Fact-driven Logical Reasoning

Sihu Ouyang, Zhuosheng Zhang, Hai Zhao

1 Department of Computer Science and Engineering, Shanghai Jiao Tong University
2 Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University
3 MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{oysr0926, zhangzs}@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract
Logical reasoning, which is closely related to human cognition, is of vital importance in human’s understanding of texts. Recent years have witnessed increasing attentions on machine’s logical reasoning abilities. However, previous studies commonly apply ad-hoc methods to model pre-defined relation patterns, such as linking named entities, which only considers global knowledge components that are related to commonsense, without local perception of complete facts or events. Such methodology is obviously insufficient to deal with complicated logical structures. Therefore, we argue that the natural logic units would be the group of backbone constituents of the sentence such as the subject-verb-object formed “facts”, covering both global and local knowledge pieces that are necessary as the basis for logical reasoning. Beyond building the ad-hoc graphs, we propose a more general and convenient fact-driven approach to construct a supergraph on top of our newly defined fact units, and enhance the supergraph with further explicit guidance of local question and option interactions. Experiments on two challenging logical reasoning benchmark datasets, ReClor and LogiQA, show that our proposed model, FOCAL REASONER, outperforms the baseline models dramatically. It can also be smoothly applied to other downstream tasks such as MuTual, a dialogue reasoning dataset, achieving competitive results.

1 Introduction
Logical reasoning is one of the most important skills of human intelligence [1], which accounts for human intuition about entailment of sentences and reflects the semantic relations between sentential constituents such as determiner, noun, adjective, adverb, preposition, and verb phrases [2]. Recently, there is a surging trend of research into logical reasoning ability, among which ReClor [3] and LogiQA [1] are two representative datasets introduced to promote the development of logical reasoning, where logical reasoning questions are selected from standardized exams such as GMAT and LSAT, requiring models to read and comprehend the complicated logical relationships. Similar to the standard question-answering (QA)-based MRC tasks in form, our concerned logical reasoning QA tasks contain three elements: passage, question and the candidate options as examples shown in Figure [1]

The major challenge of logical reasoning is to uncover logical structures, and reasoning with the candidate options and questions to predict the correct answer. However, it is difficult for PrLMs to capture the logical structure inherent in the texts since logical supervision is rarely available during

* Equal contribution. †Corresponding author. This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100), Key Projects of National Natural Science Foundation of China (U1836222 and 61733011), Huawei-SJTU long term AI project, Cutting-edge Machine Reading Comprehension and Language Model.

Preprint. Under review.
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.

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 entities and relations between entities. Both are emphasized by different colors: entities, relations. Key words in questions are highlighted in blue. Key options are highlighted in gray.

In terms of the concerned logical reasoning QA tasks, existing studies mainly focus on the understanding of the passage, with little attention paid to the question and candidate options. Especially when taking the PrLMs as the model backbone, an important concern is the degraded ability to recognize the negations [13, 14], which would easily fail in the understanding of questions and options for the logical reasoning tasks. For one thing, there might be negated expressions in questions, such as “weakens the argument” in Example 2 in Figure 1. For another, some of the answer options may contradict to each other.

In this work, we propose a fact-driven logical reasoning model, called FOCAL REASONER, which builds supergraphs on top of fact units as the basis for logical reasoning, to capture both global connections between facts and the local concepts or actions inside the fact. In addition, we strengthen our model by the question-option-aware interaction. Specifically, we explicitly reformulate questions with negation expressions. The reformulated questions and options are regarded as global vertex and option vertices, in parallel with fact vertices, all of which are interacted in our supergraph for deep interaction. Based on the enhanced features, our model predicts answers to logical questions after correlation with the question and options. FOCAL REASONER is evaluated on
two challenging logical reasoning benchmark datasets including LogiQA, ReClor, and one dialogue reasoning dataset Mutual for generalizability, achieving new state-of-the-art results.

2 Related Work

Machine Reading Comprehension Recent years have witnessed massive researches on Machine Reading Comprehension (MRC), which has become one of the most important areas of NLP. Despite the success of MRC models on various datasets such as CNN/Daily Mail [15], SQuAD [16], RACE [17] and so on, researchers began to rethink to what extent does the problem be solved. Nowadays, there is an increasing trend of research into the reasoning ability of machines. According to [18, 19, 20], reasoning abilities can be broadly categorized into (1) commonsense reasoning [21, 22, 23, 24]; (2) numerical reasoning [25]; (3) multi-hop reasoning [26] and (4) logical reasoning [1, 3], among which logical reasoning is essential in human intelligence but has merely been delved into. Natural Language Inference (NLI) [27, 28, 29] 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 to research more comprehensive and deep logical reasoning abilities.

Logical Reasoning in MRC There are two main kinds of features in language data that would be the necessary basis for logical reasoning: 1) knowledge: global facts that keep consistency regardless of the context, such as commonsense, mostly derived from named entities; 2) non-knowledge: local facts or events that may be sensitive to the context, mostly derived from linguistics. Existing works have made progress in improving logical reasoning ability [4, 5, 6, 7, 8, 30], however, these approaches are barely satisfactory as they mostly focus on the global facts such as typical entity or sentence-level relations and use ad-hoc graphs to model them, which are obviously not sufficient. In this work, we strengthen the basis for logical reasoning by unifying both types of the features as facts. Different from previous studies that focus on the knowledge components, we propose a fact-driven logical reasoning framework that builds supergraphs on top of fact units to capture both global connections between entity-aware facts and the local concepts or events inside the fact.
3 Approaches

In this section, we will describe our method in detail. The overall architecture of the model is shown in Figure 3. We first construct a fact chain from the raw text based on the logical triplets extracted. Then we conduct reasoning over the fact chain with question-option guided approaches to learn and update the features, which are further incorporated in answer prediction.

3.1 Preprocessing

Triplets Extraction. Triplets can well represent the logical facts inherent in the context. To keep the framework generic, we use a fairly simple fact extractor based on the dependency parsing of each sentence. After that, we extract the subject, the predicate, and the object tokens to get the "Entity-Predicate-Entity" triplets corresponding to each sentence in the context.

Question Reformulation. The questions in logical reasoning vary. We believe that handling with question type may compensate for the weakness of PrLM in dealing with negations, thus conducing to the prediction performance. The negation detection is composed of two parts. Firstly, we use TextBlob to do sentiment analysis. If the question is detected to be negative, we directly consider the question to have negations. Secondly, if any pre-defined negative words are among or immediately before the extracted "NP" or "VP", we identify it as a negation connective. And the pre-defined negative words include {"not", "n't", "unable", "no", "few", "little", "neither", "none of "}.

3.2 Supergraph Construction

With these fact triplets obtained from 3.1, we are able to construct the supergraph. Firstly, the fact units are organized in the form of Levi graph [31], which turns entities and predicates all into nodes. An original fact unit in the form of $F = (V, E, R)$, where $V$ is the set of the entities, $E$ is the set of edges connected between entities, 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 entities be intermediately connected via relations. As for $R_L$, previous works such as [32, 33] designed three types of edges $R_L = \{default, reverse, self\}$ 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 edges towards the predicates.

After getting fact units $F_i$, we extract the coreference relations $E_C$ among entities in fact units using Huggingface neuralcoref. Fact units with the same pronoun references are connected using undirected edges. To better encode the structure information of questions and options (details are specified in Section 3.3.2), a global atom is added and connected with all the other atoms. The final supergraph is denoted as $S = (F_i, E_C)$. An example is shown in Figure 4.

3.3 Encoder

3.3.1 Context Encoder

We initialize our context encoder $F_C(\cdot)$ with a transformer-based pre-trained encoder, i.e., RoBERTa-large [34]. Question, context and option are concatenated and then fed into the encoder. If the question is detected to contain negative meanings, we add a special token <neg>. In a whole, we get the hidden representation as following:

$$\{h_{c,0}, ..., h_{c,l_c+1}, h_{q,0}, ..., h_{q,l_q+1}, h_{o,0}, ..., h_{o,l_o+1}\} = F_C(\{x_{c,0}, ..., x_{c,l_c+1}, x_{q,0}, ..., x_{q,l_q}, x_{o,1}, ..., x_{o,l_o+1}\})$$

(1)

where $x_{c,0} = <s>$, $x_{c,l_c+1} = <</s>$, $x_{q,0} = <pos>/<neg>$ and $h_i \in \mathbb{R}^d$, $d$ is the hidden size.

3.3.2 Super Graph Encoder

Graph Initialization. We first utilize $F_C(\cdot)$ to encode each token in nodes $V_L$, and then we use the averaged hidden state as the initial representation of the original word of each atom, because

---

https://github.com/sloria/TextBlob
https://github.com/huggingface/neuralcoref
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. 

Various species of animals from the same era as dinosaurs and similar to them ... did not become extinct when the dinosaurs did.

Which one of the following ... weakens the argument?

PrLMs like RoBERTa take subwords as input while our triplets extraction performs in word-level. For the global QA-context node, we also averaged the embeddings of tokens in question and option for initialization. We also use a one-hot embedding layer to encode the relations between two nodes.

**Graph Attention Network** Based on the relational graph convolutional network \([35]\) and given the initial representation \(h_i^{(l)}\) for every node \(v_i\), the feed-forward or the message-passing process can be written as:

\[
 h_i^{(l+1)} = \text{ReLU}(\sum_{r \in R_L} \sum_{v_j \in N_r(v_i)} \frac{1}{c_{i,r}} w_r^{(l)} h_j^{(l)}),
\]

(2)

where \(N_r(v_i)\) 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 \(l\).

For information passing control, we introduce the gating mechanism \([32]\), which calculates a value between 0 and 1.

\[
 g_i^{(l)} = \sigma(h_i^{(l)} W_{r,g}),
\]

(3)

where \(W_{r,g}^{(l)}\) is a learnable parameter under relation type \(r\) of the \(l\)-th layer. Finally, the forward process of gated GCN can be represented as:

\[
 h_i^{(l+1)} = \text{ReLU}(\sum_{r \in R_L} \sum_{v_j \in N_r(v_i)} g_{i,r}^{(l)} \frac{1}{c_{i,r}} w_r^{(l)} h_j^{(l)}),
\]

(4)

Through the graph encoder \(F_G(\cdot,\cdot)\), we then obtain the hidden representations of nodes in fact units as:

\[
 \{h_0^F, \ldots, h_m^F\} = F_G(\{v_{L,0}, \ldots, v_{L,m}\}, E_L).
\]

(5)

These features are further concatenated to get the final node representation of the super graph:

\[
 \{h_0^S, \ldots, h_m^S\} = F_G(\{h_0^F, \ldots, h_m^F\}, E_C).
\]

(6)

For node features on the super graph, it is fused via the attention and gating mechanism with the original representations of the context encoder. Specifically, denote the original whole sequence representation after context encoder as \(H^C\), we apply attention mechanism to append the super graph representation to the original one:

\[
 \tilde{H} = \text{Attn}(H^C, K_F, V_F),
\]

(7)

where \(K_F, V_F\) are packed from the learned representations of the super graph. Intuitively, this information may play an auxiliary effect during the answer prediction. Therefore, we compute \(\lambda \in [0, 1]\) to weigh the expected importance of super graph representation of each source word:

\[
 \lambda_1 = \sigma(W_\lambda \tilde{H} + U_\lambda H^C),
\]

(8)

where \(W_\lambda\) and \(U_\lambda\) are model parameters. We then fuse \(H^C\) and \(\tilde{H}\) to learn an effective representation:

\[
 H = H^C + \lambda \tilde{H} \in \mathbb{R}^{4 \times d}.
\]

(9)
3.3.3 Question-Option-aware Interaction

Since options have their inherent logical relations, contradictions may dwell in different options, which can be leveraged to aid answer prediction. Inspired by [36], we use an attention-based mechanism to gather option correlation information, and then the correlated option representation again interacts with question and passage representations.

Firstly, we define the attention operation which is frequently used in the following formulae. Given input matrices $U \in \mathbb{R}^{d \times N}$ and $V \in \mathbb{R}^{d \times M}$, the attention weight function $Attn(_i)$ is defined as:

$$A = Attn(U, V; v) = \frac{\exp(s_{ij})}{\sum_i \exp(s_{ij})}|_{i,j},$$  (10)

where $s_{ij} = v^T[U_i; V_j; U_i \circ V_j]$, $v \in \mathbb{R}^{3d}$ is a learnable hyperparameter. $d$ is the hidden dimension size. We will first fix an option and let it interact with all other options one-by-one to collect the pairwise correlation information. Then all the information will be fused to get an option-wise representation via gating mechanism. Specifically for an option $O_i$, the information it get by interaction with option $O_j$ is calculated as:

$$O_{i,j} = [O_i^a - O_j^a \cdot Attn(O_i^a, O_j^a; v); O_i^a \circ O_j^a \cdot Attn(O_i^a, O_j^a; v)],$$  (11)

where $O_i^a$ is the representation of the concatenation for the $i$-th option and question after the context encoder. Then the option-wise information are gathered to fuse the option correlation information, which is defined as:

$$\hat{O}_i = \tanh(W_c[O_i^a; \{O_{ij}^a\}_{i \neq j}^c] + b_c),$$  (12)

where $W_c \in \mathbb{R}^{d \times 7d}$ and $b_c \in \mathbb{R}^{d}$. Finally, a gating mechanism is used to fuse the option features with the obtained option correlation features to produce the advanced option features:

$$O_{i,k} = g_{i;k} \circ \hat{O}_{i;k} + (1 - g_{i;k}) \circ \hat{O}_{i;k},$$  (13)

where the $g_{i;k} = \sigma(W_g[O_{i;k}^a; \hat{Q}_{i,k} ; \hat{Q}] + b_g) \in \mathbb{R}^d$ is the $i$-th column of gate $g$.

The obtained option feature is further correlated with document feature via co-attention. The final representation corresponding to the $i$-th option is denoted as $O_i^f \in \mathbb{R}^{4 \times d}$.

3.4 Hierarchical Decoder

To better incorporate the information obtained above, apart from getting the original pooled context-attended representation $h^C \in \mathbb{R}^{4 \times d}$, we combine the attended vectors $O_i^f$ and $H$ from the previous encoder through a fusing layer.

$$E_1 = \text{ReLU}(\text{FC}([h^C, H, h^C - H, h^C \circ H])),$$
$$E_2 = \text{ReLU}(\text{FC}([h^C, H, h^C - O_i^f, h^C \circ O_i^f])),$$
$$P = \sigma(\text{FC}([E_1, E_2])),$$
$$C = P \circ H + (1 - P) \circ O_i^f \in \mathbb{R}^{4 \times d}.$$  (14)

Then another linear layer is applied for final prediction

$$z = W_z C + b_z \in \mathbb{R}^4.$$  (15)

We seek to minimize the cross entropy loss by

$$\mathcal{L}_{ans} = -\log \text{softmax}(z)_l,$$  (16)

where $l$ indicate the correct decision.

Logical Fact Regularization  

Inspired by [37], the embedding of the tail entity should be close to the embedding to the head entity plus a relation-related vector in the hidden representation space. Without loss of generality, we assume that in our settings, the summation of the subject vector and the relation vector should be close to the object vector as much as possible, i.e.,

$$v_{subject} + v_{relation} \rightarrow v_{object}.$$  (17)
In order to make the logical facts more of factual correctness, we introduce a regularization for the extracted logical facts based on the hidden states of the sequence \( h_i \) where \( i = 1, \ldots, L \) and \( L \) is the total length of the sequence. The regularization is defined as:

\[
L_{\text{fr}} = \sum_{k=1}^{m} (1 - \cos(h_{\text{sub}_k} + h_{\text{rel}_k}, h_{\text{obj}_k})),
\]

where \( m \) is the total number of logical fact triplets extracted from the context as well as the option and \( k \) indicates the \( k \)-th fact triplet.

**Training Objective.** During training, the overall loss for answer prediction is:

\[
\mathcal{L} = \alpha \mathcal{L}_{\text{ans}} + \beta \mathcal{L}_{\text{fr}},
\]

where \( \alpha \) and \( \beta \) are two parameters. In our implementation, we set \( \alpha = 1.0 \) and \( \beta = 0.5 \).

### 4 Experiments

#### 4.1 Datasets

We conducted the experiments on three datasets. Two for specialized logical reasoning ability testing: Reading Comprehension dataset requiring logical reasoning (ReClor) [3] and LogiQA [11] and one for logical reasoning in dialogues: Multi-Turn Dialogue Reasoning (MuTual) [38].

**ReClor** ReClor contains 6,138 multiple-choice questions modified from standardized tests such as GMAT and LSAT, which are randomly split into train/dev/test sets with 4,638/500/1,000 samples respectively. It contains multiple logical reasoning types. The held-out test set is further divided into EASY and HARD subsets based on the performance of BERT-based model [39].

**LogiQA** LogiQA consists of 8,678 multiple-choice questions collected from National Civil Servants Examinations of China and are manually translated into English by experts. The dataset is randomly split into train/dev/test sets with 7,376/651/651 samples correspondingly. LogiQA also contains various logical reasoning types.

**MuTual** MuTual has 8,860 dialogues annotated by linguist experts and high-quality annotators from Chinese high school English listening comprehension test data. It is randomly split into train/dev/test sets with 7,088/886/886 samples respectively. There more than 6 types of reasoning abilities reflected in MuTual. MuTual\textsuperscript{+} is an advanced version, where one of the candidate responses is replaced by a safe response (e.g., "could you repeat that?") for each example.

#### 4.2 Implementation Details

We fine-tune RoBERTa-large as the backbone pre-trained language model for FOCAL REASONER, which contains 24 hidden layers with hidden size 1,024. The overall model is end-to-end trained and updated by Adam [40] optimizer with an overall learning rate 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. For graph encoders, we implement it using DGL\textsuperscript{[4]} 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 is 16 and an overall dropout rate 0.1 on NVIDIA Tesla V100 GPU, which takes around 5 hours for ReClor and 10 hours for LogiQA.

#### 4.3 Results

Table 1 and Table 2 show the results on ReClor, LogiQA, and MuTual, respectively. All the best results are shown in bold. Based on our implemented baseline models (basically consistent with public results), we observe dramatic improvements on both of the logical reasoning benchmarks,

[4]: https://www.dgl.ai/
Table 1: Experimental results of our model compared with baseline models on ReClor and LogiQA dataset. Test-E and Test-H denote Test-Easy and Test-Hard respectively. We performed Pitman’s permutation test \cite{41} and found that our model significantly outperformed the baseline (p<0.05).

| Model                  | ReClor       | LogiQA       |
|------------------------|--------------|--------------|
|                        | Dev | Test | Test-E | Test-H | Dev | Test |
| Human \cite{3}          | -   | 63.00 | 57.10 | 67.20 | -   | 86.00 |
| BERT-Large \cite{3}     | 33.80 | 49.80* | 72.00* | 32.30 | 34.10* | 31.03* |
| XLNet-Large \cite{3}    | 62.00 | 56.00 | 75.70 | 40.50 | -   | -    |
| RoBERTa-Large \cite{3}  | 62.60 | 55.60 | 75.50 | 40.00 | 35.02 | 35.33 |
| DAGN \cite{6}           | 65.20 | 58.20 | 76.14 | 44.11 | 35.48 | 38.71 |
| DAGN (Aug) \cite{6}     | 65.80 | 58.30 | 75.91 | 44.46 | 36.87 | 39.32 |
| FOCAL REASONER\cite{6}  | 66.80 | 58.90 | 77.05 | 44.64 | 41.01 | 40.25 |

Table 2: Experimental results of our model compared with baseline PrLM on MuTual dataset.

e.g., on ReClor test set, FOCAL REASONER achieves +4.2\% on dev set and +3.3\% on the test set. FOCAL REASONER also outperforms the prior best system DAGN\cite{6}, reaching 77.05\% on the EASY subset, and 44.64\% on the HARD subset. The boost suggests that FOCAL REASONER makes better use of logical structure inherent in the given context to perform reasoning than existing methods.

In addition, Table 3 lists 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”. This means that our model can well interpret the question type from the question statement and make the correct choice corresponding to the question.

On the dialogue reasoning dataset MuTual, we also achieve quite a jump compared with the RoBERTa-base LM\cite{6}. This verifies our model’s generalizability on other downstream task settings. Our model can be smoothly applied to other tasks that require reasoning to boost performance.

## 5 Analysis

### 5.1 Ablation Study

Table 4 summarizes the ablation study conducted on each of our model’s components using the ReClor dev set.

Supergraph reasoning: The first key component is the supergraph reasoning. We ablate the global atom and erase all the edges connected with it. The results suggest that the global atom indeed provides a better message propagation effect, leveraging performance from 64.6\% to 66.8\%. We also find that replacing the initial

---

\footnote{For a fair comparison, we only compare to public literatures with the same PrLM RoBERTa-large. The test results can be found at https://eval.ai/web/challenges/challenge-page/503/leaderboard/1347}

\footnote{Since there are no official results on RoBERTa-large LM, we use RoBERTa-base LM instead for consistency.}
Table 5: Accuracy on the dev set of ReClor corresponding to 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.

| Model                | S   | W   | I   | CMP  | ER  | P   | D   | R   | IF  | MS  |
|----------------------|-----|-----|-----|------|-----|-----|-----|-----|-----|-----|
| RoBERTa\textsubscript{large}  | 61.70 | 47.79 | 39.13 | 63.89 | 50.77 | 50.00 | 56.25 | 56.67 | 56.67 |
| DAGN                | 63.83 | 46.02 | 39.13 | 69.44 | 57.14 | 53.85 | 62.50 | 62.39 | 56.67 |
| FOCAL \textsc{Reasoner} | 65.96 | 51.33 | 43.48 | 72.22 | 67.86 | 53.85 | 62.50 | 62.39 | 60.0 |

Table 4: Ablation results on the dev set of ReClor.

| Question reformulation: | Accuracy |
|------------------------|----------|
| FOCAL \textsc{Reasoner} | 66.8 ±0.13 |
| supergraph reasoning | 66.4 ±0.16 |
| w/o global atom | 64.6 ±0.32 |
| w/o coreference | 64.8 ±0.24 |
| w/o logical fact regularization | 64.2 ±0.12 |
| QA context atom → Q atom | 63.7 ±0.19 |
| w/o edge type | 62.8 ±0.26 |
| fact unit → named entity | 65.2 ±0.16 |
| Question Reformulation | 63.5 ±0.52 |
| w/o question direction | 65.2 ±0.16 |
| Interactions | 63.5 ±0.52 |

Table 6: Distribution of fact unit number on dev set of the training datasets.

| Dataset | [0, 3) | [3, 6) | [6, 9) | [9, 12) | [12, ∞) |
|---------|--------|--------|--------|--------|--------|
| ReClor  | 37.2%  | 48.6%  | 12.6%  | 0.6%   | 1.2%   |
| LogiQA  | 47.5%  | 37.5%  | 10.9%  | 3.5%   | 0.6%   |

We display the accuracy of RoBERTa-large, DAGN and our proposed FOCAL \textsc{Reasoner} in Figure 5.2. We find that our model outperforms baseline models on all the divided subsets, which demonstrates the effectiveness of our model on different numbers of fact units. Specifically, for ReClor, FOCAL \textsc{Reasoner} performs better when there are more fact units in the context, while for LogiQA, FOCAL \textsc{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 include ours struggle when the number of fact units is above some certain thresholds, i.e., the logical structure is more complicated, calling for better mechanisms to cope with complex logical structures.

5.3 Interpretability: a Case Study

We aim to interpret FOCAL REASONER’s reasoning process by analyzing the node-to-node attention weights induced in the supergraph in Figure 6. We can see that our FOCAL REASONER can well bridge the reasoning between context, question and option. Specifically, in the graph, ”students rank 30%” attends strongly to ”playing improve performance”. Under the guidance of question to select the option that weakens the statement and option interaction, our model is able to tell that ”students rank 30% can play” mostly undermines the conclusion that ”playing improves performance”.

A recent survey in a key middle school showed that high school students in this school have a special preference for playing football, and it far surpasses other balls. The survey also found that students who regularly play football are better at academic performance than students who do not often play football. This shows that often playing football can improve students’ academic performance.

1. students have preferences
2. preference playing football
3. it surpasses balls
4. who play football
5. students better performance
6. who play football
7. playing improve performance
8. students rank 30%
9. students play football

Which of the following can weaken the above conclusion most?

A. Only high school students who are ranked in the top 30% of grades can often play football.
B. Regular football can exercise and maintain a strong learning energy.
C. Often playing football delays the study time.
D. Research has not proved that playing football can contribute to intellectual development.

Figure 6: An example of how our model reasons to get the final answer.

6 Conclusion

In this work, we propose a novel method named FOCAL REASONER for logical reasoning in the machine reading comprehension task. Our method not only better uncovers the logical structures within the context, which can be a general method for other sophisticated reasoning tasks but also better captures the logical interactions between context and options. The experimental results verify the effectiveness of our method. In the future, we intend to design more elaborate mechanisms to cope with different question types and logical types as well as combine the symbolic and neural approaches.
References

[1] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In Christian Bessiere, editor, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[2] Lucja Iwańska. Logical reasoning in natural language: It is all about knowledge. Minds and Machines, 3(4):475–510, 1993.

[3] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In International Conference on Learning Representations (ICLR), April 2020.

[4] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In North American Chapter of the Association for Computational Linguistics (NAACL), 2021.

[5] Hongyu Ren and Jure Leskovec. Beta embeddings for multi-hop logical reasoning in knowledge graphs. Advances in Neural Information Processing Systems, 33, 2020.

[6] Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. DAGN: Discourse-aware graph network for logical reasoning. In NAACL, 2021.

[7] Siddharth Krishna, Alexander J Summers, and Thomas Wies. Local reasoning for global graph properties. In European Symposium on Programming, pages 308–335. Springer, Cham, 2020.

[8] Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. AR-LSAT: Investigating Analytical Reasoning of Text. arXiv e-prints, page arXiv:2104.06598, April 2021.

[9] Ndapandula Nakashole and Tom Mitchell. Language-aware truth assessment of fact candidates. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1009–1019, 2014.

[10] Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. Dynamically fused graph network for multi-hop reasoning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6140–6150, 2019.

[11] Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. Cognitive graph for multi-hop reading comprehension at scale. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2694–2703, 2019.

[12] Jifan Chen, Shih-ting Lin, and Greg Durrett. Multi-hop question answering via reasoning chains. arXiv preprint arXiv:1910.02610, 2019.

[13] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4658–4664, Florence, Italy, July 2019. Association for Computational Linguistics.

[14] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics, 8:842–866, 2020.

[15] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1, pages 1693–1701, 2015.

[16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.
[17] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[18] Divyansh Kaushik and Zachary C. Lipton. 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, pages 5010–5015, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[19] Ming Zhou, Nan Duan, Shujie Liu, and Heung-Yeung Shum. Progress in neural nlp: Modeling, learning, and reasoning. Engineering, 6(3):275–290, 2020.

[20] Danqi Chen, Jason Bolton, and Christopher D. Manning. 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), pages 2358–2367, Berlin, Germany, August 2016. Association for Computational Linguistics.

[21] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. Communications of the ACM, 58(9):92–103, 2015.

[22] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In International Conference on Learning Representations, 2019.

[23] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In NAACL-HLT (1), 2019.

[24] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual 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 2391–2401, 2019.

[25] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, 2019.

[26] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2369–2380, 2018.

[27] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, 2015.

[28] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, 2018.

[29] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4885–4901, 2020.

[30] Siyuan Wang, Wanjun Zhong, Duyu Tang, Zhongyu Wei, Zhihao Fan, Daxin Jiang, Ming Zhou, and Nan Duan. Logic-driven context extension and data augmentation for logical reasoning of text. arXiv preprint arXiv:2105.03659, 2021.

[31] Friedrich Wilhelm Levi. Finite geometrical systems: six public lectures delivered in February, 1940, at the University of Calcutta. University of Calcutta, 1942.
[32] Diego Marcheggiani and Ivan Titov. Encoding sentences with graph convolutional networks for semantic role labeling. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1506–1515, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[33] Daniel Beck, Gholamreza Haffari, and Trevor Cohn. 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), pages 273–283, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv e-prints, page arXiv:1907.11692, July 2019.

[35] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne vanden Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, The Semantic Web, pages 593–607. Springer International Publishing, 2018.

[36] Qiu Ran, Peng Li, Weiwei Hu, and Jie Zhou. Option comparison network for multiple-choice reading comprehension, 2019.

[37] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 26. Curran Associates, Inc., 2013.

[38] Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. Mutual: A dataset for multi-turn dialogue reasoning. In Proceedings of the 58th Conference of the Association for Computational Linguistics. Association for Computational Linguistics, 2020.

[39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics.

[40] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[41] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker’s guide to testing statistical significance in natural language processing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1383–1392. Association for Computational Linguistics, 2018.