang, D.; Wang, Y.; and Zhang, Y. 2020. LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. In Bessiere, C., ed., *IJCAI-20*, 3622–3628. Main track.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, arXiv:1907.11692.

Lv, S.; Guo, D.; Xu, J.; Tang, D.; Duan, N.; Gong, M.; Shou, L.; Jiang, D.; Cao, G.; and Hu, S. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *AAAI*, volume 34, 8449–8456.

Marcheggiani, D.; and Titov, I. 2017. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1506–1515. Copenhagen, Denmark: Association for Computational Linguistics.

Nakashole, N.; and Mitchell, T. 2014. Language-aware truth assessment of fact candidates. In *ACL*, 1009–1019.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *ACL*, 4885–4901.

Peterson, P. L. 1997. *Fact proposition event*, volume 66.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *ACL*, 784–789.

Ran, Q.; Li, P.; Hu, W.; and Zhou, J. 2019. Option Comparison Network for Multiple-choice Reading Comprehension. *arXiv preprint arXiv:1903.03033*.

Reichertz, J. 2004. 4.3 Abduction, deduction and induction in qualitative research. *A Companion to*, 159.

Ren, H.; and Leskovec, J. 2020. Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs. *NeurIPS*, 33.

Rogers, A.; Kovaleva, O.; and Rumshisky, A. 2020. A primer in bertology: What we know about how bert works. *TACL*, 8: 842–866.

Sachan, M.; and Xing, E. 2016. Machine comprehension using rich semantic representations. In *ACL*, 486–492.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; vanden Berg, R.; Titov, I.; and Welling, M. 2018. Modeling Relational Data with Graph Convolutional Networks. In Gangemi, A.; Navigli, R.; Vidal, M.-E.; Hitzler, P.; Troncy, R.; Hollink, L.; Tordai, A.; and Alam, M., eds., *The Semantic Web*, 593–607.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2017. Bidirectional Attention Flow for Machine Comprehension. In *ICLR 2017*.

Song, L.; Wang, Z.; Yu, M.; Zhang, Y.; Florian, R.; and Gildea, D. 2018. Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks. *arXiv preprint arXiv:1809.02040*.

Sugawara, S.; Stenetorp, P.; Inui, K.; and Aizawa, A. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *AAAI*, volume 34, 8918–8927.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *NAACL-HLT (1)*.

Wang, S.; Zhong, W.; Tang, D.; Wei, Z.; Fan, Z.; Jiang, D.; Zhou, M.; and Duan, N. 2021. Logic-Driven Context Extension and Data Augmentation for Logical Reasoning of Text. *arXiv preprint arXiv:2105.03659*.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL-HLT*, 1112–1122.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *EMNLP*, 2369–2380.

Yasunaga, M.; Ren, H.; Bosselut, A.; Liang, P.; and Leskovec, J. 2021. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In *NAACL*.

Yu, W.; Jiang, Z.; Dong, Y.; and Feng, J. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *ICLR*.

Zhang, Z.; Wu, Y.; Zhou, J.; Duan, S.; Zhao, H.; and Wang, R. 2020. SG-Net: Syntax-Guided Machine Reading Comprehension. In *AAAI*.

Zhang, Z.; Yang, J.; and Zhao, H. 2020. Retrospective Reader for Machine Reading Comprehension. *arXiv preprint arXiv:2001.09694*.

Zhong, W.; Wang, S.; Tang, D.; Xu, Z.; Guo, D.; Wang, J.; Yin, J.; Zhou, M.; and Duan, N. 2021. AR-LSAT: Investigating Analytical Reasoning of Text. *arXiv e-prints*, arXiv:2104.06598.

Zhong, W.; Xu, J.; Tang, D.; Xu, Z.; Duan, N.; Zhou, M.; Wang, J.; and Yin, J. 2020. Reasoning Over Semantic-Level Graph for Fact Checking. In *ACL*.

Zhou, M.; Duan, N.; Liu, S.; and Shum, H.-Y. 2020. Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering*, 6(3): 275–290.