QA Reasoning Enhancement Model Based on the Fusion of Dictionary and Hierarchical Directed Graph

Yuhang Bie[1], Meiling Liu[1(✉)], and Jiyun Zhou[2]

[1] Northeast Forestry University, Harbin, China
mlliu@nefu.edu.cn
[2] Johns Hopkins University, Baltimore, USA

Abstract. In the study of question and answer system, using pre-training and pre-training language model and knowledge graph for joint reasoning still faces two challenges: one is how to effectively solve the problem of knowledge lack in the pre-training stage; the other is how to capture more local evidence after external knowledge enhancement. To address these two pain points, This paper presents a new quiz inference enhancement model——Inference Enhancement Model (InferEM), From the perspective of knowledge enhancement and the interpretability of the reasoning process, First, the dictionary information is integrated into the pre-training through synonym replacement, translation enhancement, Improve the model's ability to predict low-frequency words; Then, we propose the hierarchical digraph method, Using the hierarchical directed graph (HD-GNN) extracted from the knowledge graph, Query related neighbors' attention selection strongly related edges capture local information, Enhance the reliability of the evidence chain. In this paper, we evaluate the newly proposed model InferEM model on two dataset benchmarks in the field of common sense quiz, which outperforms the existing single quiz inference model and the existing pre-trained language and knowledge graph joint model.

Keywords: Q&A Knowledge reasoning · Hierarchical directed graph · Knowledge Graph

1 Introduction

In the big system of instant communication on the internet, for those seeking solutions to problems, the faster the problem is solved, the better. Their needs are timely and accurate. Sometimes, solving question-and-answer type problems requires daily experience, while others require specialized knowledge. This is difficult for most people to achieve, so more and more self-help QA systems have emerged. Recently, many QA tasks have gradually become more difficult, requiring not only machine understanding of the content of the question, but also relationship reasoning of entities and their relationships through the use of external knowledge [1–4], posing great challenges to QA models. For example,

A. Zaslavsky et al. (Eds.): MobiQuitous 2023, LNICST 594, pp. 485–499, 2024.
https://doi.org/10.1007/978-3-031-63992-0_32

the question in Fig. 1 requires a model to reason about the entities mentioned, inferring implicit relationships between concepts. Background knowledge like "Where can I stand on a river to see water falling without getting wet?" may be difficult for machines to understand, but it is common sense for humans, and using various knowledge to help understand the meaning of pre-training is a key human ability.

Common sense QA evaluates whether machines can understand pre-training like humans by asking questions that rely on common sense knowledge behind the answers. Initially, models relied on pre-trained language models for prediction and transferred learning based on a large amount of initial knowledge, showing excellent performance for lower difficulty QA. However, as difficulty increases, the shortcomings of pre-trained language models gradually become apparent, as the vocabulary of the model is greatly challenged by the increase in difficulty, resulting in unsatisfactory QA performance. Later, the emergence of knowledge graphs solved the problem of knowledge shortage, and the appearance of a large amount of structured data elevated the accuracy of reasoning to a new level. Of course, knowledge graphs are not without their drawbacks. The construction of knowledge graphs is too large, and the process of internal retrieval can be extremely time-consuming, leading to problems with unidentifiable neighbor nodes with too few neighbors, resulting in incomplete local information affecting the final results. Recently, some researchers have proposed a solution that combines knowledge graphs and pre-trained language models to overcome the shortcomings of low accuracy of a single model's reasoning. Although the fusion of pre-trained language models and knowledge graphs has improved the accuracy of reasoning to a certain extent, the difficulties of each still remain unresolved.

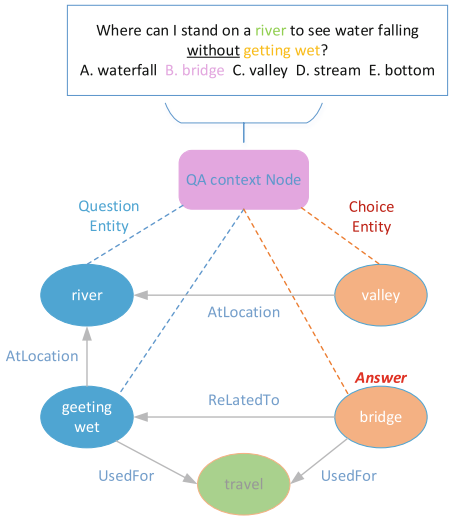

Fig. 1. Given the QA context, our goal is to derive answers through joint inference of language and knowledge maps.

In this paper, improvements were made to address the shortcomings of the two models mentioned above. Specifically, the weaknesses of GNN were addressed, and its pre-training was enhanced. The Inference Enhancement Model (Inference Enhancement Model, InferEM) was proposed, which improves on the fusion model in two ways: first, by improving the handling of low-frequency words and implicit knowledge prediction in the pre-training language model; second, by recognizing that there are limitations to capturing local evidence and that directed subgraphs are more advantageous than knowledge graphs and provide interpretability. The main contributions of this paper are as follows:

Dt-RoBERTa using the description information of the dictionary, three ways are adopted to improve the prediction ability of low-frequency words. At the same time, the self-attention mechanism is improved. One is the pool output connection, the second is the external attention, and the third is the layered external attention, which greatly improves the model efficiency and generalization ability.

We propose a hierarchical directed graph method, through the encoding and reasoning mechanism design of local information, it adopts multi-dimensional features and adopts different neighbor sampling strategies for each dimension, thus encoding different local information of nodes. The two model uses the reversible module and multi-head attention mechanism to capture and model each inference route and obtain the best inference results by choosing different inference routes. Through the choice of inference route, the model can make full use of local information, as well as global information, improving the reasoning ability.

2 Relation Work

Usually, knowledge in question answering can be implicitly encoded in large pre-trained language models on unstructured text [5, 6] or explicitly represented in structured knowledge graphs, such as Freebase [7] and ConceptNet [8], where entities are represented as nodes and their relationships are represented as edges. Recently, pre-trained language models have achieved significant success in many question answering tasks [9, 10]. However, while pre-trained language models cover a wide range of knowledge and have achieved great success, knowledge at the commonsense level is still insufficient, resulting in significant gaps between the inferred results and reality, with the most significant being the deviation in joint reasoning. In addition, pre-trained language models perform well in text-based unstructured prediction but perform poorly in structured reasoning (e.g., handling negation) [11]. On the other hand, knowledge graphs are suitable for structured reasoning [12, 13] and can make interpretable predictions, such as by providing reasoning paths [16], but may lack coverage and fail to capture implicit information effectively [15, 16]. Previous work [14, 17, 18] retrieved subgraphs from the knowledge graph by obtaining the topic entity (the knowledge graph entity mentioned in the question-answer pair) and its few-hop neighbors. However, this introduces many entity nodes that are semantically irrelevant to the question-answer pair, especially when the number of topic entities or hops increases.

Existing fusion inference methods [14, 19–21] of pre-trained language models and knowledge graphs treat them as two independent modalities. They separately apply pre-trained language models to question-answer pairs, but pre-trained language models lack

sufficient external knowledge for training and structured knowledge is not applicable to training pre-trained language models. In addition, the success in modeling graph-structured data introduces GNNs to capture subgraph structures in knowledge graphs, which solves the shortcomings of directly retrieving from the knowledge graph, while the path structure of subgraphs is conducive to the interpretability of reasoning. R-GCN [22], CompGCN [23], KE-GCN [24], and QA-GNN [25] propose updating entity representations by aggregating all neighbors of each layer through paths. However, they cannot distinguish the structural roles of different neighbors, and they are also not interpretable.

Although these methods have solved the challenges to some extent in question answering tasks, there are still two aspects to be addressed regarding the insufficient knowledge of pre-trained language models and the inability to effectively capture local information in subgraphs. Combining pre-trained language models and knowledge graphs for inference presents two challenges: (1) how to effectively solve the knowledge deficiency problem with pre-trained language models and perform joint reasoning given a question-answer pair; (2) how to strengthen the capture of local evidence behind rich external knowledge. To address these two challenges, this paper proposes the InferEM model and conducts extensive experiments on two question answering datasets, showing significant improvements over previous models.

3 Approach

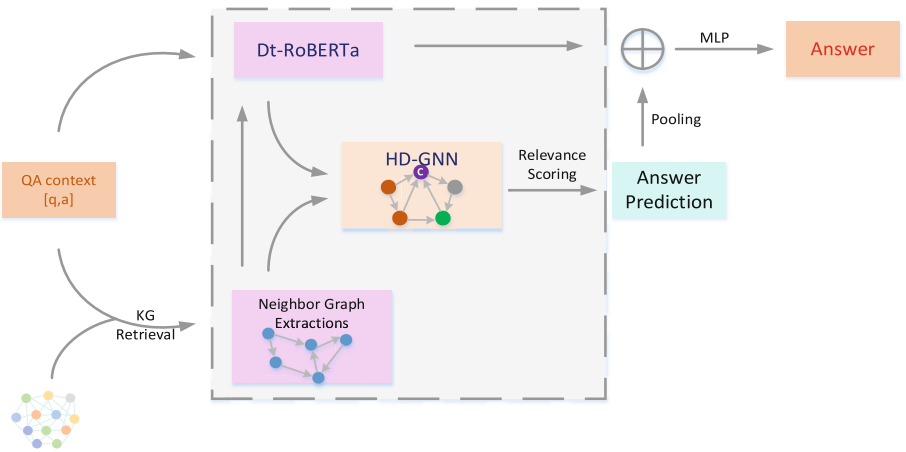

Fig. 2. Q&A inference enhancement model diagram

As shown in Fig. 2, given a question q and an answer option a, they are concatenated to form a question pair (question and answer choice) [q;a]. To perform reasoning on the given question pair using the knowledge from both the pre-trained language model and knowledge graph, InferEM works as follows. First, a node $\mathbf{C}$ representing the question pair is introduced and connected to the topic entities $\tilde{\mathrm{V}}_{\mathrm{q,a}}$, resulting in a pre-fusion graph

$\mathcal{G}_w$ with two knowledge sources (Sect. 3.1).Then, the pre-trained language model is enhanced with knowledge for the question pair using a dictionary (Sect. 3.2) to obtain its representation, and a Initial Graph $\mathcal{G}_w$ is retrieved from the knowledge graph. To adaptively capture the relationship between the question pair node and each other node, a correlation score is computed for each pair using the pre-trained language model and added as an additional feature to each node. Then, a hierarchical directed graph module based on attention is proposed, which performs layered $\mathcal{G}_w$ message passing on the pre-fusion graph (Sect. 3.3). Finally, the pre-trained language model, hierarchical directed graph, and pooled work graph representation are used for final prediction.

3.1 Initial Graph

To design a joint inference space for these two knowledge sources, we need to connect them in a common graph structure. This paper introduces a question-to-pair context node $\complement$ to represent the question-to-pair context, and uses two new relationship types γ_q^c and γ_a^c will be connected to the subgraph $\mathcal{G}_s$ on $\varphi_{q,a}$, each subject entity in $\complement$. These relationship types capture the relationship between relevant entities in the question-pair context and the knowledge graph, depending on whether the entity is found in the question or answer section of the question-to-answer context. Since this subgraph is only a preliminary construction of the connections between entities, this paper calls the pre-joint graph $\mathcal{G}_w = (\beta_w, \varphi_w)$.

$$\varphi_w = \varphi_s \cup \{\complement\} \tag{1}$$

$$\beta_w = \beta_s \cup \left(\complement, \gamma_q^c, e\right) e \in \varphi_q \cup \left(\complement, \gamma_a^c, e\right) e \in \varphi_q \tag{2}$$

Each node in $\mathcal{G}_w$ is associated with one of the four types: $\Gamma = \{\complement, Q, A, O\}$, each type represents the context node $\complement$, node in φ_q, node in φ_a and other nodes. This article represents the text of the context node and the knowledge graph node $e \in \varphi_s$ (entity name) as ($\complement$) and $text(e)$.

In this paper, we initialize the node embedding by answering a pre-trained language model representation of the context

$$\complement^{plm} = f_{enc}((\complement)) \tag{3}$$

Perfect each node on the $\mathcal{G}_s$ through its entity embedding.

3.2 Pre-trained Language Model Based on Dictionary Knowledge Enhancement

The advantage of Dt-RoBERTa lies in using the lexicon description information [36] as an external knowledge to enhance the pre-training language model, in order to improve the reasoning ability of the model and the ability to deal with problems without background knowledge. Solved the model of low frequency words in the corpus, through the low-frequency word dictionary and the corresponding low frequency word definition to enhance training pre-training language model, and introduced for the word level and sentence level of two special tasks, these information can help the model to learn the

relationship between the entity and semantics, so as to improve the performance of the model on the common sense reasoning task.

This method is mainly improved from two aspects: one is to use the dictionary knowledge to enhance the semantic information and improve the prediction ability of the model to the low frequency words; the other is to adopt a new self-attention mechanism to improve the efficiency and generalization ability of the model.

Several aspects of improving the ability to predict low-frequency words by using dictionary knowledge:

(1) **Synonym replacement**: Use synonyms in the dictionary to expand the semantic radiation range of low-frequency words. During the pre-training process, Dt-RoBERTa will randomly replace some words in the original text. If these replaced words have similar words in the dictionary, the synonym in the dictionary will be used to replace the original words, so that the model can better learn the semantic information of low-frequency vocabulary.

(2) **Translation enhancement**: A multilingual dictionary is used to translate some low-frequency words from the source language to the target language to use the broader representation of the target language. This approach can both expand the semantic radiation range of the model to low-frequency words and increase the effect on translanguage tasks.

(3) **Word sense disambiguation:** In the pre-training process, word sense disambiguation will be taken into account, that is, a common word can have multiple semantics. It will select the appropriate descriptions in multiple dictionaries to assist in the analysis of the semantic relations of words. This way can enhance the model to understand the polysemy in the low-frequency vocabulary and improve the generalization ability.

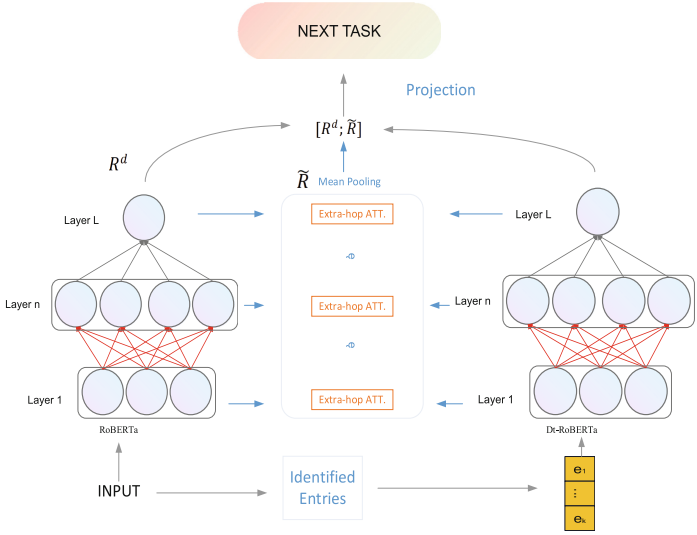

Fig. 3. Layered external hop attention

In fine-tuning, this paper uses the pre-trained-Dt-RoBERTa as a plug-in knowledge base to retrieve the implicit knowledge of identified entries in the input sequence and inject the retrieved knowledge into the input, enhancing its representation through a novel external hop attention mechanism. Specifically, the dictionary entries are first identified from a given input, then the corresponding entry information is retrieved using Dt-RoBERTa as a knowledge base, and finally the retrieved entry information is injected into the original input to obtain an enhanced representation of the downstream tasks. To better utilize the retrieved tacit knowledge in downstream tasks, three different knowledge injection mechanisms are introduced: Layered external hop attention, as shown in Fig. 3.

Layered external hop attention. To further improve the performance, the extra-hop attention of the last layer is extended to each inner layer, making it a more sensible $\tilde{R}$ layer. Attention scores were calculated for each layer, and finally with their average for implicit input knowledge injection. Specifically, the layered external jump attention can be described as follows:

$$\tilde{R}_{\mathrm{l}} = \sum_{i=1}^{\mathrm{M}} ATT\left(R_l, e_i^l\right) \tag{4}$$

$$\tilde{R} = \frac{1}{L} \sum_{l=1}^{L} \tilde{R}_{\mathrm{l}} \tag{5}$$

where, $\tilde{R}$ represents the weighted sum of the layer l output of Dt-RoBERTa. For the implicit final $\tilde{R}$ obtained by Eq. (5), a similar approach is provided for the downstream tasks using $[R^d; \tilde{R}]$.

3.3 A Hierarchical Directed Graph Based on the Graph Neural Network

In order to improve the pre-joint graph, this paper changes the GNN model and improve the directed graph. Relational paths show strong transferable and interpretable inference power on the knowledge graph [27–29]. However, because the nodes in the pathway are only connected in order, they are limited in capturing the more complex dependencies in the knowledge graph. The GNN-based methods can learn different subgraph structures. However, none of the existing methods can effectively learn subgraph structures that are both interpretable and inductive as rules. Therefore, this paper proposes a new structure to capture important local evidence, using the following two methods: local information coding and inference mechanism design.

Local information coding: In the feature representation of nodes, not only the global information between nodes, but also the local information of nodes themselves, is taken into account. Specifically, the model employs multidimensional features and different neighbor sampling strategies for each dimension, thus encoding different local information of the nodes. In this way, the feature vectors of nodes can contain more rich local information, thus improving the ability of the model to capture local information.

Reasoning mechanism design: a mechanism based on the selection of reasoning route is adopted. From the original language information to the reasoning result, each

reasoning route will be modelled. Each node from linguistic information to inference results can be regarded as a tiny inference route. The model uses the reversible module and multi-head attention mechanism to capture and model each inference route, and obtain the best inference results by choosing different inference routes. Through the choice of inference route, the model can make full use of local information, as well as global information, improving the reasoning ability.

Inspired by RED-GNN [39], in this paper, we introduce a special type of directed graph-hierarchical graph. The hierarchical graph is a directed graph with only one source node (s) and one convergence node (t). All edges are directed, with nodes between successive layers and pointing from layer m to layer $m+1$. A general approach is used to augment the triplet [23, 28] with inverse and identity relations. Then, all relational paths of length less than or equal to L between e_q and e_a can be represented as relational paths of length $e_q \rightarrow_{r^1} \cdot \rightarrow_{r^2} \cdot \cdots \rightarrow_{r^L} e_a$. In this way, they can be formed as paths in hierarchical graphs with a single source entity e_q and a converging entity e_a. This structure preserves all the relational pathways between e_q and e_a and maintains the subgraph structure. Based on this result, a hierarchical graph is introduced.

The hierarchical graph $\mathcal{G}_{\mathrm{e_q,e_a|L}}$ is a hierarchical graph with the source entity e_q and the sink entity e_a The entities in the same layer are different from each other. Any path from e_q to e_a in the hierarchy graph is a relational path L of length $e_q \rightarrow_{r^1} \cdot \rightarrow_{r^2} \cdot \cdots \rightarrow_{r^L} e_a$, where r^m connects the entities in the $m-1$ layer with the entities in the m layer. If there is no relational path connecting e_q and e_a, define $\mathcal{G}_{\mathrm{e_q,e_a|L}} = 0$.

Inspired by the inference ability of relational pathways, the goal of this paper is to perform knowledge graph inference using directed graphs. However, unlike learning structurally simple relational paths with the sequential model [27, 30], how to efficiently construct and efficiently learn from directed graphs is challenging.

In this paper, $\beta^m_{e_q,e_a|L}$ is the edge, $\varphi^m_{e_q,e_a|L} = e_o|(e_s, r, e_o) \in \beta^m_{e_q,e_a|L}$ is an entity in the r layer of the directed graph $\mathcal{G}_{\mathrm{e_q,e_a|L}} = \beta^1_{e_q,e_a|L} \otimes \cdots \otimes \beta^L_{e_q,e_a|L}$, where hierarchical connectivity. Deine the joint operator as

$$\mathcal{G}_{e_{q1},e_{a1}|L} \cup \mathcal{G}_{e_{q2},e_{a2}|L} = \beta^L_{e_{q1},e_{a1}|L} \cup \beta^L_{e_{q2},e_{a2}|L} \otimes \ldots \otimes \beta^L_{e_{q1},e_{a1}|L} \cup \beta^L_{e_{q2},e_{a2}|L} \tag{6}$$

Given an entity e, represent $\hat{\beta}^m_e$, $\overset{\smile}{\beta}{}^m_e$, and φ^m_e as sets of outer, inner, and entities, from which e is seen walking through the r jump path.

Here, we show how GNN can be improved to learn efficiently and efficiently from directed graphs. Extracting the subgraph structure and then learning the subgraph representation is a common practice for subgraph encoding in the literature, such as GraphSage [31] and GraIL [32]. Given a query triad (e_q, r_q, e_a), subgraph coding generally includes three processes:

(1) Exthe neighbors of e_q and e_a;
(2) Take the intersection to construct the subgraph;
(3) Message messaging and use graph level representation as subgraph encoding;

When working on the directed graph $\mathcal{G}_{e_q,e_a|L}$, first, the neighborhoods of e_q and e_a are obtained in the knowledge graph. Second, take the intersection of the neighborhoods of e_q and e_a to derive the directed graph $\mathcal{G}_{e_q,e_a|L}$, layer by layer. Third, if the digraph is

empty, set $\mathcal{G}_{e_q,e_a|L}$ to 0. Otherwise, the messaging is delivered layer by layer. Since e_a is a single junction entity, the final layer $h^L_{e_a}(e_q, r_q)$ is used as a subgraph representation to encode the directed graph $\mathcal{G}_{e_q,e_a|L}$. This paper names this solution as Hierarchical diagrams.

When using a different $e_a \in \varphi$, but the same query $(e_q, r_q, ?)$ To evaluate (e_q, r_q, e_a), the e_q of the adjacent edge $\widehat{\beta}^m_{e_q}$, $m = 1...L$ is shared. The following observations were made with:

The edge set visible from e_q by m step $\widehat{\beta}^m_{e_q}$ equals $\cup_{e_a \in \varphi} \beta^m_{e_q,e_a|L}$, namely $\beta^m_{e_q}$ is the union of layer r edges in the directed graph between e_q and all entities $e_a \in \varphi$. Due to the existence of sharing the same edge set, the overlap problem can affect the computational cost, and the common method to solve this problem is dynamic programming. This has been used to aggregate node representations on large-scale graph [31], or in the representation of the knowledge graph propagation problem. Inspired by the efficient low books of dynamic programming, this paper recursively constructs the directed graph between e_q and all entities e_o as

$$\mathcal{G}_{e_q,e_0|m} = U_{(e_s,r,e_o)\in\widehat{\beta}^m_{e_q}} \mathcal{G}_{e_q,e_s|m-1} \otimes \left\{(e_s, r, e_0) \in \widehat{\beta}^m_{e_q}\right\} \tag{7}$$

Once the $m-1$ of all entities $e_s \in \varphi^{m-1}_{e_q}$ in the $\mathcal{G}_{e_q,e_s|m-1}$ layer, $\mathcal{G}_{e_q,e_s|m-1}$ indicates updated to encode by combining $\mathcal{G}_{e_q,e_0|m}$ with shared edges $(e_s, r, e_o) \in \widehat{\beta}^r_{e_q}$ in layer m. Based on the above statement and formula (7), multiple directed diagrams can be effectively encoded recursively in $\widehat{\beta}^r_{e_q}$.

3.4 Accuracy and Interpretability Design of QA Reasoning

Dt-RoBERTa has used a large amount of lexicon knowledge as pre-training data that can help the model understand semantic relationships and thereby improve the model performance on common sense inference tasks. Adding hierarchical directed graphs to pre-training can provide a richer and organized knowledge representation, thus further improving model performance and interpretability.

A hierarchial directed graph is a graphical structure used to represent a conceptual hierarchy. In the common sense reasoning task, concept hierarchy is very important for reasoning because it can help the model understand the relationships and hierarchy between concepts and thus reasoning.

Given the different triples sharing the same hierarchical graph, the local evidence used for inference in this paper is different. In order to obtain knowledge of query-related from the hierarchical graph and find interpretable local evidence, this paper uses the attention mechanism [33], encoding r_q into attention weights to control the importance of different edges in $\mathcal{G}_{e_q,e_0|m}$. The messaging function is specified as

$$h^m_{e_o}(e_q, r_q) = \delta\left\{W^m \cdot \sum_{(e_s,r,e_o)\in\widehat{\mathcal{E}}^m_{e_q}} \alpha^m_{e_s,r,e_o|r_q}\left[h^{m-1}_{e_s}(e_q, r_q) + h^m_r\right]\right\} \tag{8}$$

The attention weights $\alpha^m_{e_s,r,e_o|r_q}$ are at the edge (e_q, r, e_a) are

$$\alpha^m_{e_s,r,e_q|r_q} = \sigma\left(\left(w^m_\alpha\right)^\top \mathrm{ReLU}\left(W^m_\alpha \cdot \left(h^{m-1}_{e_s}\left(e_q, r_q\right) \oplus h^m_r \oplus h^m_{r_q}\right)\right)\right) \tag{9}$$

where $w_\alpha^m \in \mathbb{R}^{d_\alpha}$, $W_\alpha^m \in \mathbb{R}^{d_\alpha \times 3d}$ and $\oplus$ are the connection operator. Use the Sigmod function σ to ensure that multiple edges can be selected in the same neighborhood.

After converging the L layer through (8), the representation $h_{e_a}^m(e_q, r_q)$ can encode the basic information for scoring (e_q, r_q, e_a). Therefore, a simple scoring function was used

$$f(e_q, r_q, e_a) = w^\top h_{e_a}^L(e_q, r_q) \tag{10}$$

where the $w^\top \in \mathbb{R}^d$. Link the multiclass log loss [34] to each triplet (e_q, r_q, e_a), i. e

$$\sum\nolimits_{(e_q, r_q, e_a) \in \mathcal{T}_{\text{tra}}} \left\{ -f\left(e_q, r_q, e_a\right) + \log\left(\sum\nolimits_{\forall e \in \varphi} e^{f(e_q, r_q, e)} \right) \right\} \tag{11}$$

The first part in formula (11) is the score of the positive triad $\left(e_q, r_q, e_a\right)$ in $\mathcal{T}_{\text{tra}}$, which is the set of training queries, the second part contains the same query $\left(e_q, r_q, ?\right)$ The fraction of all the triples of the vs. The model parameters $\theta = \left\{\{W^m\}, \{w_\alpha^m\}, \{W_\alpha^m\}, \{h_r^m\}, w\right\}$ are randomly initialized and optimized by random gradient descent [35] minimization (12). If a set of relational pathways is strongly correlated with the query triplet, they can be interpretable by identifying the attention weights in the Hierarchical diagrams.

4 The Experiment

All experiments were written using the PyTorch framework and run on an NVIDIA A100 GPU with 80GB of memory. The test values of this model are referenced using the common measure of common knowledge and answers, namely accuracy (Acc).

4.1 Data Set

This paper evaluates InferEM on three question answer datasets: CommonsenseQA [1] and OpenBookQA [4].

CommonsenseQA is a five-choice question and answer task that requires common sense reasoning and contains 12,102 questions. The test set of CommonsenseQA is not publicly available, and the model prediction evaluation can only be published every two weeks through the official leaderboard. Therefore, the main experiment on the in-house (IH) data segmentation used in [14] by Lin et al.

OpenBookQA is a four-choice question and answer task that requires reasoning using basic scientific knowledge and contains 5,957 questions. This paper uses the official data segmentation from the [4] of Mihaylov et al.

4.2 Knowledge Graph

For CommonsenseQA and OpenBookQA, this paper uses ConceptNet, a general domain knowledge graph, as a structured knowledge source $\mathcal{G}$ for this paper. It has a total of 799,273 nodes and 2,487,810 edges. The node embeddings are initialized using the entity embedding prepared by Feng et al. [20], which applies the pre-trained trained language

model to all triples in ConneceptNet, and then obtains the ensemble representation of each entity.

Given each question and answer pair, this paper retrieved subgraph $\mathcal{G}_s$ from $\mathcal{G}$ following the pre-processing step described in Feng et al. Then, the $\mathcal{G}_s$ was trimmed to maintain the first 200 nodes from the model of Sect. 3.1.

4.3 Baseline

For fine-tuned pre-trained language models, pre-trained language models without the knowledge graph are used for comparison. Testing on the CommonsenseQA dataset using RoBERTa-larga [9] and Dt-RoBERTa and on the OpenBookQA dataset using RoBERTa-larga, AristoRoBERTa [37] and Dt-RoBERTa.

For knowledge graph and pre-trained language models, compare existing knowledge graph and pre-trained language models, sharing the same knowledge graph architecture, but using different graph neural network models. Comparing the following state-of-the-art baseline models: (1) Relation Network [38], (2) RGCN [22], (3) GconAttn [19], (4) KagNet [14], (5) MHGRN [20], and (6) QA-GNN [25]. (1), (2), (3) the GNN of the perceptual relationship of KGs, and (4) and (5) use the path in the knowledge graph for reasoning. QA-GNN uses question and answer pairs to score correlations and update them jointly. QA-GNN is the existing top performance model in the framework of this knowledge graph and pre-trained language model. For a fair comparison, the paper uses the same pre-trained language model across all baselines and the models in this paper. The key difference between InferEM and these models is that (1) none of them focuses on the importance of external knowledge of the pre-trained language model, resulting in the failure to accurately predict implicit information; (2) previous models are not good at capturing local evidence, which causes less local information on relationship edges.

4.4 Main Results

Tables 1 and 2 show the results for CommonsenseQA and OpenBookQA, respectively. On both datasets, this paper observed consistent improvement over fine-tuned pre-trained and existing knowledge map and pre-trained language model models, for example, on CommonsenseQA, 5.7% over RoBERTa and 0.82% over the previous best-knowledge map and pre-trained language model QA-GNN. Performance over the QA-GNN model shows that InferEM better uses KGs to perform joint inference than existing knowledge graph and pre-trained language model methods.

CommonsenseQA Accuracy comparison of the internal segmentation. Since the official test is hidden, this paper reports here the accuracy of internal Dev (IHdev) and Test (IHtest) after the data segmentation of Lin et al. The experimental results show that InferEM* (no pre-trained language model using RoBERTa-large) is better than the other models. The InferEM was more significantly.

The test accuracy of the OpenBookQA was compared. The first row of the table uses RoBERTa-lag, AristoRoBERTa, and Dt-RoBERTa as additional input for the quiz pair. The results showed that Dt-RoBERTa performed the best. The first column of the table is a comparison of the baseline of the current graph structure model. The experimental results show that the same pre-trained language model InferEM* effect outperforms the

Table 1. Performance comparison on Commonsense QA in-house split.

Methods	IHdev-Acc.(%)	IHtest-Acc.(%)
RoBERTa-large (w/o KG)	73.07	68.69
+ RGCN	72.69	68.41
+ GconAttn	72.61	68.59
+ KagNet	73.47	69.01
+ RN	74.57	69.08
+ MHGRN	74.45	71.11
+ QA-GNN	76.54	73.41
+ InferEM*	76.87	73.74
InferEM(Ours)	**77.13**	**74.56**

Table 2. Comparing the testing accuracy of OpenBook QA with pre training

Methods	RoBERTa-large	AristoRoBERTa	Dt-RoBERTa
RGCN	62.45	74.60	78.11
GconAtten	64.75	71.80	78.34
RN	65.20	75.35	79.86
MHGRN	66.85	80.60	81.17
QA-GNN	67.80	82.77	83.29
InferEM*(Ours)	**68.43**	**83.31**	**84.17**

previous methods and outperforms the latest QA-GNN model. Moreover, InferEM was also higher than the existing models, with the best performance.

4.5 Ablation Experiments

As shown in the ablation experiment in Fig. 4, it is concluded that the Dt-RoBERTa pre-trained language model with lexicon knowledge enhancement can improve the lack of pre-training knowledge, and the performance of Dt-RoBERTa is stronger than other models in prediction.

Depth of the InferEM model. In Fig. 5, this paper shows the use of different layer L on the x-axis to test the effect on accuracy and the coverage of test triples (e_q, r_q, e_a). e_a is visible in e_q. Intuitively, when L is increased, more triples will be overlaid and the path or subgraph between e_q and e_a will contain more information but will be more difficult to learn. As shown in the figure, when L is too small, for as $L \leq 2$, the performance of InferEM is relatively poor, mainly due to the limited information encoded in the hierarchical graph in such a small range. When $L \geq 3$, InferEM achieves the best

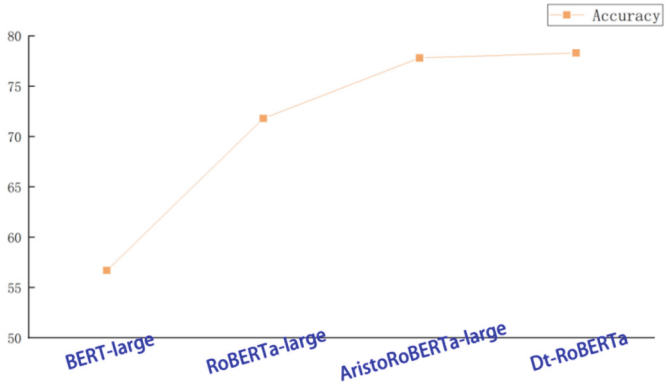

Fig. 4. Accuracy corresponding to different pre trained models

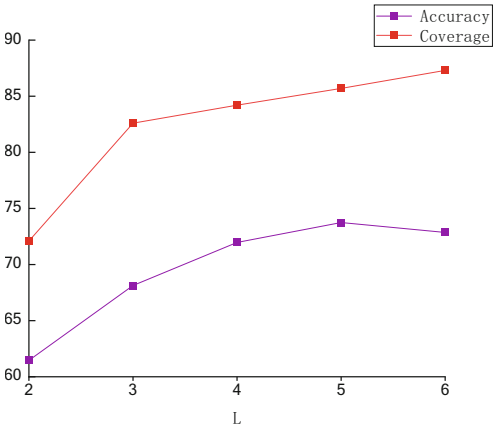

Fig. 5. The impact of hierarchy on accuracy and coverage

performance, where hierarchical graphs can contain richer information, and important information of inference can be learned efficiently.

5 Conclusion

This paper presents InferEM, an end-to-end question answering model utilizing knowledge graph and pre-trained language models. The key innovations of this paper include two points: first, to improve the vocabulary knowledge of pre-trained language model, especially the case of insufficient low frequency words, using Dt-RoBERTa to improve the predictive reasoning ability of the model; and second, to solve the problem of capturing insufficient internal evidence, the D-GNN model is used to improve GNN and interpretable. Through quantitative and qualitative analysis, this paper shows that InferEM has significant improvement over the existing pre-trained language model, knowledge map and pre-trained language model, and has the ability to explain and structural reasoning.

The following work, on the one hand, is to train the Chinese corpus, to compensate for the field and improve the robustness of the model; on the other hand, to develop more interpretable models to transform the reasoning process into a natural language form to better explain the decision of the model. In addition, the knowledge graph can also be extended and updated, because the knowledge graph is the basis of the common sense reasoning model and plays a key role.

Acknowledgment. This work was supported by National Natural Science Foundation of China 61702091 and Heilongjiang Natural Science Foundation Project LH2022F002.

References

1. TaLMor, A., Herzig, J., Lourie, N., et al.: Commonsenseqa: a question answering challenge targeting commonsense knowledge. arXiv preprint arXiv:1811.00937 (2018)
2. Sap, M., Rashkin, H., Chen, D., et al.: Socialiqa: commonsense reasoning about social interactions. arXiv preprint arXiv:1904.09728 (2019)
3. Clark, P., Cowhey, I., Etzioni, O., et al.: Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457 (2018)
4. Mihaylov, T., Clark, P., Khot, T., et al.: Can a suit of armor conduct electricity? a new dataset for open book question answering. arXiv preprint arXiv:1809.02789 (2018)
5. Petroni, F., Rocktäschel, T., Lewis, P., et al.: Language models as knowledge bases?. arXiv preprint arXiv:1909.01066 (2019)
6. Bosselut, A., Rashkin, H., Sap, M., et al.: COMET: Commonsense transformers for automatic knowledge graph construction. arXiv preprint arXiv:1906.05317 (2019)
7. Bollacker, K., Evans, C., Paritosh, P., et al.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1247–1250 (2008)
8. Speer, R., Chin, J., Havasi, C.: C. N. 5.5: an open multilingual graph of general knowledge. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 4444–4451, December 2016
9. Liu, Y., Ott, M., Goyal, N., et al.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
10. Raffel, C., Shazeer, N., Roberts, A., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)
11. Kassner, N., Schütze, H.: Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. arXiv preprint arXiv:1911.03343 (2019)
12. Ren, H., Hu, W., Leskovec, J.: Query2box: reasoning over knowledge graphs in vector space using box embeddings. arXiv preprint arXiv:2002.05969 (2020)
13. Ren, H., Leskovec, J.: Beta embeddings for multi-hop logical reasoning in knowledge graphs. Adv. Neural. Inf. Process. Syst. **33**, 19716–19726 (2020)
14. Lin, B.Y., Chen, X., Chen, J., et al.: Kagnet: knowledge-aware graph networks for commonsense reasoning. arXiv preprint arXiv:1909.02151 (2019)
15. Bordes, A., Usunier, N., Garcia-Duran, A., et al.: Translating embeddings for modeling multi-relational data. Adv. Neural Inf. Process. Syst. **26** (2013)
16. Guu, K., Miller, J., Liang, P.: Traversing knowledge graphs in vector space. arXiv preprint arXiv:1506.01094 (2015)
17. Bao, J., Duan, N., Yan, Z., et al.: Constraint-based question answering with knowledge graph. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 2503–2514 (2016)

18. Sun, H., Dhingra, B., Zaheer, M., et al.: Open domain question answering using early fusion of knowledge bases and text. arXiv preprint arXiv:1809.00782 (2018)
19. Wang, X., Kapanipathi, P., Musa, R., et al.: Improving natural language inference using external knowledge in the science questions domain. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 7208–7215 (2019)
20. Feng, Y., Chen, X., Lin, B.Y., et al.: Scalable multi-hop relational reasoning for knowledge-aware question answering. arXiv preprint arXiv:2005.00646 (2020)
21. Lv, S., Guo, D., Xu, J., et al.: Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05 pp. 8449–8456 (2020)
22. Schlichtkrull, M., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: Gangemi, A., et al. The Semantic Web, ESWC 2018, LNCS, vol. 10843, pp. 593–607. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_38
23. Vashishth, S., Sanyal, S., Nitin, V., et al.: Composition-based multi-relational graph convolutional networks. arXiv preprint arXiv:1911.03082 (2019)
24. Yu, D., Yang, Y., Zhang, R., et al.: Knowledge embedding based graph convolutional network. In: Proceedings of the Web Conference, vol. 2021, pp. 1619–1628 (2021)
25. Yasunaga, M., Ren, H., Bosselut, A., et al.: QA-GNN: Reasoning with language models and knowledge graphs for question answering. arXiv preprint arXiv:2104.06378 (2021)
26. Zhao, C., Xiong, C., Rosset, C., et al.: Transformer-xh: Multi-evidence reasoning with extra hop attention (2020)
27. Das, R., Dhuliawala, S., Zaheer, M., et al.: Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. arXiv preprint arXiv:1711.05851 (2017)
28. Sadeghian, A., Armandpour, M., Ding, P., et al.: Drum: end-to-end differentiable rule mining on knowledge graphs. Adv. Neural Inf. Process. Syst. **32** (2019)
29. Yang, F., Yang, Z., Cohen, W.W.: Differentiable learning of logical rules for knowledge base reasoning. Adv. Neural Inf. Process. Syst. **30** (2017)
30. Xiong, W., Hoang, T., Wang, W.Y.: Deeppath: a reinforcement learning method for knowledge graph reasoning. arXiv preprint arXiv:1707.06690 (2017)
31. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. Adv. Neural Inf. Process. Syst. **30** (2017)
32. Teru, K., Denis, E., Hamilton, W.: Inductive relation prediction by subgraph reasoning. In: International Conference on Machine Learning, pp. 9448–9457. PMLR (2020)
33. Veličković, P., Cucurull, G., Casanova, A., et al.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
34. Lacroix, T., Usunier, N., Obozinski, G.: Canonical tensor decomposition for knowledge base completion. In: International Conference on Machine Learning, pp. 2863–2872. PMLR (2018)
35. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
36. Chen, Q., Li, F.L., Xu, G., et al.: DictBERT: Dictionary description knowledge enhanced language model pre-training via contrastive learning. arXiv preprint arXiv:2208.00635 (2022)
37. Clark, P., Etzioni, O., Khot, T., et al.: From 'F' to 'A'on the NY regents science exams: an overview of the aristo project. AI Mag. **41**(4), 39–53 (2020)
38. Santoro, A., Raposo, D., Barrett, D.G., et al.: A simple neural network module for relational reasoning. Adv. Neural Inf. Process. Syst. **30** (2017)
39. Zhang, Y., Yao, Q.: Knowledge graph reasoning with relational digraph. In: Proceedings of the ACM Web Conference, vol. 2022, pp. 912–924 (2022)