

Joint Reasoning with Knowledge Subgraphs for Multiple Choice Question Answering

Qin Zhang^a, Shangsi Chen^a, Meng Fang^b and Xiaojun Chen^{a,*}

^aCollege of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong, China

^bDepartment of Computer Science, University of Liverpool, Liverpool, L69 7ZX, UK

ARTICLE INFO

Keywords:

Multi-choice question answering

Multiple knowledge graphs

Joint reasoning

ABSTRACT

Humans are able to reason from multiple sources to arrive at the correct answer. In the context of Multiple Choice Question Answering (MCQA), knowledge graphs can provide subgraphs based on different combinations of questions and answers, mimicking the way humans find answers. However, current research mainly focuses on independent reasoning on a single graph for each question-answer pair, lacking the ability for joint reasoning among all answer candidates. In this paper, we propose a novel method *KMSQA*, which leverages multiple subgraphs from the large knowledge graph *ConceptNet* to model the comprehensive reasoning process. We further encode the knowledge graphs with shared Graph Neural Networks (GNNs) and perform joint reasoning across multiple subgraphs. We evaluate our model on two common datasets: CommonsenseQA (CSQA) and OpenBookQA (OBQA). Our method achieves an exact match score of 74.53% on CSQA and 71.80% on OBQA, outperforming all eight baselines.


1. Introduction

Question answering (QA) is an important application of natural language processing (NLP). It is human nature to answer questions by utilizing comprehensive knowledge. Similarly, external knowledge can also provide additional information to QA systems for reasoning and answering questions. There are two primary sources of external knowledge are: large pre-trained language models (LMs) (Petroni et al., 2019; Bosselut et al., 2019b) and structured knowledge graphs (KGs) (Bollacker et al., 2008; Speer et al., 2017). It is possible to use these resources to enhance the performance of quality assurance systems by providing a wider range of information to draw from when answering questions.

On the one hand, pre-trained language models (LMs) have been shown to improve question answering systems and perform well (Liu et al., 2019b) due to their strong knowledge encoding abilities and a large amount of unstructured text on which they have been trained (Bosselut et al., 2019a). Although pre-trained LMs possess a wide range of knowledge, they perform poorly on structured reasoning tasks (Yasunaga et al., 2021; Kassner and Schütze, 2020). On the other hand, structured knowledge graphs such as *Freebase* (Bollacker et al., 2008) and *ConceptNet* (Speer et al., 2017), where nodes represent concept entities and edges denote relationships between them, can compensate for the weaknesses of pre-trained LMs (Lin et al., 2019). Models that combine LMs and knowledge graphs (KGs), such as Knowledge-aware graph networks (KagNet) (Lin et al., 2019), Multi-hop graph relation networks (MHGRN) (Feng et al., 2020), and QA-GNN (Yasunaga et al., 2021), have achieved significant success. However, existing LM+KG methods mainly focus on addressing the challenge of inconsistencies between knowledge graph embeddings and natural language embeddings, and effectively utilizing both implicit and explicit knowledge for reasoning.

Besides the challenge of inconsistencies of dense question/answer-choice representations, in this paper, we explore further on the knowledge subgraph construction and reasoning paradigm. Our idea is inspired by the way humans answer questions. When reading a question, humans compare and reason comprehensively, using all available information, especially all the given answer candidates. Besides the semantic information in each candidate, a comprehensive glance at and comparison among all candidates can help in making the final decision, since the method of elimination is a quite useful strategy for multiple choice questions. Fig. 1 illustrates the construction of the local subgraph (see Fig. 1(b) and Fig. 1(c)) for each pair of question and answer candidate¹, as well as a global knowledge subgraph (as shown

*Corresponding author

 qinzhang@szu.edu.cn (Q. Zhang); chenshangsi2021@email.szu.edu.cn (S. Chen);

Meng.Fang@liverpool.ac.uk (M. Fang); xjchen@szu.edu.cn (X. Chen)

ORCID(s): 0000-0002-1449-5046 (Q. Zhang)

¹We omit the local subgraphs for “B. Hear sounds”, “D. Arthritis” and “E. Making music” to save space.

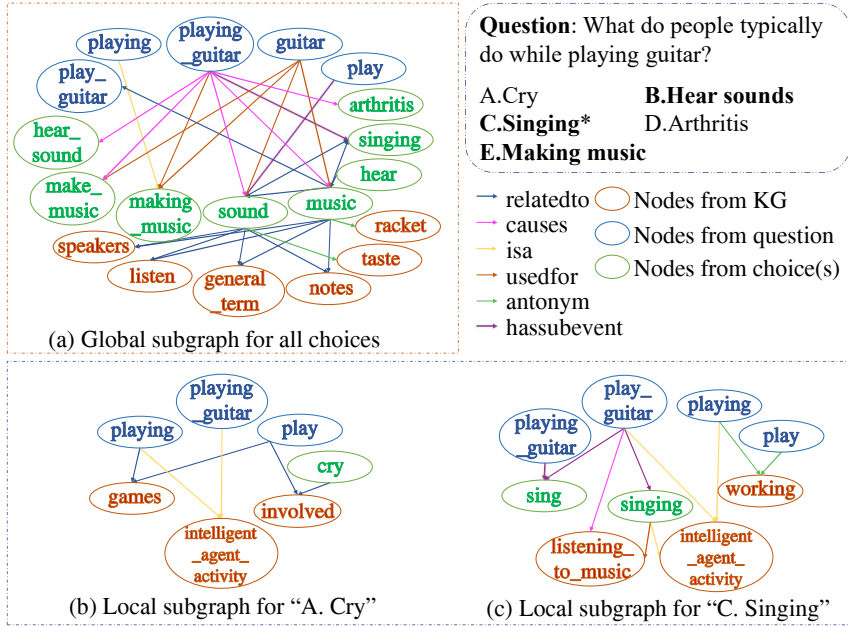


Fig. 1. An example of question answering from the CommonsenseQA dataset is shown in the right top box, where the symbol (*) marks the correct answer. The example demonstrates that there are stronger relationships among the bolded answer candidates, and weaker or no relationships among the other candidates. Fig. 1(a) is a large global subgraph including all entities that appear in the question and the five answer candidates. Fig. 1(b) is the local subgraph for candidate A, and Fig. 1(c) is the local subgraph for the correct answer C. Each local subgraph only includes the entities that appear in the question and the individual candidate. The relationships between nodes are extracted from ConceptNet. Only 20 nodes and their edge relations are displayed for the local subgraphs and 50 for the global subgraph.

in Fig. 1(a)) for the question and all the answer candidates. Combining both global and local knowledge graphs can provide multiple reasoning chains for answering a question, and enhance the ability of joint reasoning and potential information mining across multiple knowledge subgraphs. Additionally, we use a gate mechanism to control the flow of information between the multiple subgraphs.

In summary, this paper presents a novel method for addressing the problem of Multiple Choice Question Answering (MCQA) that uses a pre-trained Language Model (PLM) and multiple knowledge subgraphs to provide multi-chain reasoning and performs joint reasoning over all candidates. Our approach involves extracting two types of knowledge subgraphs from a large Knowledge Graph (KG), specifically, a local subgraph and a global subgraph. The local subgraph is constructed for a specific question-answer pair and the global subgraph is built based on the question and all its answer candidates. We then use a pre-trained Language Model (PLM) to learn the semantic features of the question-answer pair and a Graph Attention Network (GAT) to capture the relation features of topic entities in the KG subgraphs. The learnt semantic features are inserted into both subgraphs as interaction nodes respectively to implement joint reasoning over the PLM and KGs. The local subgraph focuses only on the information in the particular question-answer pair to provide a local view while the global subgraph aggregates the information in the question and all its answer candidates to offer a global view. The knowledge learned from the two subgraphs is complementary, thus joint reasoning on these subgraphs facilitates answer reasoning. Additionally, we apply a gate mechanism to control the flow of information between the multiple subgraphs.

In particular, our approach differs from most existing LM+KG methods (Yasunaga et al., 2021; Sun et al., 2022; Yasunaga et al., 2022), which mainly focus on the challenge of fusion of language models and knowledge graphs, and try to solve the problem of inconsistencies between knowledge graph embeddings and natural language embeddings, the main purpose of this paper is to explore further on the knowledge subgraph construction and answer reasoning paradigm. In terms of knowledge graph based reasoning, the existing methods either transfer a multiple choice question to several True/False questions (Feng et al., 2020; Lin et al., 2019; Lv et al., 2020; Yasunaga et al., 2021), i.e., whether candidate a_i is the right answer of the question q or not ($i = 0, \dots, N - 1$), or transfer a multiple-choice question to

an open-domain question, i.e., retrieve the knowledge subgraph and find the recommended answer then choose the most similar candidate as the final choice (Atzeni et al., 2021; Cao and Liu, 2022). Neither of these ways make full use of the inherent advantage of multiple choice question, i.e. there must exist a right answer and there is also only one right answer among the candidates. Inspired by this and how humans do multiple choice question answering, we propose our method by combining both local and global knowledge subgraphs to explore the multiple reasoning chains of answering and make full use of the elimination paradigm. The evaluation of our method on CommonsenseQA and OpenBookQA proves its superiority.

2. Related Work

2.1. Question Answering

In recent years, questions answering (QA) tasks have gained increasing interest in natural language processing (NLP) (Seonwoo et al., 2020; Zhang et al., 2022a). Some researchers focus on extractive question answering (EQA), where the answer is selected from the given context. Many EQA models (Seonwoo et al., 2020) are based on large pre-trained language models such as BERT or RoBERTa, and rely on fine-tuning with a large training dataset to adapt to EQA tasks. Additionally, some works (Deng et al., 2021) perform pre-training on existing pre-trained language models with different pre-training objectives that are closer to the final EQA task, allowing the models to be fine-tuned on smaller training datasets. Others focus on reasoning-based question answering, such as multi-choice question answering (MCQA) (Hu et al., 2021). Unlike MCRC, our target task is multi-choice question answering (MCQA), which performs reasoning over the given question context without explicit passage evidence.

Some works make full use of the information in the given context by computing the similarity between the question and each candidate through an attention mechanism (Chaturvedi et al., 2018). The Generation-enhanced MCQA model (GenMC) generates additional evidence from the question to enhance reasoning (Huang et al., 2022). However, the evidence provided by the question context is limited, which constrains the model's reasoning ability. To supplement more evidence, researchers have turned to collecting structured knowledge from *Wikipedia* and *ConceptNet* (Lin et al., 2019) or capturing unstructured information from various Internet data (Emami et al., 2018). In this paper, we leverage evidence from the given question and candidates, as well as knowledge graphs extracted from ConceptNet, to reason and predict the answer.

2.2. Knowledge Graph

In recent years, Knowledge Graphs (KGs) have been widely applied in natural language processing tasks for reasoning (Xu et al., 2020; Zhang et al., 2022b; Xie et al., 2022; Ryu et al., 2022; Xu et al., 2022). They have also been used in question answering tasks to enhance the reasoning abilities of models (Cao et al., 2019) and supplement the limited evidence provided by the question context. Some existing QA-related works (Lin et al., 2019; Feng et al., 2020; Yasunaga et al., 2021; Zhang et al., 2022c) have similar frameworks with our model, which perform joint reasoning over pre-trained language models and knowledge graphs. They mainly calculate the probability of one answer candidate being the correct answer via reasoning over a knowledge subgraph. Similar to the work of Yasunaga et al. (2021), we use a graph attention network to learn the representations of the subgraphs. The primary distinction of our work is that we construct an additional global knowledge subgraph based on the information of all the candidates. We utilize this knowledge subgraph to capture additional contrastive information among answer candidates and use it to revise the probability of one answer candidate.

2.3. Graph Neural Networks

Graph Neural Networks (GNNs), introduced as a generalization of recursive neural networks to directly deal with a more general class of graphs, e.g. cyclic, directed and undirected graphs, are a powerful tool for machine learning on graphs (Hu et al., 2020). The most classic models are Graph Convolution Network (GCN) (Kipf and Welling, 2016) and Graph Attention Network (GAT) (Veličković et al., 2017). There are several works that use GNNs to model the structure of text (Sun et al., 2018) or Knowledge Graphs (Wang et al., 2020). Recent studies have explored applying GNNs to KG-powered QA, where GNNs naturally fit the graph-structured knowledge and show prominent results (Santoro et al., 2017; Lin et al., 2019; Feng et al., 2020; Yasunaga et al., 2021; Sun et al., 2022; Huang et al., 2021; Wang et al., 2021). Knowledge-aware graph networks (KagNet) proposes GCN-LSTM-HPA for path-based relational graph representation (Lin et al., 2019). Multi-hop graph relation network (MHGRN) (Feng et al., 2020) extends relation networks (Santoro et al., 2017) to multi-hop relation scope and unifies both path-based models. QA-GNN (Yasunaga et al., 2021) proposes

a LM+GAT framework to joint reasoning over language and KG. Joint reasoning between LM and KG (JointLK) focuses on further effectively fusing the information from the pre-trained language model and GNN module through a dense bidirectional attention (Sun et al., 2022). Graph soft counter (GSC) has reported the research about how GNN works for common sense reasoning in question answering tasks (Wang et al., 2021). The aforementioned works only consider a single knowledge subgraph for one QA pair, whereas our work considers joint reasoning over multiple subgraphs.

3. Problem Definition

In this paper, we focus on the problem of multi-choice question answering, where each example consists of one question and multiple answer candidates. The goal is to identify the correct answer among the given candidates. An example is shown in Fig. 1 as follows:

–Question: “What do people typically do while playing guitar?”

–A: “Cry”

–B: “Hear sounds”

–C: “**Singing**”

–D: “Arthritis”

–E: “Making music”

where the correct answer is marked in bold black.

To aid in this task, we assume the availability of a large-scale knowledge graph, such as *ConceptNet* (Speer et al., 2017). The graph is designed to represent general knowledge that is useful for understanding language, and can be used to improve natural language applications by providing additional context and information about the meanings of words. For each example, we have a question q and a set of answer candidates $A = \{a_0, \dots, a_i, \dots, a_{N-1}\}$, where i denotes the index of one candidate, and N is the total number of answer candidates. The notation used in this paper is summarized in Table 1.

4. Method

Each QA pair is analyzed using two subgraphs constructed from *ConceptNet* in order to better understand the knowledge graph facts relevant to the question and candidates: a local subgraph which allows the probability of a single candidate being correct to be determined; and a global subgraph which facilitates joint reasoning among all candidates and guides the adjustment of the probability of the final prediction. By combining the pre-trained LM and the gate mechanism, we are able to take advantage of the information provided by large-scale unstructured texts and specific structured multi-chain reasoning.

4.1. Constructing Subgraphs

We construct local subgraph G_i and global subgraph G^c based on the question q , candidates $A = \{a_0, \dots, a_i, \dots, a_{N-1}\}$, and the external knowledge graph *ConceptNet* (Speer et al., 2017). In light of the fact that *ConceptNet* includes multilingual entities, it is essential to extract the English knowledge graph G_{en} from *ConceptNet* and generate a vocabulary match pattern consisting of lemma for each word. Afterwards, we match the topic entities in q and each a_i respectively using lemmas in the match pattern to generate question-node set V^q and answer-node set V_i^a . For example, in Fig. 1, choice C “Singing” has been matched with words “sing” and “singing” from *ConceptNet* since they derive from the same lemma “sing”.

We then search the English knowledge graph G_{en} with V^q and V_i^a to extract all two-hop neighbor nodes as extra-node set V_i^{ex} . Meanwhile, the relationships between the node pairs in $V^q \cup V_i^a \cup V_i^{ex}$ from the English knowledge graph G_{en} are extracted together to generate the edge set E_i . Finally, a local subgraph G_i is constructed by using $V^q \cup V_i^a \cup V_i^{ex}$ and E_i (Lin et al., 2019), including the topic entities found in the question q and candidate answer a_i . Similarly, the global subgraph G^c is constructed with $V^q \cup V^A \cup V^{ex}$ and E , which includes the topic entities extracted from q and A .

4.2. Reasoning over Subgraphs

A description of our reasoning on subgraphs is presented in Fig. 2. In order to obtain representation \mathbf{lm}_i^{qa} for each QA pair $[q; a_i]$, we use a pre-trained language model such as BERT or RoBERTa. Meanwhile, we construct

Table 1

Here are notations and their definitions used in this paper.

Notations	Definitions
α_{xy}	The attention weight for \mathbf{m}_{xy}
a_i	The i-th answer candidate text of the given question q
A	The text set of all answer candidates of q
D	The dimension of embeddings or hidden states in graphs
E_i	The edge set of nodes in $V^q \cup V_i^a \cup V_i^{ex}$
E	The edge set of nodes in $V^q \cup V^A \cup V^{ex}$
G^c	The global subgraph constructed from G^{en} using nodes in V^q, V^A, V^{ex} and edges in E
G^{en}	The knowledge subgraph including all English entities in the larger knowledge graph <i>ConceptNet</i>
G_i	The local subgraph constructed from G^{en} using nodes in V^q, V_i^a, V_i^{ex} and edges in E_i
\mathbf{g}^c	The graph representation of global subgraph G^c
\mathbf{g}_i	The graph representation of local subgraph G_i
\mathbf{g}'_i	The integrated graph representation of \mathbf{g}_i and \mathbf{g}^c
\mathbf{g}_i^{qa}	The node graph representation of the QA pair $[q; a_i]$ in the last layer
\mathbf{h}_x^l	The hidden state of node x in l-th layer
\mathbf{h}^T	The transpose of a vector \mathbf{h}
L	The number of layers in the QA model
\mathbf{lm}_i^{qa}	The output embedding of text $[q; a_i]$ from a pre-trained language model
\mathbf{m}_{xy}	The message passing from node x to node y
\mathbf{M}_y^l	The embedding of the aggregate message from neighbor nodes of y in the l-th layer
N	The number of answer candidates for q
$P(a_i q)$	The predicted probability of the answer candidate a_i being the correct answer of q
q	The given question text
$[q; a_i]$	The concatenation of texts q and a_i , called QA pair
$[q; a_0; \dots; a_{N-1}]$	The concatenation of texts q and A , called global context
\mathbf{r}_{xy}	The edge (relation) type embedding between two nodes
\mathbf{u}_x	The node type embedding of node x
V_i^a	The node set consisting of all entities in a_i
V^A	The node set consisting of all entities nodes in A
V_i^{ex}	The node set consisting of all two-hop neighbors between nodes in V^q and V_i^a
V^{ex}	The node set consisting of all two-hop neighbors between nodes in V^q and V^A
V^q	The node set consisting of all entities in q
V_{y_nbhd}	The nodes set of all neighbor of node y
\oplus	The concatenation of embeddings

multiple subgraphs from the external knowledge graph: a local subgraph G_i for each QA pair $[q; a_i]$, and a shared global subgraph G^c for the global context $[q; a_0; \dots; a_{N-1}]$.

We combine G^c with each G_i respectively to establish multiple chains for better reasoning. To achieve jointly reasoning over LM and subgraphs, similar with QA-GNN (Yasunaga et al., 2021), \mathbf{lm}_i^{qa} is inserted as a new node (QA pair node) to G_i and G^c respectively. Further, we integrate the representations of the global subgraph G^c and each local subgraph G_i using the gate mechanism to control information passing between the subgraphs. Finally, the probability score of each candidate is obtained through MLP networks.

We share a common 5-layer GNN model (Yasunaga et al., 2021) for subgraphs G_i and G^c , and we set different maximum node numbers for updating the state of the nodes and passing messages between the nodes. For updating the state of node y in layer (l), we firstly aggregate the message of its neighbor nodes V_{y_nbhd} with the self-attention

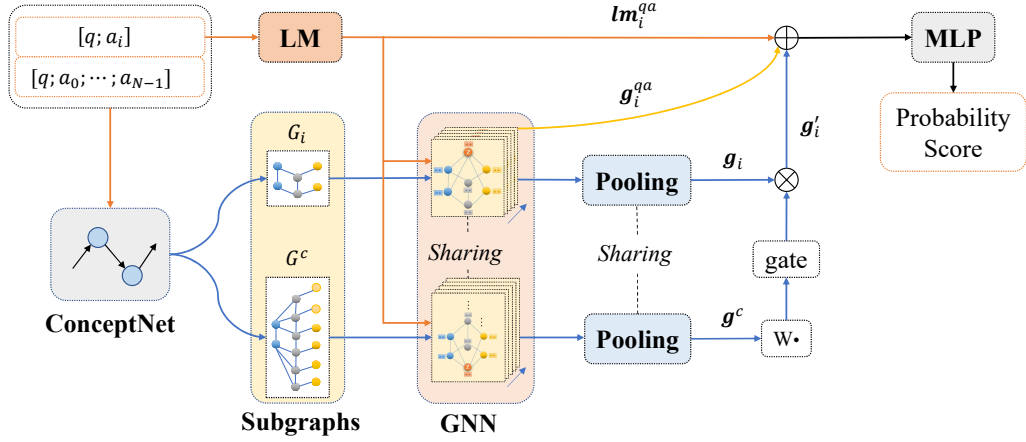


Fig. 2. An overview of the MKSQA model. The knowledge subgraphs are extracted from *ConceptNet* based on QA pair $[q; a_i]$ and global context $[q; a_0; \dots; a_{N-1}]$, and the QA pair is encoded by LM as a new node for joint reasoning with subgraphs. As a next step, a shared GNN model is used to pass messages, update node states, and obtain graph representations in the pooling layer. Following the application of a gate to graph representations for the fusion of information between subgraphs, we combine the information from the QA pair with the subgraphs to predict the outcome.

mechanism (Vaswani et al., 2017):

$$\mathbf{M}_y^{(l-1)} = f_n \left(\sum_{x \in V_{y_nbhd}} \alpha_{xy} \mathbf{m}_{xy} \right), \quad (1)$$

Here, f_n is a 2-layer MLP: $R^D \rightarrow R^D$, α_{xy} is an attention weight for $\mathbf{m}_{xy} \in R^D$ where \mathbf{m}_{xy} represents the message passing from *node* x to *node* y . Specifically, \mathbf{m}_{xy} is conducted with three components:

$$\mathbf{m}_{xy} = f_m(\mathbf{h}_x \oplus \mathbf{u}_x \oplus \mathbf{r}_{xy}), \quad (2)$$

where $\mathbf{h}_x \in R^D$ is the hidden state embedding of *node* x , $\mathbf{u}_x \in R^{D/2}$ is the node type embedding and $\mathbf{r}_{xy} \in R^D$ is the edge (relation) type embedding. A total of four types of nodes have been created: question nodes, answer nodes, extra nodes, and QA pair nodes, as well as 34 types of edge relationships have been collected from G_{en} and have been listed in MHGRN (Feng et al., 2020). $f_m : R^{2.5D} \rightarrow R^D$ is a linear transformation, and \oplus denotes concatenation operation. With the aggregated message embedding $\mathbf{M}_y^{(l-1)} \in R^D$, we can update node embedding $\mathbf{h}_y^{(l)} \in R^D$:

$$\mathbf{h}_y^{(l)} = \mathbf{M}_y^{(l-1)} + \mathbf{h}_y^{(l-1)}, \quad (3)$$

where $\mathbf{h}_y^{(l-1)} \in R^D$ is the embedding of *node* y obtained by the preceding layer.

Further we apply a gate function for passing the information between subgraphs. For subgraphs G_i and G^c , we have the integrated graph representation \mathbf{g}'_i :

$$\mathbf{g}'_i = gate(W \cdot \mathbf{g}^c + b) \cdot \mathbf{g}_i, \quad (4)$$

where we use *sigmoid* as $gate(\cdot)$ in our experiments, W and b are learnable parameters. $\mathbf{g}_i \in R^D$ and $\mathbf{g}^c \in R^D$ are the graph representations of G_i and G^c respectively after pooling, actually an attention mechanism, defined as:

$$\mathbf{g}_i = pooling \{ \mathbf{h}_v^L | v \in V_i \}, \quad (5)$$

$$\mathbf{g}^c = pooling \{ \mathbf{h}_v^L | v \in V^c \}, \quad (6)$$

$$pooling = \sum_{v \in V} softmax \left(\frac{(\mathbf{h}_v^L)^T \cdot \mathbf{lm}_i^{qa}}{\sqrt{D}} \right), \quad (7)$$

4.3. Prediction & Learning

To figure out the probability of each candidate being the right answer, a one-layer MLP $f(\cdot)$ is followed:

$$P(a_i|q) \propto f(\mathbf{g}'_i \oplus \mathbf{g}_i^{qa} \oplus \mathbf{lm}_i^{qa}). \quad (8)$$

where \mathbf{lm}_i^{qa} is the representation of QA pair $[q; a_i]$, $\mathbf{g}_i^{qa} = \mathbf{h}_{qa}^L$ is the state of QA pair node in the last layer L.

Finally, cross-entropy loss is used to optimize our MKSQA model. A complete explanation of our quality assurance model can be found in Algorithm 1.

Algorithm 1 MKSQA Algorithm

Input: global context $[q; a_0; \dots; a_{N-1}]$, large knowledge graph *ConceptNet*

Output: probability P

- 1: From *ConceptNet*, extract knowledge graph G_{en} to generate match patterns and topic entities
 - 2: Use QA pair and global context $[q; a_i]$ and $[q; a_0; \dots; a_{N-1}]$ to construct local subgraph G_i and global subgraph G^c respectively following Section 4.1
 - 3: Encode each QA pair $[q; a_i]$ with a pretrained model (i.e., RoBERTa-Large) to obtain \mathbf{lm}_i^{qa}
 - 4: **for** $i = 0$ to $N - 1$ **do**
 - 5: Apply \mathbf{lm}_i^{qa} as a node to subgraphs G_i and G^c respectively
 - 6: **for** $l = 1$ to L **do**
 - 7: Update node hidden states and pass messages between nodes in local subgraph G_i based on Eq. (3)
 - 8: **end for**
 - 9: Output $H_i^L = \{h_v^L, v \in V_i\}$, the hidden states of nodes in the final layer L
 - 10: **for** $l = 1$ to L **do**
 - 11: Update node hidden states and pass messages between nodes in global subgraph G^c based on Eq. (3)
 - 12: **end for**
 - 13: Output $H_c^L = \{h_v^L, v \in V^c\}$, the hidden states of nodes in the final layer L
 - 14: Pooling H_i^L and H_c^L to obtain the subgraph representations \mathbf{g}_i and \mathbf{g}^c based on Eqs. (5), (6), and (7)
 - 15: Integrate subgraph representations \mathbf{g}_i and \mathbf{g}^c to gain the final graph representation \mathbf{g}' based on Eq. (4)
 - 16: Compute the probability p_i based on Eq. (8)
 - 17: **end for**
 - 18: **Return** $P = \{p_i\}, i = [0, 1, \dots, N - 1]$
-

4.4. Effectiveness Analysis of Global Subgraph

In this section, we explore and analyze how the global subgraph can enhance answer prediction.

As shown in Fig. 3, our method constructs three kinds of chains for answer reasoning: LM, local subgraphs KG_1 - KG_5 , and global subgraph KG_6 . Traditionally, graph reasoning for each question-answer (QA) pair $[q; a_i]$ is conducted over its local subgraph KG_i , specifically seen in the middle box of Fig. 3. In this way, there exist no edges among nodes of different answer candidates (yellow balls with different brightness in KG_1 - KG_5). Thereby, no information has been passed among these candidate nodes and our newly constructed global subgraph KG_6 facilitates this. Subgraph KG_6 includes the entity nodes of the question and of all candidates. In addition, we add edges for these choice nodes (dashed double arrows in KG_6) in order to facilitate the flow of information between them. We use the knowledge captured from KG_6 to supplement the knowledge learned from KG_1 - KG_5 .

Specifically, in the example given in Fig. 3, we can see candidates “ a_1 . take time” and “ a_5 . make haste” are both related to “time”, but the two candidates emphasize two almost opposing views respectively. It should be noted that candidate “ a_1 ” emphasizes spending time while candidate “ a_5 ” emphasizes reducing time. So the gap of probability values between the two candidates become larger after integrating the information in KG_6 . Consequently, the probability of candidate “ a_1 ” is greater than that of candidate “ a_5 ”, since “take time” seems more reasonable than “make haste” in order to achieve harmony. Among the candidates, “ a_2 ”, “ a_3 ”, and “ a_4 ” are more related to the question than the other two candidates. Furthermore, “ a_3 ” and “ a_4 ” have more influence on “harmony” than “ a_2 ”. This means that the candidate “ a_2 ” has the lowest probability of being selected. Finally, candidates “ a_3 ” and “ a_4 ” are the most relevant to the question, and the probability value for candidate “ a_3 ” is only slightly higher than that of candidate “ a_4 ”.

Although the contrasting information among choices facilitates the inference of the correct answer, it cannot be captured only with the local subgraphs KG_1 - KG_5 . Our global subgraph KG_6 can capture this information well by passing messages among all candidate nodes.

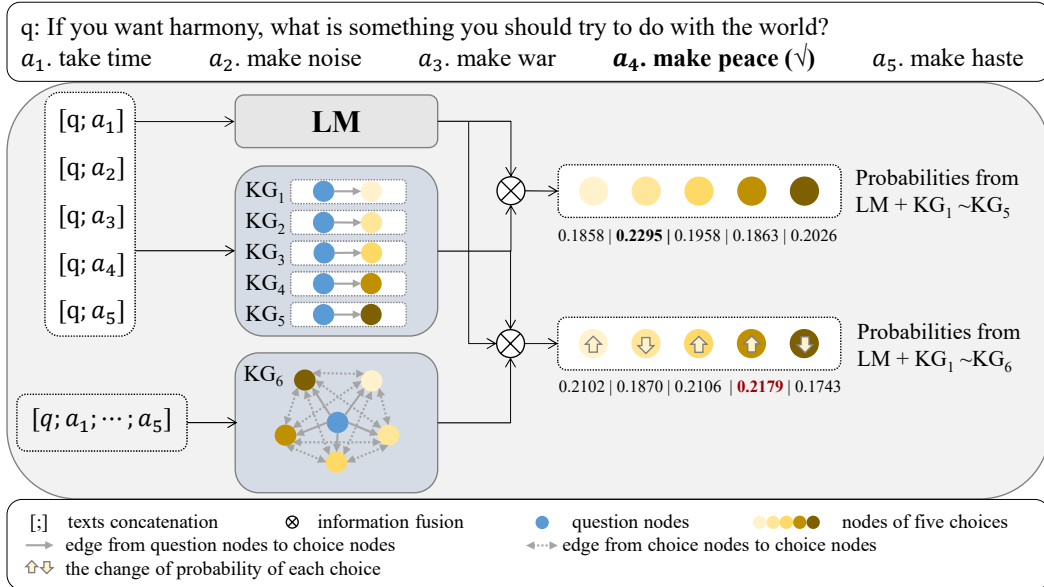


Fig. 3. The contribution of multiple subgraphs for commonsense question answering. In the subgraphs KG_1 - KG_6 , we do not show the intermediate nodes between the question nodes and the answer candidate/choice nodes. The two dashed boxes on the right of the above figure depict the prediction results for each candidate before and after fusing the information in KG_6 , respectively.

4.5. Connection to Existing Related Methods

Connection to LM+KG based Methods. Joint reasoning with LM and KG has been widely implemented in Multi-choice QA research and has gained prominent performance, such as QA-GNN (Yasunaga et al., 2021), JointLK (Sun et al., 2022), GeaseLM (Zhang et al., 2022c), DRAGON (Yasunaga et al., 2022), et al. QA-GNN (Yasunaga et al., 2021) considers the embedding of question-answer (QA) pair learnt from PLM as a new node, and utilizes it to achieve information interaction and joint reasoning between LM and KG. JointLK (Sun et al., 2022) and GeaseLM (Zhang et al., 2022c) implement the information bidirectional interaction between LM and KG in each layer of them. DRAGON (Yasunaga et al., 2022) pretrains a deeply joint language-knowledge foundation model from text and KG at scale. Our paper is most related to QA-GNN. However, QA-GNN mainly focuses on the problem of the mutually updating of the LMs and GNNs and tries to unify the representations of these two different source models. Differently, we focus on the strategy of knowledge subgraph construction and multiple chains of answer reasoning.

Connection to LM based Methods. To solve multiple choice question answering, some methods only rely on LMs to get external knowledge. For example, Muppet (Aghajanyan et al., 2021) pre-finetunes PLM on around 50 datasets, over 4.8 million total labeled examples between language model pre-training and fine-tuning so that the model can learn more knowledge from many different tasks. To generate additional clues from the given context of a question, GenMC ((Huang et al., 2022) employs an encoder-decoder generator. Subsequently, the question together with the generated clues are fed into the shared encoder with the generator to predict the answer. Thus, there is a difference between our method and these types of approaches. In our experiments, we apply both LM and KG simultaneously, and we do not make comparisons between them due to the differences in their settings.

Other External Knowledge Utilization. Other sources of external information and sources can also be used to improve the question answering systems, such as Wiktionary2 as well as more labeled datasets for question answers

Besides the pretrained LMs and KGs, some other external information and sources can also be used to improve the question answering systems, such as Wiktionary² and more labeled question-answer datasets (Khashabi et al., 2020; Aghajanyan et al., 2021). For example, DEKCOR (Xu et al., 2021d) extracts descriptions of question-answer mentioned concepts from Wiktionary and encodes them together with the question, answer candidate and corresponding concepts using ALBERT (Lan et al., 2019). KEAR (Xu et al., 2021c) introduces Wiktionary and labeled training data (CommonsenseQA and 16 related QA datasets) to strengthen the capability of model’s answer prediction. UnifiedQA (Khashabi et al., 2020) unifies the format of model input for diverse types question answering tasks (including MCQA tasks) into a single one, to enable the model to learn the knowledge in all tasks. Our method is different with these methods and we do not utilize any other external data except the LM and KG.

5. Experiments

We conduct experiments to validate the performance of our model MKSQA. The experiments are performed on a device equipped with Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz and two GeForce RTX 2080Ti GPUs. All the algorithms are implemented using Pytorch and trained with an RAdam optimizer (Liu et al., 2019a). In order to tune the hyperparameters, development sets are used, and testing results are reported based on the best epoch of development set. Additionally, we set the dimension (D=200) and the number of layers (L=5). We set the dimension (D=200) and number of layers (L=5) of our module. To prevent model overfitting, dropout is applied to each layer with dropout rate 0.2. The parameters of the model are optimized by RAdam (Liu et al., 2019a), with batch size 128, gradient clipping as 1.0, and learning rate as 1e-5, 1e-3 for the LM and GNN components respectively.

5.1. Datasets

We evaluate our model on two common datasets of multi-choice question answering: CommonsenseQA (CSQA) (Talmor et al., 2019) and OpenBookQA (OBQA) (Mihaylov et al., 2018). For CSQA set, given that the official test dataset can’t be accessed publicly, we use the IH data split (Lin et al., 2019) to perform the experiments. In the IH data set, the development set is equal to the official one with 1221 question samples. A total of 1241 samples have been selected from the official training set by Lin et al. (2019) for use as the IH test set, and the remaining samples are used as the IH training set. In the case of OBQA, we use official data. Table 2 gives the experiment data scale and split of the two datasets.

Table 2

The scale and split of experimental data. In our experiments, we use the IH data splits of CSQA (since the official test set of CSQA is not publicly available) and the official data set (OF) of OBQA.

Dataset	Source	Train Set	Dev Set	Test Set	Number of Answer Candidates
CSQA	OF	9741	1221	1140	5
CSQA	IH	8500	1221	1241	5
OBQA	OF	4957	500	500	4

5.2. Baselines

We compare our method with several baselines, including fine-tuned LM and LM+KG models. For fine-tuned LM models, we use RoBERTa-Large (Liu et al., 2019b) which obtains better performance than BERT-Base LM and BERT-Large LM (Feng et al., 2020). For LM+KG models, we compare our method with KagNet (Lin et al., 2019), MHGRN (Feng et al., 2020), QA-GNN (Yasunaga et al., 2021), JointLK (Sun et al., 2022), GSC (Wang et al., 2021), and GreaseLM (Zhang et al., 2022c) respectively.

Roberta-large: Roberta-large is a common baseline for multi-choice question answering tasks, which is directly fine-tuned on the downstream tasks. It does not rely on any extra knowledge, such as knowledge graph (KG) or generated facts (Liu et al., 2019b).

²<https://www.wiktionary.org>

RN: Relation network (RN) is designed to capture the core common properties for relation reasoning (Santoro et al., 2017).

KagNet: KagNet combines the GCN and LSTM modules. It reasons for answers based on reasoning path on the KG.

MHGRN: MHGRN utilizes KGs to achieve multi-hop multi-relation reasoning for multi-choice question answering tasks (Feng et al., 2020).

QA-GNN: QA-GNN conducts a joint reasoning over the LM and KG in a form of pipeline, by integrating the QA pair information into knowledge graphs as a new graph node (Yasunaga et al., 2021).

JointLK: JointLK performs a similar joint reasoning over the LM and KG, yet through a dense bidirectional attention module (Sun et al., 2022).

GSC: GSC treats the KG as a simple edge counter and experimental results demonstrate that it performs well (Wang et al., 2021).

GreaseLM: GreaseLM fuses effectively the encoded representations from the LM and KG by passing them to a multi-layer module for deep interaction operations (Zhang et al., 2022c).

It is important to note that the above methods only utilize one knowledge graph for reasoning for each QA pair independently, in contrast with our method. They ignore the cross information among answer candidates.

Table 3

The test accuracy comparison with main baselines on the IH data split of CommonsenseQA dataset.

Methods	IHtest-acc.(%)
RoBERTa-Large (Liu et al., 2019b)	68.69 (± 0.56)
+KagNet (Lin et al., 2019)	69.01 (± 0.76)
+MHGRN (Feng et al., 2020)	71.11 (± 0.81)
+QA-GNN (Yasunaga et al., 2021)	73.41 (± 0.92)
+GreaseLM (Zhang et al., 2022c)	74.20 (± 0.40)
+JointLK (Sun et al., 2022)	74.43 (± 0.83)
+GSC (Wang et al., 2021)	74.48 (± 0.41)
+MKSQA (Ours)	74.53 (± 0.52)

Table 4

The test accuracy comparison with main baselines on the official data of OpenBookQA dataset.

Methods	OFtest-acc.(%)
RoBERTa-Large (Liu et al., 2019b)	64.80 (± 2.37)
+RN (Santoro et al., 2017)	65.20 (± 1.18)
+MHGRN (Feng et al., 2020)	66.85 (± 1.19)
+GSC (Wang et al., 2021)	70.33 (± 0.81)
+JointLK (Sun et al., 2022)	70.34 (± 0.75)
+QA-GNN (Yasunaga et al., 2021)	70.58 (± 1.42)
+MKSQA (Ours)	71.80 (± 0.51)

5.3. Main Results

Tables 3 and 4 illustrate the main comparison results with our baselines. We can see our method outperforms all baselines on both CommonsenseQA and OpenBookQA datasets. The improvement over QA-GNN shows that using multiple knowledge subgraphs can benefit reasoning for question answering. Meanwhile, the standard deviations are narrower than baselines' on both datasets, which means the performance of our model is more stable³.

³Although the AristoRoberta model has been proved performing better in multi-choice question answering tasks (Xu et al., 2021a; Yan et al., 2021; Xu et al., 2021b). We do not evaluate our model based on it due to it is not publicly available.

We make clear that there are methods on the official leaderboards perform better than our method, but the experiment settings and external knowledge used are not totally same. For example, much larger pre-trained language models (such as T5-11B) are employed (Khashabi et al., 2020; Huang et al., 2022), or ensemble systems are designed (Lan et al., 2019; Xu et al., 2021d), or more extra facts are used to augment reasoning in addition to knowledge graphs (Xu et al., 2021c,d), or generator and retrieval models are used to generate additional clues to enhance answer reasoning (GenMC (Huang et al., 2022)). More details can be found in Section 4.5. The SOTA methods on the leaderboards: CPACE⁴ on CommonsenseQA and X-Reasoner⁵ on OpenBookQA are undocumented. Therefore, we did not compare our model with these two methods.

5.4. Ablation Studies

To better analyze the performance of our model, we conduct a series of ablation studies in terms of the effectiveness of the global subgraph, gate design, and the size of subgraphs G_i and G^c , respectively.

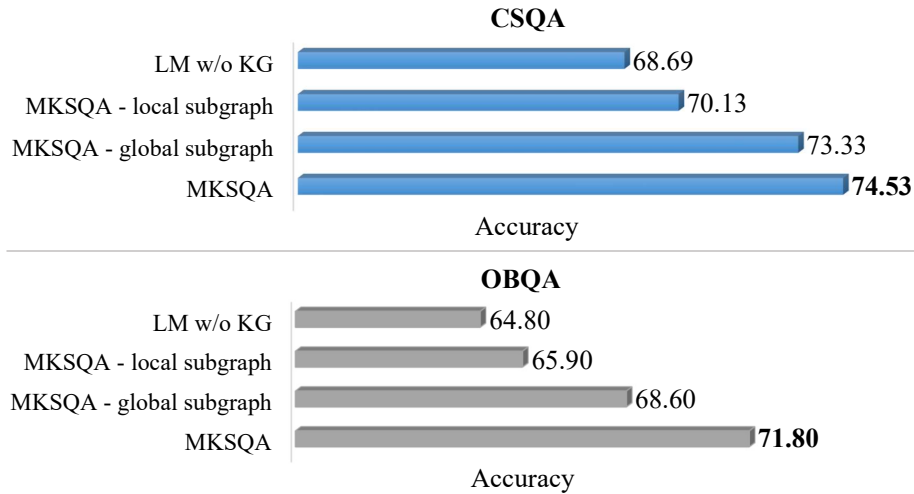


Fig. 4. The effectiveness of local and global subgraphs on CSQA and OBQA. Method LM only uses a single pre-trained LM (RoBERTa-Large) to predict the answer without an external knowledge graph. We set the parameter `max_node_num` as 400 for subgraph G_i and 500 for global subgraph G^c .

Effectiveness of the global subgraph. To unambiguously study how the global subgraph G^c acts on the performance, we take away the local and global subgraphs from our method respectively to observe the impact of the remaining parts on the answer prediction. As shown in Fig. 4, the results indicate that both subgraphs contribute to the answer prediction. It is evident from this that G^c is capable of providing extra information and views on the questions and candidates, and contributing to the reasoning process. Constructing global subgraph G^c normally relates to more nodes compared with subgraph G_i . The limitation of global subgraph size controls the information extracted from the external KG. Later on, a further analysis is conducted to analyze the influence of size limitation of each subgraph.

Effectiveness of the gate mechanism. We carry out a series of exploratory experiments to evaluate different gate designs for processing subgraph information. Specifically, we design six groups of experiments with different input data as shown in Table 5. In Table 5, the result shows our model brings better performance when processing subgraph information. It also shows there exist gaps between the model QA-GNN+ G^c and our model MKSQA on both datasets, which demonstrates that the simple combination of subgraphs G_i and G^c may not bring better performance.

Analysis for subgraph size. Constructing global subgraph G^c normally relates to more nodes compared with subgraph G_i . The limitation of global subgraph size controls the information extracted from the external KG. Fig. 5 shows the results of MKSQA under different size limitation of global subgraph, while the size of subgraph G_i is fixed at no more than 400 nodes. We can see a bigger global subgraph size brings benefits to the model at its early phase, since

⁴CPACE: <https://www.tau-nlp.sites.tau.ac.il/csqa-leaderboard>.

⁵X-Reasoner: https://leaderboard.allenai.org/open_book_qa/submission/cbf0f581jc49vlquujg.

Table 5

The test accuracy for different gate mechanisms on datasets CSQA and OBQA with the sizes of subgraph G_i and G^c as (400,500) respectively. Here, $(\mathbf{g}_i \oplus \mathbf{g}_i^{qa} \oplus \mathbf{lm}_i^{qa})$ equals to QA-GNN method; $(\mathbf{g}_i \oplus \mathbf{g}^c \oplus \mathbf{g}_i^{qa} \oplus \mathbf{lm}_i^{qa})$ equals to (QA-GNN+ G^c) method, and $(gate(\mathbf{g}^c) \cdot \mathbf{g}_i \oplus \mathbf{g}_i^{qa} \oplus \mathbf{lm}_i^{qa})$ equals to our model MKSQA.

Methods	CSQA	OBQA
$\mathbf{g}_i \oplus \mathbf{g}_i^{qa} \oplus \mathbf{lm}_i^{qa}$	73.33	68.60
$\mathbf{g}_i \oplus \mathbf{g}^c \oplus \mathbf{g}_i^{qa} \oplus \mathbf{lm}_i^{qa}$	73.81	69.20
$gate(\mathbf{lm}_i^{qa}) \cdot \mathbf{g}_i \oplus \mathbf{g}_i^{qa} \oplus \mathbf{lm}_i^{qa}$	73.41	68.00
$gate(\mathbf{g}_i^{qa}) \cdot \mathbf{g}_i \oplus \mathbf{g}_i^{qa} \oplus \mathbf{lm}_i^{qa}$	72.06	70.80
$gate(\mathbf{g}) \cdot \mathbf{g}_i \oplus \mathbf{g}_i^{qa} \oplus \mathbf{lm}_i^{qa}$	72.76	68.40
$gate(\mathbf{g}^c) \cdot \mathbf{g}_i \oplus \mathbf{g}_i^{qa} \oplus \mathbf{lm}_i^{qa}$	74.53	71.80

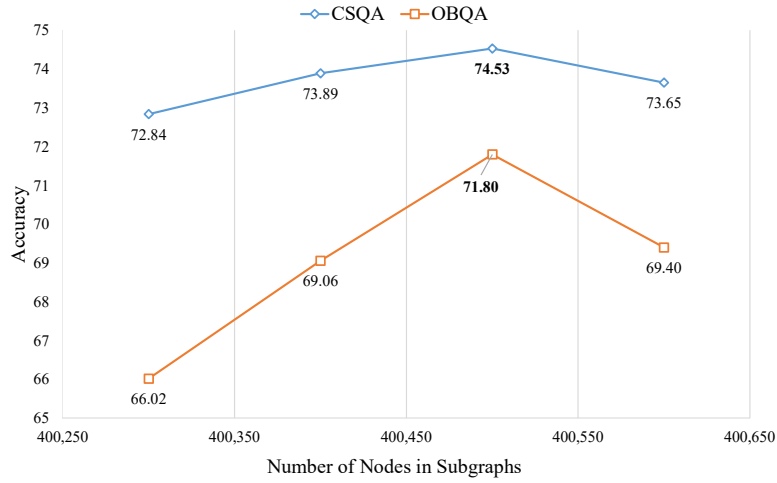


Fig. 5. The results of MKSQA under different size limitation of global subgraph, while the size limitation of subgraph G_i is fixed at 400 nodes.

more information is introduced to the model. As the size of the global subgraph continues to grow, the performance declines due to the irrelevant nodes involved.

Based on the results of this exploration, we limit the size of the subgraph G_i and global subgraph G^c (i.e., 400, 500). To ensure a fair comparison, we also fine tuned the size limitation of the subgraph G_i in the method QA-GNN. An illustration of the results can be found in Fig. 6. The original setting in QA-GNN is 200, and we can see the performance is better when the maximum node number increased to 300 on CSQA dataset. However, it is still worse than our method (i.e., QA-GNN: 73.65% VS MKSQA: 74.53%).

Table 6

The ablation experiments about the number of answer candidates given a question on CSQA dataset. The numbers in round brackets indicate the degree of increasing in terms of accuracy when the number of candidates is reduced by one.

Number of answer candidates	IHtest-acc.(%)
5	74.53
4	76.15 (+1.62)
3	80.26 (+4.11)
2	88.24 (+7.98)

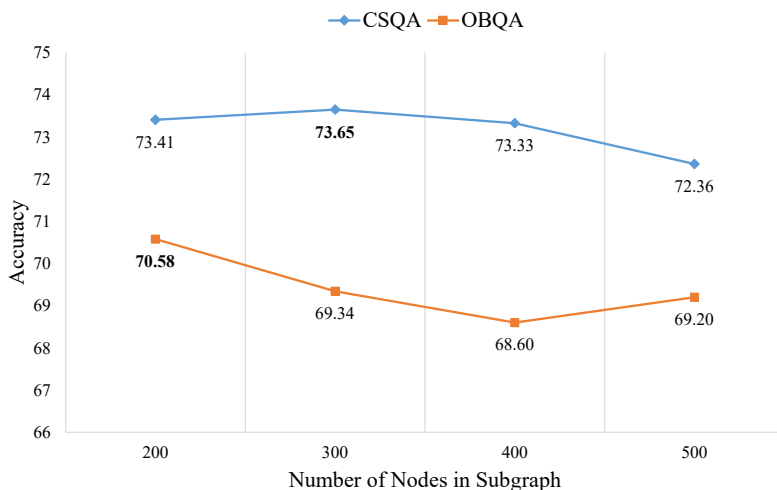


Fig. 6. The results of QA-GNN under different size limitation of subgraph G_i .

5.5. Study on Size of Candidate Set

In this section, we further explore and study the influence of the size of answer candidate set. We conduct the experiments on the number of answer candidates based on CSQA dataset and the results are shown in Table 6. In each experiment, we first ensure the right answer is included, then randomly choose the required number of distractors from the original candidate set to meet the requirements for the total number of answer candidates. The selected candidates and the correct answer compose a new candidate set for the question. Based on the new set, we reconstructed our global and local subgraphs and train MKSQA model. We can see MKSQA makes a big difference for the same questions but with different size of candidates. It has been observed that the smaller the number of candidates, the higher the prediction accuracy. There is a possibility that this is due to the fact that fewer candidates can make the elimination strategy more effective.

6. Conclusion

In this paper, we propose a novel method MKSQA based on LM+KG, which is intended to perform a joint reasoning among all candidates to answer multi-choice questions. We retrieve multiple subgraphs from an external knowledge graph to establish multiple chains for reasoning, so that we can make full use of KGs and potential knowledge in question and candidates. A shared GNN model is trained to learn the representations of the subgraphs and the QA pair, and a gate mechanism is applied to promote information fusion between subgraphs. To the best of our knowledge, MKSQA is the first work to consider multiple knowledge subgraphs, especially the global subgraph for Multiple Choice Question Answering, and to take both local and global candidate information into account when reasoning for each question-answer pair. In addition, our method achieved an exact match score of 74.53% on CSQA and 71.80% on OBQA, which represents an outstanding performance. We believe our simple method could be served as a baseline for the proposed task, and invoke better models in the future works.

Acknowledgment

This research was supported in part by NSFC under Grant no. 62206179 and no. 92270122, in part by Guangdong Provincial Natural Science Foundation under grant no. 2022A1515010129 and in part by Shenzhen Research Foundation for Basic Research, China, under Grant no. JCYJ20210324093000002.

References

Aghajanyan, A., Gupta, A., Shrivastava, A., Chen, X., Zettlemoyer, L., Gupta, S., 2021. Muppet: Massive multi-task representations with pre-finetuning, in: Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online

- and Punta Cana, Dominican Republic. pp. 5799–5811. URL: <https://aclanthology.org/2021.emnlp-main.468>, doi:10.18653/v1/2021.emnlp-main.468.
- Atzeni, M., Bogojeska, J., Loukas, A., 2021. Sqaer: Scaling question answering by decoupling multi-hop and logical reasoning. *Advances in Neural Information Processing Systems* 34, 12587–12599.
- Bollacker, K.D., Evans, C., Paritosh, P.K., Sturge, T., Taylor, J., 2008. Freebase: a collaboratively created graph database for structuring human knowledge, in: SIGMOD.
- Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., Choi, Y., 2019a. COMET: Commonsense transformers for automatic knowledge graph construction, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy*. pp. 4762–4779. URL: <https://aclanthology.org/P19-1470>, doi:10.18653/v1/P19-1470.
- Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Çelikyilmaz, A., Choi, Y., 2019b. Comet: Commonsense transformers for automatic knowledge graph construction. *ArXiv abs/1906.05317*.
- Cao, X., Liu, Y., 2022. Relmkg: reasoning with pre-trained language models and knowledge graphs for complex question answering. *Applied Intelligence*, 1–15.
- Cao, Y., Fang, M., Tao, D., 2019. BAG: Bi-directional attention entity graph convolutional network for multi-hop reasoning question answering, in: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota. pp. 357–362. URL: <https://aclanthology.org/N19-1032>, doi:10.18653/v1/N19-1032.
- Chaturvedi, A., Pandit, O., Garain, U., 2018. CNN for text-based multiple choice question answering, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Melbourne, Australia. pp. 272–277. URL: <https://aclanthology.org/P18-2044>, doi:10.18653/v1/P18-2044.
- Deng, X., Su, Y., Lees, A., Wu, Y., Yu, C., Sun, H., 2021. ReasonBERT: Pre-trained to reason with distant supervision, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. pp. 6112–6127. URL: <https://aclanthology.org/2021.emnlp-main.494>, doi:10.18653/v1/2021.emnlp-main.494.
- Emami, A., De La Cruz, N., Trischler, A., Suleman, K., Cheung, J.C.K., 2018. A knowledge hunting framework for common sense reasoning, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium. pp. 1949–1958. URL: <https://aclanthology.org/D18-1220>, doi:10.18653/v1/D18-1220.
- Feng, Y., Chen, X., Lin, B.Y., Wang, P., Yan, J., Ren, X., 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online. pp. 1295–1309. URL: <https://aclanthology.org/2020.emnlp-main.99>, doi:10.18653/v1/2020.emnlp-main.99.
- Hu, L., Zou, D., Guo, X., Qi, L., Tang, Y., Song, H., Yuan, J., 2021. Four-way bidirectional attention for multiple-choice reading comprehension, in: *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 238–243. doi:10.1109/SMC52423.2021.9658632.
- Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., Leskovec, J., 2020. Open graph benchmark: Datasets for machine learning on graphs, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), *Advances in Neural Information Processing Systems*, Curran Associates, Inc.. pp. 22118–22133. URL: <https://proceedings.neurips.cc/paper/2020/file/fb60d411a5c5b72b2e7d3527cfc84fd0-Paper.pdf>.
- Huang, Y., Fang, M., Cao, Y., Wang, L., Liang, X., 2021. DAGN: Discourse-aware graph network for logical reasoning, in: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Online. pp. 5848–5855. URL: <https://aclanthology.org/2021.naacl-main.467>, doi:10.18653/v1/2021.naacl-main.467.
- Huang, Z., Wu, A., Zhou, J., Gu, Y., Zhao, Y., Cheng, G., 2022. Clues before answers: Generation-enhanced multiple-choice QA, in: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Seattle, United States. pp. 3272–3287. URL: <https://aclanthology.org/2022.naacl-main.239>, doi:10.18653/v1/2022.naacl-main.239.
- Kassner, N., Schütze, H., 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly, in: *Association for Computational Linguistics*, Association for Computational Linguistics, Online. pp. 7811–7818. URL: <https://aclanthology.org/2020.acl-main.698>, doi:10.18653/v1/2020.acl-main.698.
- Khashabi, D., Min, S., Khot, T., Sabharwal, A., Taffjord, O., Clark, P., Hajishirzi, H., 2020. UNIFIEDQA: Crossing format boundaries with a single QA system, in: *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Online. pp. 1896–1907. URL: <https://aclanthology.org/2020.findings-emnlp.171>, doi:10.18653/v1/2020.findings-emnlp.171.
- Kipf, T., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. *ArXiv abs/1609.02907*.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., 2019. Albert: A lite bert for self-supervised learning of language representations. *ArXiv abs/1909.11942*.
- Lin, B.Y., Chen, X., Chen, J., Ren, X., 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China. pp. 2829–2839. URL: <https://aclanthology.org/D19-1282>, doi:10.18653/v1/D19-1282.
- Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J., 2019a. On the variance of the adaptive learning rate and beyond. *ArXiv abs/1908.03265*.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019b. Roberta: A robustly optimized bert pretraining approach. *ArXiv abs/1907.11692*.
- Lv, S., Guo, D., Xu, J., Tang, D., Duan, N., Gong, M., Shou, L., Jiang, D., Cao, G., Hu, S., 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 8449–8456. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/6364>, doi:10.1609/aaai.v34i05.6364.
- Mihaylov, T., Clark, P., Khot, T., Sabharwal, A., 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium. pp. 2381–2391. URL: <https://aclanthology.org/D18-1260>, doi:10.18653/v1/D18-1260.

- Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.H., Riedel, S., 2019. Language models as knowledge bases? ArXiv abs/1909.01066.
- Ryu, D., Shareghi, E., Fang, M., Xu, Y., Pan, S., Haf, R., 2022. Fire burns, sword cuts: Commonsense inductive bias for exploration in text-based games, in: Association for Computational Linguistics, Association for Computational Linguistics, Dublin, Ireland. pp. 515–522. URL: <https://aclanthology.org/2022.acl-short.56>, doi:10.18653/v1/2022.acl-short.56.
- Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T., 2017. A simple neural network module for relational reasoning. *Advances in neural information processing systems* 30.
- Seonwoo, Y., Kim, J.H., Ha, J.W., Oh, A., 2020. Context-aware answer extraction in question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 2418–2428. URL: <https://aclanthology.org/2020.emnlp-main.189>, doi:10.18653/v1/2020.emnlp-main.189.
- Speer, R., Chin, J., Havasi, C., 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence* 31. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/11164>, doi:10.1609/aaai.v31i1.11164.
- Sun, H., Dhingra, B., Zaheer, M., Mazaitis, K., Salakhutdinov, R., Cohen, W., 2018. Open domain question answering using early fusion of knowledge bases and text, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium. pp. 4231–4242. URL: <https://aclanthology.org/D18-1455>, doi:10.18653/v1/D18-1455.
- Sun, Y., Shi, Q., Qi, L., Zhang, Y., 2022. JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States. pp. 5049–5060. URL: <https://aclanthology.org/2022.naacl-main.372>, doi:10.18653/v1/2022.naacl-main.372.
- Talmor, A., Herzig, J., Lourie, N., 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), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4149–4158. URL: <https://aclanthology.org/N19-1421>, doi:10.18653/v1/N19-1421.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems* 30.
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y., 2017. Graph attention networks. URL: <https://arxiv.org/abs/1710.10903>, doi:10.48550/ARXIV.1710.10903.
- Wang, H., Ren, H., Leskovec, J., 2020. Entity context and relational paths for knowledge graph completion. ArXiv abs/2002.06757.
- Wang, K., Zhang, Y., Yang, D., Song, L., Qin, T., 2021. Gnn is a counter? revisiting gnn for question answering. ArXiv abs/2110.03192.
- Xie, Z., Zhu, R., Liu, J., Zhou, G., Huang, J.X., 2022. An efficiency relation-specific graph transformation network for knowledge graph representation learning. *Information Processing & Management* 59, 103076. URL: <https://www.sciencedirect.com/science/article/pii/S0306457322001777>, doi:<https://doi.org/10.1016/j.ipm.2022.103076>.
- Xu, W., Deng, Y., Zhang, H., Cai, D., Lam, W., 2021a. Exploiting reasoning chains for multi-hop science question answering. ArXiv abs/2109.02905.
- Xu, W., Zhang, H., Cai, D., Lam, W., 2021b. Dynamic semantic graph construction and reasoning for explainable multi-hop science question answering, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online. pp. 1044–1056. URL: <https://aclanthology.org/2021.findings-acl.90>, doi:10.18653/v1/2021.findings-acl.90.
- Xu, Y., Fang, M., Chen, L., Du, Y., Zhou, J., Zhang, C., 2022. Perceiving the world: Question-guided reinforcement learning for text-based games, in: Association for Computational Linguistics, Association for Computational Linguistics, Dublin, Ireland. pp. 538–560. URL: <https://aclanthology.org/2022.acl-long.41>, doi:10.18653/v1/2022.acl-long.41.
- Xu, Y., Fang, M., Chen, L., Du, Y., Zhou, J.T., Zhang, C., 2020. Deep reinforcement learning with stacked hierarchical attention for text-based games. *Advances in Neural Information Processing Systems* 33, 16495–16507.
- Xu, Y., Zhu, C., Wang, S., Sun, S., Cheng, H., Liu, X., Gao, J., He, P., Zeng, M., Huang, X., 2021c. Human parity on commonsenseqa: Augmenting self-attention with external attention. ArXiv abs/2112.03254.
- Xu, Y., Zhu, C., Xu, R., Liu, Y., Zeng, M., Huang, X., 2021d. Fusing context into knowledge graph for commonsense question answering, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online. pp. 1201–1207. URL: <https://aclanthology.org/2021.findings-acl.102>, doi:10.18653/v1/2021.findings-acl.102.
- Yan, J., Raman, M., Chan, A., Zhang, T., Rossi, R., Zhao, H., Kim, S., Lipka, N., Ren, X., 2021. Learning contextualized knowledge structures for commonsense reasoning, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online. pp. 4038–4051. URL: <https://aclanthology.org/2021.findings-acl.354>, doi:10.18653/v1/2021.findings-acl.354.
- Yasunaga, M., Bosselut, A., Ren, H., Zhang, X., Manning, C.D., Liang, P., Leskovec, J., 2022. Deep bidirectional language-knowledge graph pretraining. ArXiv abs/2210.09338.
- Yasunaga, M., Ren, H., Bosselut, A., Liang, P., Leskovec, J., 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online. pp. 535–546. URL: <https://aclanthology.org/2021.naacl-main.45>, doi:10.18653/v1/2021.naacl-main.45.
- Zhang, Q., Chen, S., Xu, D., Cao, Q., Chen, X., Cohn, T., Fang, M., 2022a. A survey for efficient open domain question answering. arXiv preprint arXiv:2211.07886.
- Zhang, Q., Weng, X., Zhou, G., Zhang, Y., Huang, J.X., 2022b. Arl: An adaptive reinforcement learning framework for complex question answering over knowledge base. *Information Processing & Management* 59, 102933. URL: <https://www.sciencedirect.com/science/article/pii/S0306457322000565>, doi:<https://doi.org/10.1016/j.ipm.2022.102933>.
- Zhang, X., Bosselut, A., Yasunaga, M., Ren, H., Liang, P., Manning, C.D., Leskovec, J., 2022c. Greaselm: Graph reasoning enhanced language models for question answering. ArXiv abs/2201.08860.