Modeling Hierarchical Logical Reasoning Chains

Jialin Chen, Zhuosheng Zhang, Hai Zhao

Shanghai Jiao Tong University
sjtuchenjl@sjtu.edu.cn, zhangzs@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

Machine reading comprehension poses new challenges over logical reasoning, which aims to understand the implicit logical relationships entailed in the given contexts and perform inference over them. Due to the complexity of logic, logical relationships exist at different granularity levels. However, most existing methods of logical reasoning individually focus on either entity-aware or discourse-based information but ignore the hierarchical relationships that may even have mutual effects. In this paper, we propose a holistic graph network (HGN) which deals with context at both discourse level and word level, as the basis for logical reasoning, to provide a more fine-grained relationship extraction. Specifically, a dual-level attention mechanism, including node-level and type-level attention, is leveraged in our graph attention model to interact with different nodes, which can be interpreted as bridges in the reasoning process. Experimental results on two logical reasoning QA benchmark datasets, ReClor and LogiQA, show the effectiveness of our method, and in-depth analysis verifies its capability to understand complex logical relationships.

1 Introduction

Machine reading comprehension (MRC) is a challenging task that requires machines to answer a question according to given passages (Hermann et al. 2015; Rajpurkar et al. 2016; Rajpurkar, Jia, and Liang 2018; Lai et al. 2017). A variety of datasets have been introduced to push the development of MRC to a more complex and more comprehensive pattern, such as conversational MRC (Reddy, Chen, and Manning 2019; Choi et al. 2018), multi-hop MRC (Yang et al. 2018), and commonsense reasoning (Davis and Marcus 2015; Bhagavatula et al. 2020; Talmor et al. 2019; Huang et al. 2019). In particular, some recent multi-choice MRC datasets pose even greater challenges to the logical reasoning ability of models (Yu et al. 2020; Liu et al. 2020) which are not easy for humans to do well, either. Firstly, all the supporting details needed for reasoning are provided by the context, which means there is not available additional commonsense or knowledge. Secondly, it is a task of answer selection rather than answer retrieval, which means the best answer is chosen according to their logical fit with the given context and the question, rather than retrieved from the context directly. Most importantly, the relationships entailed in the contexts are much more complex than that of previous MRC datasets owing to the complexity of logic.

Example (taken from ReClor dataset)
Context: Most lecturers who are effective teachers are eccentric, but some non-eccentric lecturers are very effective teachers. In addition, every effective teacher is a good communicator.
Question:
Which one of the following statements follows logically from the statements above?
Options:
A: Most lecturers who are good communicators are eccentric.
B: Some non-eccentric lecturers are effective teachers but are not good communicators.
C: All good communicators are effective teachers.
D: Some good communicators are eccentric. ✓

Figure 1: An example from Reclor dataset.

The problems of logical reasoning from MRC tasks are usually to find a statement that supports or undermines the context or to find the flaw or principle of the context. An example is shown in Figure 1. As humans, to solve such kinds of problems, we usually go through the following steps. Firstly, we divide the context into sections to capture the logical relationship between each clause, such as transition, continuity, contrast, and so on. Also, we ignore parentheses, clauses, etc., which are not important for the understanding of the context. Secondly, we extract the important elements (words or phrases) in the context, namely, the objects and topics described by the context, and construct the logical chain with the topic as the elements. Finally, we need to compare the answer statement to the mentioned part in the context and assess its logical fit with the given context.

Most existing methods of logical reasoning only focus on either entity-aware or discourse-based information but ignore the hierarchical relationships that may even have mutual effects (Yu et al. 2020; Liu et al. 2020; Wang et al. 2021; Huang et al. 2021). Inspired by human reasoning processes, in this paper, we model logical reasoning chains based on a newly proposed holistic graph network (HGN) that incorporates the information of element discourse units (EDU) (Gao et al. 2020; Ouyang, Zhang, and Zhao 2021) and key phrases (KPH) extracted from context and answer. Thus, we focus on both EDU and KPH to discover their individual features

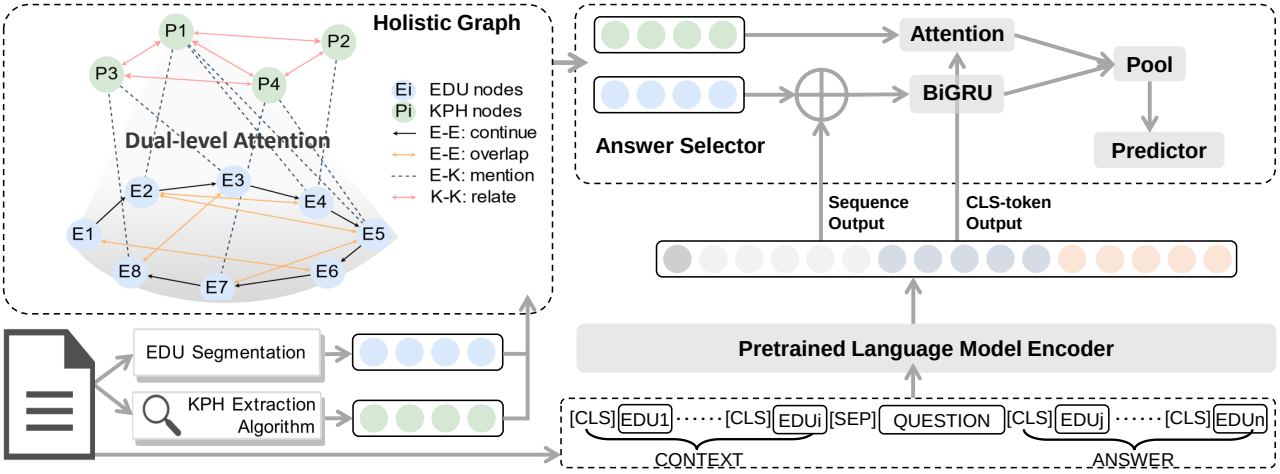

Figure 2: An overview of our proposed holistic graph-based reasoning model.

and interactive relationships. To model logical reasoning chains, we apply the proposed holistic graph with two types of nodes—EDU and KPH—and reasonable edge connection rules to learn both hierarchical features and effects of interactions between different granularity levels. Then we apply a dual-level attention mechanism to update the features, considering the contributions of node features and type features at the same time. Experimental results show that our model has a strong capability to retrospect the reasoning process, which corresponds with the reasoning chains by demonstrating performance improvement over a strong baseline.

2 Problem Statement

For multi-choice MRC tasks with logical reasoning, we aim to find the best answer among four given options based on a piece of context which entails some logical relationships. The questions are usually to find the answer that supports or weakens the context the most or to find the flaw or principle of the context. Figure 1 shows an example from ReClor dataset (Yu et al. 2020) which requires logical reasoning ability to make the correct predictions. We observe that the logical relationship in this kind of reasoning dataset is complicated and often occurs between several important phrases ("effective teachers, non-eccentric, eccentric, good communicator") in Figure 1). Formally, given a natural language context C containing m tokens $\{t_1, t_2, \ldots, t_m\}$, a question Q, and four potential answers A=$\{A_1, A_2, A_3, A_4\}$, we use EDU and KPH extraction algorithm to process C plus $A_i (i = 1, 2, 3, 4)$ sequence. $\{P_j\}_j$ are extracted KPH nodes and $\{E_j\}_j$ are EDU nodes.

3 Methodology

Given a context, a question and four potential answers, we concatenate them to get four $\{C, Q, A_i\}$ pairs. To incorporate the principle of human inference into our method, we propose a holistic graph network (HGN) as shown in Figure 2. Our model works as follows. First, we use EDU and KPH extraction algorithm to get necessary nodes from the given pairs. Then, we apply a simple rule-based method to construct a logical chain for every set of $\{C, Q, A_i\}$. Intuitively, if P_i is mentioned by E_j, then there is an interaction between the two different nodes. If there is a direct connection between P_i and P_j, then there is a link between the two nodes. Based on these extracted nodes and relations, we combine the logical chain with KPH and EDU interaction information to construct the holistic graph. The process of constructing holistic graph is shown in Figure 3. Dual-level attention mechanism is used to further capture the relationship between $\{E_j\}$ and $\{P_j\}$ extracted from the context and candidate answers at both node level and type level. Finally, we make the answer prediction using attention layer, bidirectional gating recurrent unit (BiGRU) and Pooling layer.

3.1 Logical Chain Construction

Element Discourse Units (EDU) We use clause-like text spans delimited by logical relations to construct the rhetorical structure of texts. These clause-like discourses can be regarded as element units that reveal the overall logic and emotional tone of the text. For example, conjunctions like "`because`" indicate a causal relationship, which means the following discourse is likely to be the conclusion we need to pay attention to; "`however`" indicates a transition relationship, which means the following discourse provides an opposite situation. Parenthesis and clauses like "`who are effective teachers`" in Figure 3 play a complementary or limiting role in context. Also, punctuation indicates a pause or end of a sentence, containing semantic transition and turning point implicitly. We use an open segmentation tool, SEGBOT (Li, Sun, and Joty 2018), to identify the element discourse units (EDUs) from the concatenation of `context` and `answer`, ignoring the question whose structure is simple. Conjunctions, punctuation and the beginning of clauses are usually the segment points.

To get the initial embedding of EDUs, we insert an external `[CLS]` symbol at the start of each discourse, and add a `[SEP]` symbol at the end of every type of inputs. Then

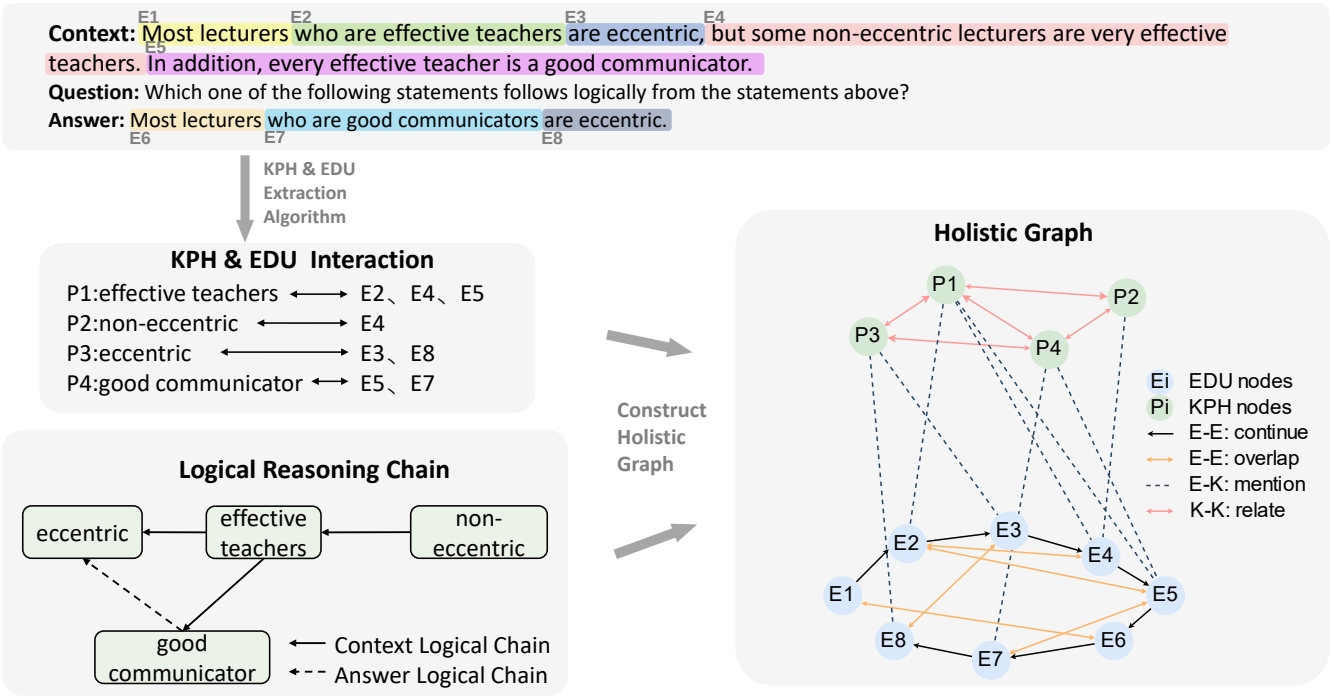

Figure 3: Process of constructing the holistic graph based on KPH-EDU interaction information and Logical Reasoning Chain.

we use RoBERTa to encode the concatenated tokens and the encoded `[CLS]` token represents the following EDU. Therefore, we get the initial embedding of EDU, denote to $\{E_1, \ldots, E_n\}$.

Key Phrase (KPH) Key Phrases, including keywords here, play an important role in context. They are usually the object and principle of a context. We use the sliding window to generate n-gram word list, filtering according to the Stopword list, POS tagging, the length of the word, and whether it contains any number.[1]

The filtering process is based on the following two main criteria: (1) If the n-gram contains a stop word or a number, then delete it. (2) If the length of word is less than the threshold value m, delete it, and if the n-gram length is 1, then only the noun, verb, and adjective are retained.

Then, we calculate the TF-IDF feature of each n-gram, and select top-k n-gram as key phrases. We restore the selected tokens and retrieve the original expressions containing the key phrase from the original text. For example, as in Figure 3, "`eccentric`" is one of the KPHs, while we retrieve the original expression "`eccentric`" and "`non-eccentric`" from the original text. [2]

Given the token embedding sequence $\{t_1, \ldots, t_n\}$ of a KPH with length n, its initial embedding is obtained by

$$P_i = \frac{1}{|K_i|} \sum_{l \in K_i} t_l, \tag{1}$$

where K_i is the set of tokens that belong to the KPH.

Holistic Graph Construction Formally, every input sample is a triplet that consists of a context, a question and a candidate answer. EDU and KPH nodes are extracted in the above way. We propose a Holistic Graph Network (HGN) to model the logical reasoning chain. For example, Figure 3 shows the process of constructing holistic graph of the example in Figure 1. The Holistic Graph includes two types of nodes: EDU Nodes (in blue) and KPH Nodes (in green). For edge connections, there are four distinct types of edges between pairs of nodes.

- EDU-EDU continue: the two nodes are contextually associated in the context and answer. This type of edge is directional.
- EDU-EDU overlap: the two nodes contain the same KPH. This type of edge is bidirectional.
- EDU-KPH mention: the EDU node mentions the KPH. This type of edge is bidirectional.
- KPH-KPH relate: the two nodes are semantically related. We define two types of semantic relationships. One is that the two KPHs are retrieved by the same n-gram as described above. The other one is that the Cosine similarity between the two KPH nodes is greater than a threshold. This type of edge is bidirectional. We believe that semantically-related-type connections can capture the information of word pairs like synonyms and antonyms.

[1]The stop list is derived by the open-source toolkit Gensim: https://radimrehurek.com/gensim/. The POS tagging is derived by the open-source toolkit NLTK: https://www.nltk.org/.

[2]The complete extraction algorithm is given in Appendix.

3.2 Dual-level Attention Mechanism

Considering a specific node in the holistic graph, neighboring nodes in the same type may carry more useful and more important information, thus affecting each other in a direct way. In the process, the neighboring nodes in the different types may also interact with each other. For example, node `P4` interacts with node `E5` and `E7` which helps model to better locate where logical relationships occur. To capture both the node-level and type-level attention, we apply a Dual-level Attention Mechanism to the update of the graph network's representations.

Graph Preliminary Formally, consider a graph $G = \{V, E\}$, where V and E represent the sets of nodes and edges respectively. A is the adjacency matrix of the graph. $A_{ij} > 0$ means there is an edge from the i-th node to the j-th node. We introduce $A' = A + I$ to take self-attention into account. In order to avoid changing the original distribution of the feature when multiplying with the adjacency matrix, we normalize A', set $\tilde{A} = D^{-\frac{1}{2}} A' D^{-\frac{1}{2}}$ where D is the degree matrix of the graph. $D = diag\{d_1, d_2 \ldots, d_n\}$, d_i is the number of edges attached to the i-th node.

Now, we calculate the attention score from node v' to node v in the following steps.

Type Attention Vector We use $T(\tau)$ to represent all nodes that belong to type τ, and $N(v)$ to represent all neighboring nodes that are adjacent to v. T is the set of types. Assume that node v belongs to $T(\tau)$, h_μ is the feature of node μ, h_τ is the feature of type τ which is computed by

$$h_\tau = \sum_{\mu \in T(\tau)} \tilde{A}_{v\mu} W h_\mu. \tag{2}$$

Using the feature of type and node v, we compute the attention score of type τ as:

$$e_\tau = \sigma(\mu_\tau^T \cdot [W h_v \parallel W_\tau h_\tau]). \tag{3}$$

Then, type-level attention weights α_τ is obtained by normalizing the attention scores across all the types T with the softmax function. σ is an activate function such as leaky-ReLU.

$$\alpha_\tau = \frac{\exp(\sigma(\mu_\tau^T \cdot [W h_v \parallel W_\tau h_\tau]))}{\sum_{\tau' \in T} \exp(\sigma(\mu_{\tau'}^T \cdot [W h_v \parallel W_\tau h_{\tau'}]))}. \tag{4}$$

Node Attention Vector α_τ shows the importance of nodes in type τ to node v. While computing the attention score of node v' that is adjacent to node v, we multiply that by the type attention weights α_τ (assume v' belongs to type τ). Similarly, node attention weights are obtained by the softmax function across all neighboring nodes.

$$e_{vv'} = \sigma(\nu^T \cdot \alpha_\tau [W h_v \parallel W h_{v'}]), \tag{5}$$

$$\alpha_{vv'} = \frac{\exp(e_{vv'})}{\sum_{i \in N(v)} \exp(e_{vi})}, \tag{6}$$

where $\parallel$ is the concatenation operator and $\alpha_{vv'}$ is the attention weight from node v' to v.

Update of Node Representation Let $h_v^{(l)}$ be the representation of the node v at the l-th layer. Then the layer-wise propagation rule is as follows:

$$h_v^{(l+1)} = \sigma(\sum_{v' \in N(v)} \alpha_{vv'} W h_{v'}^{(l)}). \tag{7}$$

3.3 Answer Selector

To predict the best answer that fits the logic entailed in the context, we extract the node representations of the last layer of the graph network and feed them into the downstream predictor. For the nodes of type EDU, since the node order implies the order of node occurrence in the context, we align them with the output of sequence embedding and add to it as a residual part. Then, they are fed into a bidirectional gating recurrent unit (BiGRU) to extract interaction information.

$$\tilde{H}_E = \text{BiGRU}(H_E + H_{sent}) \in \mathbb{R}^{l \times d}, \tag{8}$$

where $H_E = [h_{v'_1}, h_{v'_2}, \ldots, h_{v'_l}] \in \mathbb{R}^{l \times d}$, v'_i belongs to type EDU. l and d are the sequence length and the feature dimension respectively. H_{sent} is the output of sequence embedding.

For the nodes of type KPH, we first expand the embedding of the first `[CLS]` token to size $1 \times d$, denoted as H_c. Then, we feed the embedding of `[CLS]` token and features of type KPH node $H_K = [h_{v_1}, h_{v_2}, \ldots, h_{v_n}] \in \mathbb{R}^{n \times d}$ (v_i belongs to type KPH) into an attention layer.

$$\begin{aligned} \alpha_i &= w_\alpha^T [H_c \parallel h_{v_i}] + b_\alpha \in \mathbb{R}^1, \\ \tilde{\alpha_i} &= \text{softmax}(\alpha_i) \in [0, 1], \\ \tilde{H}_c &= W_c \sum_i \tilde{\alpha_i} h_{v_i} + b_c \in \mathbb{R}^{1 \times d}, \end{aligned} \tag{9}$$

where $\tilde{\alpha_i}$ is the attention weight of node feature h_{v_i}. w_α, b_α, W_c, and b_c are parameters.

The output of BiGRU and the output of attention layer are concatenated and go through a pooling layer, followed by an MLP layer as the predictor. We take a weighted sum of the concatenation as the pooling operation. The predictor is a two-layer MLP with a tanh activation. Specially, coarse-grained and fine-grained features are further fused here to extract more information.

$$\begin{aligned} \tilde{H} &= W_p [\tilde{H}_E \parallel \tilde{H}_c], \\ p &= \text{MLP}(\tilde{H}) \in \mathbb{R}, \end{aligned} \tag{10}$$

where W_p is a learnable parameter, $\parallel$ is the concatenation operator. For each sample, we get $P = [p_1, p_2, p_3, p_4]$, p_i is the probability of i-th answer predicted by model.

The training objective is defined as cross entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_i^N \log \text{softmax}(p_{y_i}), \tag{11}$$

where y_i is the ground-truth choice of sample i. N is the number of samples.

Model	ReClor				LogiQA	
	Dev	Test	Test-E	Test-H	Dev	Test
In literature						
Human (Yu et al. 2020)(Liu et al. 2020)	-	63.00	57.10	67.20	-	86.00
RoBERTa$_{\text{BASE}}$ (Yu et al. 2020)	55.00	48.50	71.10	30.70	-	-
RoBERTa$_{\text{LARGE}}$ (Yu et al. 2020)	62.60	55.60	75.45	40.00	35.02	35.33
DAGN (Huang et al. 2021)	65.20	58.20	76.14	44.11	35.48	38.71
DAGN (Aug) (Huang et al. 2021)	65.80	58.30	75.91	**44.46**	36.87	39.32
Our implementation						
RoBERTa$_{\text{BASE}}$	55.20	48.80	70.90	31.42	33.25	32.72
RoBERTa$_{\text{LARGE}}$	62.50	55.50	75.40	40.20	35.33	35.45
HGN$_{\text{BASE}}$	**56.30** (↑1.1)	**51.40** (↑2.6)	**75.23** (↑4.33)	**32.68** (↑1.26)	**39.48** (↑6.23)	**35.03** (↑2.31)
HGN$_{\text{LARGE}}$	**66.40** (↑3.9)	**58.70** (↑3.2)	**77.73** (↑2.33)	43.75 (↑3.55)	**40.09** (↑4.76)	**39.87** (↑4.42)

Table 1: Experimental results (Accuracy: %) of our model compared with baseline models on ReClor and LogiQA datasets. Test-E and Test-H denote Test-Easy and Test-Hard subclass of the ReClor dataset respectively. Our model significantly outperforms the baselines based on t-test (p-values < 0.05). The development set of ReClor dataset and the test set of LogiQA dataset are used to calculate the p-values.

4 Experiment

4.1 Dataset

Our evaluation is based on two logical reasoning MRC benchmarks, ReClor (Yu et al. 2020) and LogiQA (Liu et al. 2020). 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 the BERT-based model (Devlin et al. 2019). Compared with ReClor, LogiQA has more instances (8678 in total) and is randomly split into train/dev/test sets with 7,376/651/651 samples respectively. It derives from expert-written questions for testing human logical reasoning ability (Liu et al. 2020). The statistics of the EDU numbers on the two datasets are shown in Table 2.

	ReClor			LogiQA		
	Min	Avg	Max	Min	Avg	Max
Context	1	7.858	17	1	7.169	20
Answer	1	2.574	13	1	1.839	14

Table 2: Statistics of the EDU numbers

4.2 Implementation Details

Our model is implemented using Pytorch and based on the Transformers Library (Wolf et al. 2020). We use RoBERTa (Liu et al. 2019) as our backbone model. Adam (Kingma and Ba 2015) is used as our optimizer. For ReClor, the batch size is 24 and the initial learning rate is 1×10^{-5}. For LogiQA, the batch size is 2 and the initial learning rate is 4×10^{-6}. The best threshold for defining semantic relevance is 0.5. We run 15 epochs and select the model that achieves the best result in validation. Our models are trained on one 32G NVIDIA Tesla V100 GPU. The training time is around half an hour for each epoch. Our source codes are available at `Anonymous`.[3]

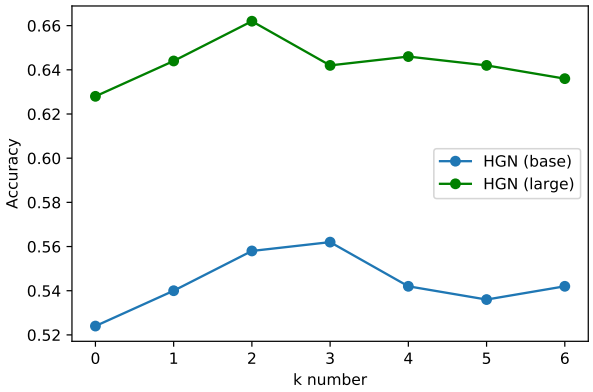

Figure 4: Dev accuracy v.s. number of KPH nodes.

4.3 Main Result

Table 1 presents the detailed results on the development set and the test set of both ReClor and LogiQA dataset. We fine-tune RoBERTa$_{\text{BASE}}$ and RoBERTa$_{\text{LARGE}}$ as the backbones respectively. In Table 1, we observe consistent improvements over the the baselines. HGN$_{\text{BASE}}$ reaches 51.4% of test accuracy on ReClor, outperforming the baseline on both Easy subset and Hard subset, and also reaches 35.03% of test accuracy on LogiQA, outperforms other existing models. HGN$_{\text{LARGE}}$reaches 58.7% of test accuracy on ReClor, therein 77.73% on Easy subset and 43.75% on Hard subset and 39.87% on LogiQA. Compared with the backbone, our model shows great improvement over this task by better utilizing the interaction information, which is ignored by existing methods.

[3]We have uploaded our source codes as Supplemental Material, which will be publicly available.

Model	Accuracy (%)
HGN_{BASE}	**56.3**
Graph Construction	
- EDU	55.8 (↓0.5)
- KPH	53.9 (↓2.4)
- edge type: E-E continue	53.0 (↓3.3)
- edge type: E-E overlap	54.0 (↓2.3)
- edge type: E-K mention	54.2 (↓2.1)
Dual-level-attention Mechanism	
- type-level attention (i.e. GAT)	54.8 (↓1.5)
- both (i.e. GCN)	55.7 (↓0.6)
Answer Selector	
- BiGRU	53.2 (↓3.1)
- Attention layer	55.0 (↓1.3)

Table 3: Ablation results on the dev set of ReClor.

4.4 Parameter Experiments

In this section, we investigate the sensitivity of parameter k, which is the number of KPH node. Figure 4 shows the accuracies on the development set of our proposed model with different numbers of KPH nodes which are extracted according to TF-IDF weights. Clearly, $k = 2$ or $k = 3$ is an appropriate parameter for our model. This is consistent with our intuition that a paragraph will have 2 to 3 key phrases as the topic of the context. When k is too small or large, the accuracy of the model does not perform well.

4.5 Ablation

We conduct a series of ablation studies with the performance of our model (base) on Holistic Graph Construction, Dual-level-attention Mechanism and Answer Selector. Results are shown in Table 3. The dataset used here is ReClor.

Holistic Graph Construction The first key component of our proposed model is Holistic Graph Construction that contains two types of nodes and four types of edges. We remove the nodes of EDU and KPH respectively and the results show that the removal hurts the performance badly. The accuracies drop to 55.8% and 53.9%. Furthermore, we delete each type of edge respectively. The test accuracy drops to worse than the baseline. The removal of edge type destroys the integrity of the network and may ignore some essential interaction information between EDUs and KPHs.

Dual-level-attention Mechanism Dual-level-attention Mechanism helps to capture the information contained in different node types. When we remove the type-level attention, the model is equivalent to a normal Graph Attention Network (GAT), ignoring the heterogeneous information. As a result, the performance drops to 54.8%. When we remove both types of attention, the model solely updates the features of nodes according to the features of the neighboring nodes. The performance on ReClor dev set drops to 55.7%.

Answer Selector We make two changes to the answer selector module in this part: (1) deleting the BiGRU, (2) deleting the attention layer. For (1), the output of EDU features concatenates with the output of the attention layer directly and then are fed into the downstream pooling layer. For (2), we ignore the attention between the KPH features and the whole sentence-level features. The resulting accuracies of (1) and (2) drop to 53.2% and 55% respectively. It is proved that the further fusion of features with different granularity is necessary in our proposed model.

4.6 Case Study

To intuitively show how our model works, we select an example from ReClor as shown in Figure 5, whose answer is predicted correctly by our model but not by baseline models (RoBERTa). The example shows that powerful pre-trained language models such as RoBERTa may be better at dealing with sentences that have the same words. For example, the wrong answer chosen by the baseline model (RoBERTa) is another expression of the first sentence in the given context. The words are basically the same, only the order changes. The model itself does not understand the logical relationship between sentences and phrases, but only compares their common elements to do prediction, failing in logical reasoning task. In contrast, our model can not only match synonymic expressions, but also make logical inferences by separating sentences into EDUs and extracting important phrases and establishing logical relationships between them. The right part is the corresponding attention map, deriving from the last layer of our model. In particular, our model's attention map is asymmetric due to the introduction of dual-level attention mechanism. From the attention map, we observe that the significance of E_i nodes for P_i nodes is inconsistent with that of P_i nodes for E_i nodes. It depends both on the feature of nodes and the type of the updated node.

5 Related Work

5.1 Machine Reading Comprehension

MRC is an AI challenge that requires machines to answer questions based on a given passage, which has aroused great research interests in the last decade (Hermann et al. 2015; Rajpurkar et al. 2016; Rajpurkar, Jia, and Liang 2018; Lai et al. 2017). Although recent systems have reported human-parity performance on various benchmarks (Zhang et al. 2020; Back et al. 2020; Zhang, Yang, and Zhao 2021) such as SQuAD (Rajpurkar et al. 2016; Rajpurkar, Jia, and Liang 2018) and RACE (Lai et al. 2017), whether the machine has necessarily achieved human-level understanding remains controversial (Zhang, Zhao, and Wang 2020; Sugawara, Stenetorp, and Aizawa 2021). Such concern stimulates the research interests of investigating the real ability of MRC systems, and what kind of knowledge or reading comprehension skills the systems have grasped (Sugawara et al. 2020), such as logical reasoning (Iwańska 1993). Recently, there is increasing interest in improving machines' logical reasoning ability, which can be categorized into symbolic approaches and neural approaches. Notably, analytical reasoning machine (AMR) (Zhong et al. 2021) is a typical symbolic method that injects human prior knowledge to deduce legitimate solutions. Specifically, it first extracts arguments

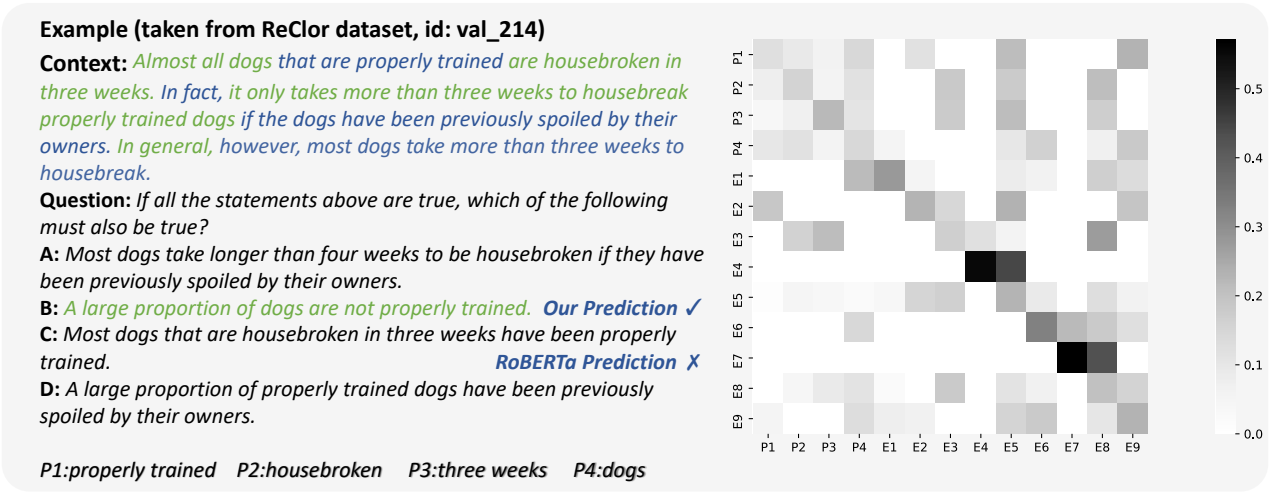

Figure 5: An example showing the logical reasoning capability of our model (Left) and the corresponding attention map (Right). EDUs are shown in different colors alternately, corresponding to E1-E9 in the attention map.

in the context, then a set of predefined logical functions are used to organize the arguments into logical constraint functions. Lastly, a tree-based reasoning algorithm is activated to deduce the answer from the previous functions. Discourse-aware graph network (DAGN) proposed by (Huang et al. 2021) also uses discourse relations to help logical reasoning, which is quite related to this work. However, our work instead incorporates key phrases (KPHs) and interactive attention between phrases and discourses to capture fine-grained features. Similar to our approach of discovering reasoning chains between element discourse and key phrases, Fang et al. (2020), proposes a hierarchical graph network (HGN) and claims that information of different granularity levels significantly helps to multi-hop QA. Our method instead avoids the incorporation of external knowledge and designs the specific pattern for logical reasoning task.

5.2 Logical Reasoning

Neural and symbolic methods have been studied for logical reasoning (Garcez et al. 2015; Besold et al. 2017; Chen et al. 2019; Ren and Leskovec 2020; Wang et al. 2021; Huang et al. 2021). Compared with the neural methods for logical reasoning, symbolic approaches rely heavily on dataset-related predefined patterns which entails massive manual labor, greatly reducing the generalizability of models. Also, it could introduce propagated errors since the final prediction depends on the intermediately generated functions. Even if one finds the gold programs, executing the program is quite a consuming work as the search space is quite large and not easy to prune. Therefore, we focus on the neural research line in this work, to capture the logic clues from the natural language texts, without the rely on human expertise and extra annotation.

Although the logical reasoning MRC task is a new task that there are only a few latest studies, we broaden the discussion to scope of the related tasks that require reasoning, such as commonsense reasoning (Davis and Marcus 2015; Bhagavatula et al. 2020; Talmor et al. 2019; Huang et al. 2019), multi-hop QA (Yang et al. 2018) and dialogue reasoning (Cui et al. 2020), to gain insights of how to model the relationships between the basic logical elements, and construct the logical chains. Previous approaches commonly consider the typical entity-level sentence-level relations which are obviously not sufficient to solve the problem (Qiu et al. 2019; Ding et al. 2019; Chen, Lin, and Durrett 2019). Besides, they heavily rely on external knowledge and fail to use attention mechanisms to capture important interaction information. In contrast, our work is inspired by human reasoning process, taking advantages of both inter-sentence elementary discourse units and intra-sentence key phrases, to construct hierarchical interactions for reasoning. The fine-grained holistic features are used for measuring the logical fitness of the candidate answer options and the given context. As our method enjoys the benefits of modeling reasoning chains from riddled texts, our model can be easily extended to other types of reasoning tasks. Especially where the given context has complex discourse structure and logical relations, like DialogQA, multi-hop QA. We left all the easy empirical verification of our method as future work.

6 Conclusion

This paper presents a novel method to guide the MRC model to better perform logical reasoning tasks. Inspired by the human reasoning process, we propose a holistic graph-based system to model hierarchical logical reasoning chains. To our best knowledge, we are the first to deal with context at both discourse level and phrase level as the basis for logical reasoning. In order to better utilize the interaction information of different types of nodes, we apply a dual-level attention mechanism to update the features, which captures the key information at different granularity levels and performs more effective denoising. On the two benchmark datasets ReClor and LogiQA, our proposed model shows effective by significantly outperforming the strong baselines.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Besold, T. R.; Garcez, A. d.; Bader, S.; Bowman, H.; Domingos, P.; Hitzler, P.; Kühnberger, K.-U.; Lamb, L. C.; Lowd, D.; Lima, P. M. V.; et al. 2017. Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*.

Bhagavatula, C.; Bras, R. L.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; Yih, W.; and Choi, Y. 2020. Abductive Commonsense Reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chen, J.; Lin, S.-t.; and Durrett, G. 2019. Multi-hop question answering via reasoning chains. *arXiv preprint arXiv:1910.02610*.

Chen, X.; Liang, C.; Yu, A. W.; Zhou, D.; Song, D.; and Le, Q. V. 2019. Neural symbolic reader: Scalable integration of distributed and symbolic representations for reading comprehension. In *International Conference on Learning Representations*.

Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; Yih, W.-t.; Choi, Y.; Liang, P.; and Zettlemoyer, L. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2174–2184. Brussels, Belgium: Association for Computational Linguistics.

Cui, L.; Wu, Y.; Liu, S.; Zhang, Y.; and Zhou, M. 2020. MuTual: A Dataset for Multi-Turn Dialogue Reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1406–1416. Online: Association for Computational Linguistics.

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.

Ding, M.; Zhou, C.; Chen, Q.; Yang, H.; and Tang, J. 2019. Cognitive Graph for Multi-Hop Reading Comprehension at Scale. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2694–2703. Florence, Italy: Association for Computational Linguistics.

Fang, Y.; Sun, S.; Gan, Z.; Pillai, R.; Wang, S.; and Liu, J. 2020. Hierarchical Graph Network for Multi-hop Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8823–8838. Online: Association for Computational Linguistics.

Gao, Y.; Wu, C.-S.; Li, J.; Joty, S.; Hoi, S. C.; Xiong, C.; King, I.; and Lyu, M. 2020. Discern: Discourse-Aware Entailment Reasoning Network for Conversational Machine Reading. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2439–2449. Online: Association for Computational Linguistics.

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*.

Hermann, K. M.; Kociský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching Machines to Read and Comprehend. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 1693–1701.

Huang, L.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2019. 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)*, 2391–2401. Hong Kong, China: Association for Computational Linguistics.

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

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

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Li, J.; Sun, A.; and Joty, S. R. 2018. SegBot: A Generic Neural Text Segmentation Model with Pointer Network. In Lang, J., ed., *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 4166–4172. ijcai.org.

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., *Proceedings of the Twenty-Ninth International Joint*

Conference on Artificial Intelligence, IJCAI 2020, 3622–3628. ijcai.org.

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.

Ouyang, S.; Zhang, Z.; and Zhao, H. 2021. Dialogue Graph Modeling for Conversational Machine Reading. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Qiu, L.; Xiao, Y.; Qu, Y.; Zhou, H.; Li, L.; Zhang, W.; and Yu, Y. 2019. Dynamically Fused Graph Network for Multi-hop Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6140–6150. Florence, Italy: Association for Computational Linguistics.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 784–789. Melbourne, Australia: Association for Computational Linguistics.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. Austin, Texas: Association for Computational Linguistics.

Reddy, S.; Chen, D.; and Manning, C. D. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7: 249–266.

Ren, H.; and Leskovec, J. 2020. Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs. *Advances in Neural Information Processing Systems*, 33.

Sugawara, S.; Stenetorp, P.; and Aizawa, A. 2021. Benchmarking Machine Reading Comprehension: A Psychological Perspective. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1592–1612.

Sugawara, S.; Stenetorp, P.; Inui, K.; and Aizawa, A. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 8918–8927.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 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)*, 4149–4158. Minneapolis, Minnesota: Association for Computational Linguistics.

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*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380. Brussels, Belgium: Association for Computational Linguistics.

Yu, W.; Jiang, Z.; Dong, Y.; and Feng, J. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhang, Z.; Wu, Y.; Zhou, J.; Duan, S.; Zhao, H.; and Wang, R. 2020. SG-Net: Syntax-Guided Machine Reading Comprehension. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 9636–9643. AAAI Press.

Zhang, Z.; Yang, J.; and Zhao, H. 2021. Retrospective Reader for Machine Reading Comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14506–14514.

Zhang, Z.; Zhao, H.; and Wang, R. 2020. Machine reading comprehension: The role of contextualized language models and beyond. *arXiv preprint arXiv:2005.06249*.

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.