Reasoning over Hierarchical Question Decomposition Tree for Explainable Question Answering

Jiajie Zhang[1*], Shulin Cao[2*], Tingjian Zhang[2], Xin Lv[2], Jiaxin Shi[3], Qi Tian[3], Juanzi Li[2†], Lei Hou[2]

[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3]Cloud BU, Huawei Technologies
{jiajie-z19, caosl19}@mails.tsinghua.edu.cn
{houlei, lijuanzi}@tsinghua.edu.cn

Abstract

Explainable question answering (XQA) aims to answer a given question and provide an explanation why the answer is selected. Existing XQA methods focus on reasoning on a single knowledge source, e.g., structured knowledge bases, unstructured corpora, etc. However, integrating information from heterogeneous knowledge sources is essential to answer complex questions. In this paper, we propose to leverage question decomposing for heterogeneous knowledge integration, by breaking down a complex question into simpler ones, and selecting the appropriate knowledge source for each sub-question. To facilitate reasoning, we propose a novel two-stage XQA framework, *Reasoning over Hierarchical Question Decomposition Tree* (RoHT). First, we build the *Hierarchical Question Decomposition Tree* (HQDT) to understand the semantics of a complex question; then, we conduct probabilistic reasoning over HQDT from root to leaves recursively, to aggregate heterogeneous knowledge at different tree levels and search for a best solution considering the decomposing and answering probabilities. The experiments on complex QA datasets KQA Pro and Musique show that our framework outperforms SOTA methods significantly, demonstrating the effectiveness of leveraging question decomposing for knowledge integration and our RoHT framework.

1 Introduction

Explainable question answering (XQA) is the task of (i) answering a question and (ii) providing an explanation that enables the user to understand why the answer is selected (Neches et al., 1985; Schuff et al., 2020). It provides a qualified way to test the reasoning ability and interpretability of intelligent systems, and plays an important role in artificial intelligence (Lu et al., 2022).

---
[*]Indicates equal contribution.
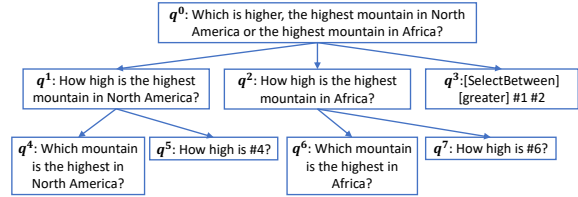[†]Corresponding author.

Figure 1: An example of Hierarchical Question Decomposition Tree (HQDT). q^i represents the index of node in its BFS ordering enumeration.

Recent work in XQA can be grouped into two directions: 1) neuro-symbolic methods (Berant et al., 2013; Liang et al., 2017; Cao et al., 2022b) translate natural language questions into formal representations (*e.g.*, SPARQL (Sun et al., 2020), KoPL (Cao et al., 2022a), lambda-DCS (Liang, 2013), *etc.*), whose execution on structured knowledge bases (KBs) gives the answer. Here, the formal representation acts as an explanation of the final answer. 2) Decompose-based models generate natural language intermediate steps that lead to the final answer (e.g., question decomposing which decomposes a complex question into sub-questions (Min et al., 2019; Perez et al., 2020; Deng et al., 2022), chain-of-thought prompting (Wei et al., 2022; Dua et al., 2022; Khot et al., 2022), etc.). Here, the intermediate steps shows the rationale of reasoning.

Although achieving significant results, both directions have key limitations. For neuro-symbolic methods, the formal representation can only be executed on KBs. However, even the largest KBs are incomplete, thus limits the recall of model. For decompose-based methods, they employ free-text corpora as the knowledge source, and the diversity of natural language makes XQA difficult. In fact, integrating knowledge from heterogeneous sources is of great importance to QA (Wolfson et al., 2020), especially for answering complex questions. Several attempts have been made for knowledge integration (e.g., KBs, text corpora) (Sun et al., 2018,

2019; Shi et al., 2021). Although promising, these graph-based methods suffer from lacking explainability or are constrained to limited reasoning capability.

Intuitively, leveraging question decomposing to integrate heterogeneous knowledge sources is a promising direction, since we can flexibly select the appropriate knowledge source for each sub-question. The challenges lie in: 1) How to determine the granularity of question decomposing, since certain complex questions can be directly answered with a knowledge source, and further decomposition increases the possibility of error. For example, in Figure 1, q^1 can be answered with the Wikipedia corpus without further decomposition. 2) How to find the optimal solution among various possible ones, since question decomposing and answering are both uncertain. For example, q^0 can also be decomposed as "Which mountains are in North America or Afirica", "*What's the height of #1*", "*[SelectAmong] [largest] #2*".

To this end, we propose a novel two-stage XQA framework *Reasoning over Hierarchical Question Decomspotion Tree*, dubbed RoHT. First, we propose to understand the complex question by building its *hierarchical question decomposition tree* (**HQDT**). In this tree, the root node is the original complex question, and each non-root node is a sub-question of its parent. The leaf nodes are atomic questions that cannot be further decomposed. Compared with existing representations that directly decompose a question into the atomic ones, e.g., QDMR (Wolfson et al., 2020), our tree structure provides the flexibility to determine solving a question whether by directly answering or further decomposing. Second, we propose **probabilistic reasoning over HQDT**, to fuse the knowledge from KB and text at different levels of the tree, and take into consideration the probability score of both tree generation and answering. The reasoning process is recursive, from the root to leaves, and constitues three steps: 1) a scheduler determines the appropriate knowledge sources for a particular question (from KB, text, or solving its children sequentially); 2) the corresponding executors output the answers with probabilities; 3) an aggregator aggregates the candidate answers from all the knowledge sources and outputs the best ones.

In evaluation, we instantiate our RoHT framework on two complex QA datasets: KQA Pro (Cao et al., 2022a), where we remove half of the triples in its KB and supplement it with Wikipedia corpus, and Musique (Trivedi et al., 2022), where we take Wikidata (Vrandecic and Krötzsch, 2014) as additional KB besides the given text paragraphs. Experimental results show that, RoHT improves the performance significantly under the KB+Text setting, by 29.7% and 45.8% EM score on KQA Pro and Musique compared with existing SOTA model. In addition, compared with the decompose-based methods, RoHT improves the SOTA by 11.3% F1 score on Musique.

Our contributions include: 1) proposing to leverage question decomposing to integrate heterogeneous knowledge sources for the first time; 2) designing a novel two-stage XQA framework RoHT by first building HQDT and then reasoning over HQDT; 3) demonstrating the effectiveness of our RoHT framework through extensive experiments and careful ablation studies on two benchmark datasets.

2 Related Work

2.1 QA over Text and KB

Over time, the QA task has evolved into two main streams: 1) QA over unstructured data (e.g., freetext corpora like Wikipedia); 2) QA over structured data (e.g., large structured KBs like DBpedia (Lehmann et al., 2015), Wikidata (Vrandecic and Krötzsch, 2014)). As structured and unstructured data are intuitively complementary information sources (Oguz et al., 2022), several attempts have been made to combines the best of both worlds.

An early approach IBM Watson (Ferrucci, 2012) combines multiple expert systems and re-ranks them to produce the answer. (Xu et al., 2016) maps relational phrases to KB and text simultaneously, and use an integer linear program model to provide a globally optimal solution. Universal schema based method (Das et al., 2017) reasons over both KBs and text by aligning them in a common embedded space. GraftNet (Sun et al., 2018) and its successor PullNet (Sun et al., 2019) incorporate free text into graph nodes to make texts amenable to KBQA methods. TransferNet (Shi et al., 2021) proposes the relation graph to model the label-form relation from KBs and text-form relation from corpora uniformly.

Although achieving promising results, these methods lack interpretability or are constrained to limited question type, *i.e.*, TransferNet shows

interpretability with transparent step transfering, however, it can only answer multi-hop questions, and cannot deal with questions that require attribute comparison or value verification. In contrast, our proposed framework shows great interpretability with HQDT and cover more question types.

2.2 Question Decomposing

For datasets, KQA Pro (Cao et al., 2022a) proposes to decompose a complex question into a multi-step program KoPL, which can be executed on KBs. BREAK (Wolfson et al., 2020) proposes to decompose questions into QDMR, which constitutes the ordered list of steps, expressed through natural language. Musique (Trivedi et al., 2022) is a QA dataset constructed by composing single-hop questions obtained from existing datasets, and thus naturally provides question decompositions.

For models, several attempts have been made for learning to decompose with weak-supervision, such as span prediction based method (Min et al., 2019), unsupervised sequence transduction method ONUS (Perez et al., 2020), AMR-based method QDAMR (Deng et al., 2022). Another line of work is to employ large language models with in-context learning, such as Least-to-most Prompting (Zhou et al., 2022), decomposed prompting (Khot et al., 2022), successive prompting (Dua et al., 2022).

Compared with existing works, we are the first to design a hierarchical question decomposition tree for integrating information from multiple knowledge sources.

3 Definition of HQDT

Formally, given a complex question, its HQDT is a tree T. Each node $q^i \in T$ represents a question. For root node, it represents the given complex question, and for non-root nodes, it represents a sub-question of its parent node. The leaf nodes are simple ("atomic") questions that cannot be decomposed. Note that HQDT is a 3-ary ordered tree. As shown in Figure 1, we enumerate the nodes of T with BFS ordering, and q^0 is the root question.

A question $q^i = \langle w_1, \cdots, w_j, \cdots, w_{|q^i|} \rangle$ can be categorized into one of the three types according to the token vocabulary: 1) natural language question (e.g., q^4: "*Which mountain is the highest in North America?*"), here, $w_j \in \mathcal{V}$, and $\mathcal{V}$ is the word vocabulary; 2) bridge question (e.g., q^5: "*How high is #4?*"), here, $w_j \in \mathcal{V} \cup \mathcal{R}$, and $\mathcal{R}$ is the reference token vocabulary. In this ques-

tion, "#4" refers to the answer of q^4, which is the sibling question of q^5; 3) symbolic operation question (e.g., q^3: "*[SelectBetween][greater] #1 #2*"), here, $w_j \in \mathcal{V} \cup \mathcal{R} \cup \mathcal{O}$, and $\mathcal{O}$ is the vocabulary of pre-defined symbolic operations, which are designed for supporting various reasoning capacity (e.g., attribute comparison and set operation) and are shown in appendix A in details. Note that all the bridge questions and symbolic operation questions are atomic questions and can only appear in leaf nodes.

For every non-leaf question q^i, we define two ordered lists:

- $q^i.children = \langle q^{st_i}, \cdots, q^{ed_i} \rangle$, which are children of q^i, successively indexed from st_i to ed_i. For example, for question q^1 in Figure 1, $q^1.children$ is $\langle q^4, q^5 \rangle$.

- $q^i.atoms = \langle a_1^i, \cdots, a_{n_i}^i \rangle$, which is a list of atomic questions deduced from the n_i leaf nodes of the sub-tree rooted by q^i, by rearranging the reference tokens. For example, for q^0 in Figure 1, its leaf nodes is $\langle q^4, q^5, q^6, q^7, q^3 \rangle$, and correspondingly, $q^0.atoms$ is $\langle q^4, \tilde{q}^5, q^6, \tilde{q}^7, \tilde{q}^3 \rangle$, with $\tilde{q}^5$ as "*How high is #1?*", $\tilde{q}^7$ as "*How high is #3*", and $\tilde{q}^3$ as "*[SelectBetween][greater] #2 #4*". The detailed deduction algorithm is in appendix B due to space limit. We also call $q^i.atoms$ the *atomic representation* of q^i.

Specially, among $q^i.children$, $q^{st_i}, \ldots, q^{ed_i-1}$ are all natural language questions, and q^{ed_i} is either a bridge question or a symbolic operation question. Answering q^i is semantically equivalent to answering sub-questions in $q^i.children$ or in $q^i.atoms$ sequentially. The last question in $q^i.children$ or $q^i.atoms$ returns the answer of q^i.

4 Methodology

Our framework RoHT is composed of two stages: 1) Building HQDT. We understand the hierarchical compositional structure of a complex question q^0 by generating its HQDT T with probability, where each question $q^i \in T$ has a score p_g^i that represents the certainty of its generation.

2) Probabilistic Reasoning over HQDT. We conduct recursive probabilistic reasoning over the HQDT from root to leaves to solve q^0. For each question q^i, we will utilize KBs, text and its child questions together to get a list R^i, which contains

answers of q^i with probabilistic scores. Finally the answer with the highest score in R^0 will be picked out as the final answer of q^0.

The details are introduced as follows.

4.1 Building HQDT

To build the HQDT for a complex question, we first generate its atomic representation, which corresponds the leaf nodes of HQDT, then generate every non-leaf nodes based on this atomic representation. We compute certainty score of each node based on the likelihood of each step of generation.

Building Leaf Nodes Given a complex question q^0, we first use a BART (Lewis et al., 2020)-based question decomposer M_θ to generate its atomic representation and output the likelihood of generation:

$$L^0, l_d = M_\theta(q^0). \quad (1)$$

Here, $L^0 = a_1^0 \langle sep \rangle a_2^0 \langle sep \rangle \ldots \langle sep \rangle a_{n_0}^0$ is the serialization of $q^0.atoms$, where $\langle sep \rangle$ is a separating token. $l_d = \Pr(L^0|q^0; \theta)$ is the likelihood of generation. Since q^0 is the root of T, each atomic question in $q^0.atoms$ corresponds to a leaf node in T (with the deterministic algorithm in Appendix C), and the certainty score of each leaf node in T is l_d.

Building Non-leaf Nodes Based on $q^0.atoms$, we can generate all the non-leaf questions in HQDT. The root question is just q^0 and thus has certainty score $p_g^0 = 1$. For every other non-leaf question q^i, its atomic representation $q^i.atoms = \langle a_1^i, \ldots, a_{n_i}^i \rangle$ can be translated from a specific subset of $q^0.atoms$ by rearranging the reference tokens. The subset can be determined by considering the reference relations of a bridge or symbolic operation question $a_j^0 \in q^0.atoms$, which corresponds to the leaf node q^{ed_i}, with other questions in $q^0.atoms$. We show the details in Appendix C. For example, $q^2.atoms$ in Figure 1 is ("*Which mountain is the highest in Africa?*", "*How high is #1?*"), and it can be obtained from (a_3^0, a_4^0) in $q^0.atoms$. Then we can use a BART-based question generator M_ϕ to generate q^i from $q^i.atoms$:

$$q^i, l_g^i = M_\phi(L^i), \quad (2)$$

where $L^i = a_1^i \langle sep \rangle a_2^i \langle sep \rangle \ldots \langle sep \rangle a_{n_i}^i$ is the serialized $q^i.atoms$, and $l_g^i = \Pr(q^i|L^i; \phi)$ is the likelihood of q^i given L^i. The certainty score of q^i is computed as:

$$p_g^i = l_d \cdot l_g^i. \quad (3)$$

Learning of Question Decomposer and Generator The question decomposer M_θ can be trained with paired $(q^0, q^0.atoms)$ data, where the atomic representation can be from either given annotation or unsupervised construction. The question generator M_ϕ can also be trained with the same data by exchanging the input and output. The details are shown in Section 5.2.

4.2 Probabilistic Reasoning over HQDT

$$f(q^i, p_g^i, G, C) \to R^i : \{(ans_j^i, p_j^i)\}, \quad (4)$$

where ans_j^i is an answer of q^i, and score p_j^i represents the certainty of ans_j^i.

As shown in Figure 3, the implementation of f contains tree steps: 1) a scheduler determines the suitable knowledge sources for a particular question, i.e., whether the question can be answered from KB, text, or by solving its child questions sequentially; 2) according to the suitable sources output by the scheduler, executors aim to get the answers with probabilities via executing on KB (KB executor) or retrieving from text (text executor), or answering the child questions (call f recursively); 3) an aggregator aggregates candidate answers from all the knowledge sources and outputs the top-k answers according to their probabilities. In the following, we will introduce their details when answering q^i.

Scheduler We formalize the scheduler as:

$$suit_{\text{kb}}, suit_{\text{text}}, suit_{\text{child}} = \text{Scheduler}(q^i, G, C), \quad (5)$$

Where $suit_{\text{kb}}$, $suit_{\text{text}}$ and $suit_{\text{child}}$ are 0/1 variables, respectively representing whether the answers of q^i are suitable to get from the KB G, the corpus C, or by solving $q^i.children$ sequentially.

Specifically, to check whether G is suitable, the scheduler employs a semantic parser (Cao et al., 2022a) M_{sp} to parse q^i into a program K with probability p_{parse}:

$$K, p_{\text{parse}} = M_{\text{sp}}(q^i). \quad (6)$$

Then it classifies the type of q^i according to the function skeleton of K. For example, the function skeleton of K in Figure 2 is "*Find-Relate-FilterConcept-SelectAmong*". If the precision of G on the questions that have the same function skeleton with K is larger than a predefined threshold γ [1], the scheduler will set $suit_{\text{kb}}$ to be 1.

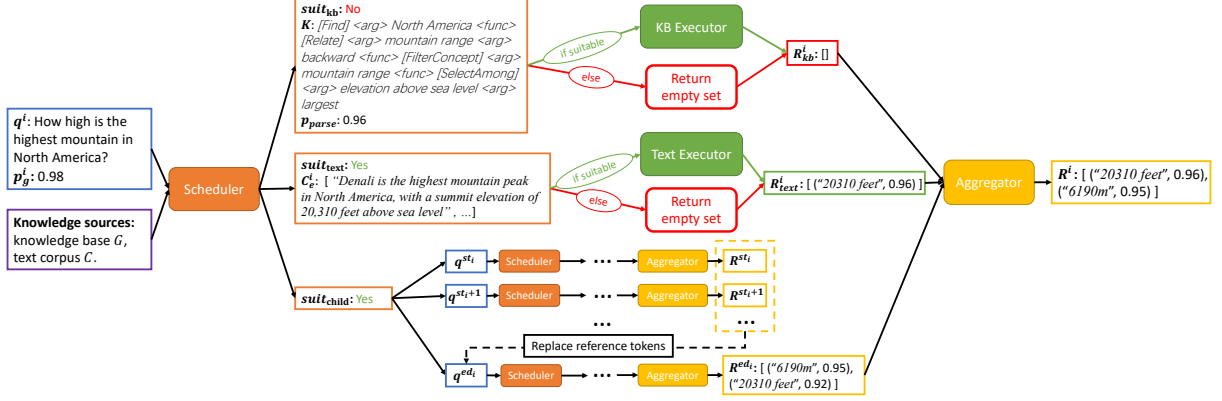[1] The precision of KB is calculated with questions in training set

Figure 2: Illustration of the recursive reasoning function f. For a question q^i, f uses the scheduler to determine suitable knowledge sources and calls executors to retrieve answers from them. f also recursively calls itself to get answers from the children of q^i. Finally the answers from different sources are fused by the aggregator.

To check whether the corpus C is suitable, the scheduler tries to find a set of evidence paragraphs for q^i. If C is too large, the scheduler will first use BM25 (Robertson and Zaragoza, 2009) to re-call dozens of most relevant paragraphs. For each paragraphs, we train a RoBERTa (Liu et al., 2019)-based selector M_{sl} to classify whether it is an evidence paragraph for q^i. Suppose the set of selected evidence paragraphs, C_e is not empty, the scheduler will set $suit_{text}$ as 1.

To make best use of knowledge from all levels, the scheduler simply set $suit_{child}$ to be 1 if q^i is a non-leaf question otherwise 0.

Executors For the KB executor, it takes the program K in Equation 6 on KB G to get the answers, and takes the parsing score p_{parse} in Equation 6 to calculate the probability score for each answer:

$$R_{kb}^i = \{(ans_{kb,j}^i, \ p_g^i \cdot p_{parse})\}. \tag{7}$$

For the text executor, it takes the selected paragraph set C_e as described above, and employs a Transformer-based reading comprehension model M_{rc} to extract answers from C_e:

$$\begin{aligned} \{(ans_{text,j}^i, p_{ex,j}^i)\} &= M_{rc}(q^i, C_e), \\ R_{text}^i &= \{(ans_{text,j}^i, \ p_g^i \cdot p_{ex,j}^i)\}. \end{aligned} \tag{8}$$

where $p_{ex,j}^i$ is the extraction probability of $ans_{text,j}^i$ given by M_{rc}.

For solving q^i by answering its children, f will recursively call itself to solve $q^{st_i}, \ldots, q^{ed_i}$ in or-der:

$$\begin{aligned} R^{st_i} &= f(q^{st_i}, p_g^{st_i}, G, C), \\ R^{st_i+1} &= f(q^{st_i+1}, p_g^{st_i+1}, G, C), \end{aligned} \tag{9}$$

$$\ldots$$

$$R^{ed_i} = f_{ref}(q^{ed_i}, p_g^{ed_i}, G, C, [R^{st_i}, \ldots, R^{ed_i-1}]),$$

and let

$$R_{child}^i = R^{ed_i}. \tag{10}$$

Here, f_{ref} is a variant of f to solve bridge and symbolic questions, which refer to the answers of their sibling questions. Suppose q^{ed_i} refers to the answers of its siblings $q^{r_1}, \ldots, q^{r_{h_i}}$ in order. If q^{ed_i} is a bridge question, f_{ref} will 1) convert q^{ed_i} into several possible natural language question $q_{nl}^1, \ldots, q_{nl}^K$ by replacing the reference tokens with every combination $((x_1^k, v_1^k), \ldots, (x_{h_i}^k, v_{h_i}^k)) \in R^{r_1} \times \cdots \times R^{r_{h_i}}$, 2) call f to solve each q_{nl}^k and 3) fuse the answers from each R_{nl}^k and select the top-k answers with the highest scores:

$$\begin{aligned} \{(ans_{nl,j}^k, \ p_{nl,j}^k)\} &= f(q_{nl}^j, p_g^i, G, C), \\ R_{nl}^k &= \{(ans_{nl,j}^k, \ Avg(p_{nl,j}^k, v_1^k, \ldots, v_{h_i}^k))\}, \\ R^{ed_i} &= Select(R_{nl}^1, \ldots, R_{nl}^K) \end{aligned} \tag{11}$$

Note that the score of answer $ans_{nl,j}^k$ is computed by averaging $p_{nl,j}^k$ and $v_1^k, \ldots, v_{h_i}^k$, instead of multiplying them, to avoid exponential shrink during recursion. If q^{ed_i} is a symbolic operation question with operation op and arguments, f_{ref} will execute simple program to apply the operation op over $R^{r_1}, \ldots, R^{r_{h_i}}$ to get R^{ed_i}. The score of each answer $ans_j^{ed_i}$ is computed as the average of $p_g^{ed_i}$ and the scores of answers in $R^{r_1}, \ldots, R^{r_{h_i}}$ used by the program to get $ans_j^{ed_i}$.

Aggregator The aggregator fuses R_{kb}^i, R_{text}^i and R_{child}^i by selecting the top-k answers with the highest scores from them. If several answers have the same surface form, only the one with the highest score will be preserved.

$$R^i = \text{Aggregator}(R_{\text{kb}}^i, R_{\text{text}}^i, R_{\text{child}}^i). \quad (12)$$

5 Experiments

5.1 Datasets

Currently, there are few high-quality complex QA datasets based on both KBs and text. Previous methods (Sun et al., 2018, 2019; Shi et al., 2021) evaluated their models on MetaQA (Zhang et al., 2018) by pairing its KB with the text corpus of WikiMovies (Miller et al., 2016). However, the questions in MetaQA are too simple since there are only 9 relations in its KB. Therefore, we conduct our experiments on two more challenging complex QA datasets: KQA Pro and Musique, and their details are as follows.

KQA Pro (Cao et al., 2022a) is a large scale complex QA dataset, including 120k diverse natural language questions up to 5 hops over KB. Its KB is a subset of Wikidata (Vrandecic and Krötzsch, 2014), and consists of 16k entities, 363 predicates, 794 concepts and 890k triple facts. For each question, KQA Pro also provides the corresponding KoPL program. To simulate the realistic case where KB is incomplete, following (Sun et al., 2019; Shi et al., 2021), we randomly discard 50% triples in the KB and take Wikipedia as supplementary text corpus.

Musique (Trivedi et al., 2022) is a multi-hop QA dataset over text, including 25k 2-4 hop questions. We evaluate our framework under Musique-Ans setting where all the questions are answerable. Its questions are carefully constructed from several single-hop QA datasets via manually composition and paraphrase, and are hard to cheat via reasoning shortcut. For each complex question, Musique gives 20 paragraphs (including annotated evidence paragraphs and distractor paragraphs) as the corpus. Specially, for each question in the training set, Musique also provides a golden atomic representation, together with the answer and the evidence paragraph of each atomic question. In addition to the given paragraphs, we choose Wikidata as the KB to acquire additional knowledge.

5.2 Implementations

KQA Pro For the experiments of KQA Pro, a key challenge is that there are no annotations for atomic representation, which are required for training the question decomposer and generator in RoHT. Because the KoPL program of a complex question follows context free grammar, every atomic question will correspond to a specific span of the program. Therefore we first split the KoPL program into sub-programs according to the grammar, then use each sub-program to generate the atomic question by applying BART model fintuned with the (KoPL, question) pairs from the original dataset. For the answers for each atomic question, we execute the corresponding sub-programs on the KB to get corresponding answers. Using these constructed atomic representations, we train two BART-base models as the question decomposer and generator, respectively.

For the scheduler, we directly use the semantic parser trained by (Cao et al., 2022a) on KQAPro, and set the precision threshold γ to be 0.7. We train a RoBERTa-large as the evidence selector via weak supervised method: for each question in the training set and constructed atomic representations, we first use BM25 to recall 10 related paragraphs from wikipedia, then take the paragraphs that contain the answer as positive samples and take other recalled paragraphs as negative samples. For the text executor, we also train a BART-large reading comprehension model on these positive samples.

Musique Since Musique provides golden atomic representation for every complex question in the training set, we directly use them to train BART-base models as question decomposer and generator. For the scheduler, we adapt semantic parser trained by (Cao et al., 2022a) on Wikidata. The KB precision threshold γ is set to be 0.4, which is determined by the top-10 types of questions with the highest precision. We train the RoBERTa selector model on complex and atomic questions in the training set together, taking annotated evidence paragraphs as positive samples and distractor paragraphs as negative samples. For the text executor, we pre-train a Longformer-large (Beltagy et al., 2020) reading comprehension model on SQUAD (Rajpurkar et al., 2016), then finetune it on complex questions and atomic questions of Musique.

Model	Overall	Multihop	Qualifier	Comparison	Logical	Count	Verify	Zero-shot
50% KB								
KVMemNN	17.72	17.63	18.53	1.39	15.48	28.38	59.30	0.06
RGCN	34.77	33.71	28.44	31.46	35.39	**39.76**	64.27	0.06
BART KoPL	38.04	33.10	29.40	51.81	29.92	33.69	60.12	29.03
RoHT^{KB}	38.94	34.16	31.54	50.91	31.61	33.69	60.4	30.52
50%KB + Text								
TransferNet	16.80	15.94	17.93	45.35	14.84	10.47	0.00	8.43
RoHT^{mix}	**46.45**	**41.76**	**41.73**	**52.21**	**41.95**	31.26	**65.45**	**38.76**

Table 1: EM results on the dev set of KQA Pro. RoHT outperforms all the baselines by a large margin and achieves the best performance on most types of questions.

5.3 Baselines

we compare RoHT with several representative methods for complex QA, including memory-based methods, graph-based methods, and XQA methods.
KVMemNN (Miller et al., 2016) stores encoded knowledge in key-value memory and iteratively reads the memory to update the query vector to conduct multi-hop reasoning.
RGCN (Schlichtkrull et al., 2018) is a variant of graph convolutional network and utilizes the graph structure of KB to tackle complex questions.
BART KoPL (Cao et al., 2022a) is a BART-based semantic parser which can convert complex question into KoPL program. It achieves over 90% accuracy on KQA Pro on the complete KB.
SA (Trivedi et al., 2022) is a two-stage model that first uses a RoBERTa-large selector to rank and select the K most relevant paragraphs with the question and then uses a Longformer-large answerer to predict answer based on selected paragraphs.
EX(SA) (Trivedi et al., 2022) is the state-of-the-art model on Musique. It first explicitly decomposes the complex question into atomic representation and then calling SA model repeatedly to answer each atomic question in order.
TransferNet (Shi et al., 2021) iteratively transfer entity scores via activated path on the relation graph that consists of both text-form relations and KB-form relations. It is existing state-of-the-art model that utilizes both KBs and text as knowledge soruces, and nearly solves MetaQA. We reimplement it on both KQA Pro and Musique, and the details are shown in Appendix D.
RoHT: RoHT^{KB}, RoHT^{text} and RoHT^{mix} denote the RoHT models that only use KB, only use text and use both KB and text, respectively.

5.4 Main Results

5.4.1 Results on KQA Pro

The experimental results for KQA Pro are shown in Table 1. When using only the incomplete KB, RoHT^{KB} model respectively improves EM by 21.22, 4.17 and 0.90 compared to KVMemNN, RGCN and BART KoPL, showing the benefit of integrating the answers of sub-questions of different levels. After adding Wikipedia as supplementary text corpus, RoHT^{mix} yields substantial improvement compared with RoHT^{KB} (7.51 on EM), demonstrating the effectiveness of utilizing knowledge from KB and text together. RoHT^{mix} also outperforms TransferNet, which is end-to-endly trained with a mixed relation graph, by a large margin (29.65 on EM). This is because unlike graph-based methods, RoHT explicitly shows the compositional structure of a complex question in natural language form via HQDT generation, and thus can retrieve answers from the KB and text with more advanced and flexible sub-modules (e.g., semantic parser and reading comprehension model). Moreover, our designed atomic operations in the HQDT also enable RoHT to solve a wide variety of complex questions: we can see that RoHT^{mix} achieves the best results on 6 types of questions among 7 types, showing comprehensive reasoning capacity.

5.4.2 Results on Musique

Table 2 presents the results on the dev set of Musique dataset. As expected, our RoHT models show significant improvement over all the baselines. With only given paragraphs, RoHT^{text} improves EM/F1 by 13.8/14.3 and 11.6/11.9 compared with SA and EX(SA), respectively; With both text and KB, the performance of RoHT^{mix} is also remarkably better than TransferNet (62.3 v.s. 10.9 on F1). Comparing RoHT^{text} and RoHT^{mix},

Model	EM	F1
Text		
SA	39.3	47.3
EX(SA)	41.5	49.7
RoHT$^{\text{text}}$	53.1	61.6
Text+KB		
TransferNet	8.6	10.9
RoHT$^{\text{mix}}$	**54.4**	**62.3**

Table 2: EM and F1 results on the dev set of Musique. Compared with state-of-the-art methods, RoHT achieves significant improvement.

Model	KQA Pro	Musique
RoHT$^{\text{mix}}$	46.5	54.4
w/o scheduler	40.7	47.0
RoAT$^{\text{mix}}$	32.3	47.6

Table 3: EM performance of RoHT$^{\text{mix}}$ with and without scheduler, and EM performance of RoAT$^{\text{mix}}$.

we can also see some benefits of supplementing the text information with KB information, though the improvement is smaller than supplementing the KB with text on KQA Pro because KBs have lower coverage than text and the semantic parser is not specially finetuned for questions of Musique. We submit the predictions of RoHT$^{\text{mix}}$ on the test set and achieve 63.6 F1 score, which significantly outperforms the best public result 52.3.

5.5 Further Analysis

5.5.1 Effect of Scheduler

To show the effect of the scheduler module, we remove it from the RoHT$^{\text{mix}}$ model, i.e, default that the KB and recalled/given text paragraphs are suitable for all questions in the HQDT, and evaluate the performance again on the dev set of KQA Pro and Musique. The results are shown in Table 3. We can see that after discarding the scheduler, the EM performance on KQA Pro and Musique drops by 5.8 and 7.4, respectively. Therefore, it is important to use the scheduler to select suitable knowledge sources for each question.

5.5.2 Effect of Hierarchical Decomposition

Many existing methods generate non-hierarchical decomposition of complex questions, similar to the atomic representation, to assist reasoning (Min et al., 2019; Wolfson et al., 2020; Deng et al., 2022). To demonstrate the superiority of hierarchical decomposition, we compare our RoHT$^{\text{mix}}$ model with
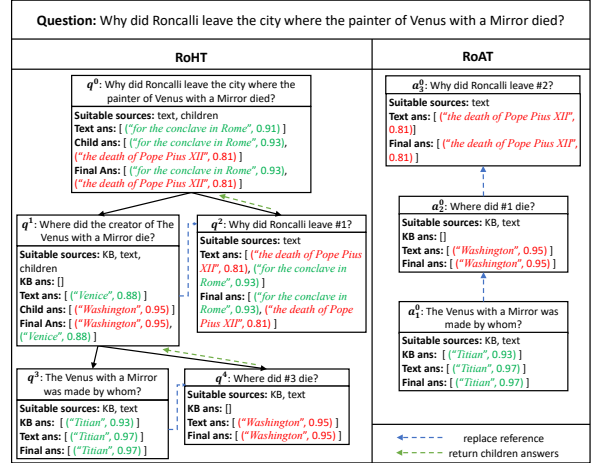

Figure 3: A case from Musique. We mark the correct answers in green and the wrong answers in red.

RoAT$^{\text{mix}}$ model, which uses the same scheduler, executors, and aggregator as RoHT$^{\text{mix}}$, but solves the complex question by directly answering the atomic questions in its atomic representation in order. As shown in Table 3, RoHT$^{\text{mix}}$ outperforms RoAT$^{\text{mix}}$ by a large margin on both KQA Pro and Musique. This is because the hierarchical structure of HQDT enables RoHT model to fuse the knowledge from KBs and text at different question levels, and to discard wrong answers via comparing the problisitic scores of answers.

To further understand the reason, we show a case from Musique in Figure 3. We can see that both RoHT$^{\text{mix}}$ and RoAT$^{\text{mix}}$ fail to answer the question *"Where did (Titian) die?"* (q^4 in the left, a_2^0 in the right). However, RoHT$^{\text{mix}}$ directly extracts the correct answer of q^1 from text and finally gets the correct answer of q^0 with the highest score, while RoHT$^{\text{mix}}$ fails to solve a_3^0 because it must rely on the wrong answer from a_2^0.

6 Conclusion

In this paper, we propose RoHT, an understanding-reasoning XQA framework that uses both a KB and a text corpus to derive answers of complex questions. RoHT first builds the HQDT for a complex question to understand its hierarchical compositional structure, then conducts recursive probabilistic reasoning over the HQDT to solve the question, integrating answers from the KB, text, and sub-questions. Experiments show that RoHT significantly outperforms previous methods. We also demonstrate the superiority of HQDT compared with non-hierarchical decomposition.

7 Limitation

Currently, RoHT framework is restricted to incorporating KBs and text. However, since RoHT retrieves answers from each knowledge source in a separate way, it could in principle utilize knowledge from more heterogeneous sources such as tables, and we will study this in future work. In addition, a device with large storage space and memory is needed for the storage and usage of Wikipeida and Wikidata.

8 Ethics Statement

The data used in this paper are drawn from publicly published datasets, encyclopedias and knowledge bases. Most of them do not involve sensitive data.

9 Acknowledgement

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. 2022a. KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6101–6119. Association for Computational Linguistics.

Shulin Cao, Jiaxin Shi, Zijun Yao, Xin Lv, Jifan Yu, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jinghui Xiao. 2022b. Program transfer for answering complex questions over knowledge bases. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8128–8140. Association for Computational Linguistics.

Rajarshi Das, Manzil Zaheer, Siva Reddy, and Andrew McCallum. 2017. Question answering on knowledge bases and text using universal schema and memory networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 358–365. Association for Computational Linguistics.

Zhenyun Deng, Yonghua Zhu, Yang Chen, Michael Witbrock, and Patricia Riddle. 2022. Interpretable amr-based question decomposition for multi-hop question answering. pages 4093–4099.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. *CoRR*, abs/2212.04092.

David A. Ferrucci. 2012. Introduction to "this is watson". *IBM J. Res. Dev.*, 56(3):1.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *CoRR*, abs/2210.02406.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. pages 7871–7880.

Chen Liang, Jonathan Berant, Quoc V. Le, Kenneth D. Forbus, and Ni Lao. 2017. Neural symbolic machines: Learning semantic parsers on freebase with weak supervision. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 23–33. Association for Computational Linguistics.

Percy Liang. 2013. Lambda dependency-based compositional semantics. *CoRR*, abs/1309.4408.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. volume abs/2209.09513.

Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1400–1409. The Association for Computational Linguistics.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. pages 6097–6109.

Robert Neches, William R. Swartout, and Johanna D. Moore. 1985. Explainable (and maintainable) expert systems. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence. Los Angeles, CA, USA, August 1985*, pages 382–389. Morgan Kaufmann.

Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Sejr Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. 2022. Unik-qa: Unified representations of structured and unstructured knowledge for open-domain question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1535–1546. Association for Computational Linguistics.

Ethan Perez, Patrick S. H. Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. pages 8864–8880.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. pages 2383–2392.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.

Hendrik Schuff, Heike Adel, and Ngoc Thang Vu. 2020. F1 is not enough! models and evaluation towards user-centered explainable question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7076–7095. Association for Computational Linguistics.

Jiaxin Shi, Shulin Cao, Lei Hou, Juanzi Li, and Hanwang Zhang. 2021. Transfernet: An effective and transparent framework for multi-hop question answering over relation graph. In *Proceedings of the*

2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 4149–4158. Association for Computational Linguistics.

Haitian Sun, Tania Bedrax-Weiss, and William W. Cohen. 2019. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2380–2390. Association for Computational Linguistics.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W. Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. pages 4231–4242.

Yawei Sun, Lingling Zhang, Gong Cheng, and Yuzhong Qu. 2020. SPARQA: skeleton-based semantic parsing for complex questions over knowledge bases. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8952–8959. AAAI Press.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics*, 10:539–554.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Tomer Wolfson, Mor Geva, Ankit Gupta, Yoav Goldberg, Matt Gardner, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Trans. Assoc. Comput. Linguistics*, 8:183–198.

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Hybrid question answering over knowledge base and free text. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2397–2407. ACL.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J. Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *Proceedings of the Thirty-Second AAAI Conference*

on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 6069–6076. AAAI Press.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed H. Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *CoRR*, abs/2205.10625.

A Atomic Operations

We design 6 atomic operations: *Verify*, *SelectBetween*, *SelectAmong*, *Count*, *Intersection*, *Union*, to support various reasoning capacity. We show their input, output, and examples in Table 4.

B Get Atomic Representation from Leaf Nodes

Algorithm 1 describes that how to get the atomic representation of a question $q^i \in T$ from the leaf nodes in the sub-tree rooted by q^i.

Algorithm 1 Get Atomic Representation from Leaf Nodes

Input: An HQDT T and a index i.
Output: $q^i.atoms$

1: **function** DFS($j, atoms, ids, n$)
2:   **if** q^j is a leaf question **then**
3:     $n \leftarrow n + 1$
4:     $ids[j] \leftarrow n$
5:     $a \leftarrow q^j$
6:     **for** k in **GetRefTokens**(q^j) **do**
7:       **if** q^k is a leaf question **then**
8:         $a \leftarrow$ **ModifyRefTokens**($a, k, ids[k]$)
9:       **else**
10:        $a \leftarrow$ **ModifyRefTokens**($a, k, ids[ed_k]$)
11:     $atoms$.append(a)
12:     **return**
13:   **for** $k \leftarrow st_j, \ldots, ed_j$ **do**
14:     **Dfs**(k)
15:
16: $q^i.atoms \leftarrow []$
17: $ids \leftarrow$ empty dict
18: **Dfs**($i, q^i.atoms, ids, 0$)

C Pseudocode for Building HQDT

Algorithm 2 shows the pseudocode for generating the HQDT of a complex question with probability.

D Reimplementation of TransferNet

To reimplemente TransferNet, we build the mixed relation graphs that consist of both label-form relations (i.e., KB triples) and text-form relations for KQA Pro and Musique, respectively, and train the models with the open source code. We show the details of graph building as follows.

KQA Pro We follow the method used by the original paper on MetaQA to build the relation graph of KQA Pro. As mentioned in Section 5.2, we use half of its KB triples as the label form. We constructe the text form by extracting sentences from Wikipedia. Following the original paper, we use exact match of surface forms for entity recognition

---

Algorithm 2 Generation of HQDT

Input: a complex question q^0, a question decomposer M_θ, a question generator M_ϕ.
Output: a list T representing the HQDT, where element (q^i, p_g^i, fa_i) in T denote a sub-question q^i, certainty score of q^i and the father of q^i, respectively.

1: **function** REARRANGEREFTOKENS(ar)
2:   $atoms \leftarrow []$
3:   $ids \leftarrow$ empty dict
4:   $h \leftarrow 0$
5:   **for** (i, a_i) in ar **do**
6:     $h \leftarrow h + 1$
7:     $ids[i] \leftarrow h$
8:     **for** k in **GetRefTokens**(a_i) **do**
9:       $a_i \leftarrow$ **ModifyRefTokens**($a_i, k, ids[k]$)
10:    $atoms$.append(a_i)
11:   **return** $atoms$
12:
13: $([a_1^0, \ldots, a_{n_0}^0], l_d) \leftarrow M_\theta(q^0)$
14: $n \leftarrow n_0$
15: $T \leftarrow []$
16: **for** $i \leftarrow 1, 2, \ldots, n_0$ **do**
17:   $(q^i, p_g^i) \leftarrow (a_i^0, l_d)$
18:   $ar^i \leftarrow [(i, a_i^0)]$
19:   **if** a_i^0 contains referring tokens **then**
20:     $r_1, \ldots, r_h \leftarrow$ **GetRefTokens**(a_i^0)
21:     $n \leftarrow n + 1$
22:     $ar^n \leftarrow []$
23:     **for** $j \leftarrow r_1, \ldots, r_h, i$ **do**
24:       **if** fa_j has been identified **then**
25:         $q^i \leftarrow$ **ModifyRefTokens**(q^i, j, fa_j)
26:         $j \leftarrow fa_j$
27:       $fa_j \leftarrow n$
28:       T.append((q^j, p_g^j, fa_j))
29:       ar^n.extend(ar^j)
30:     $q^n.atoms \leftarrow$ **RearrangeRefTokens**(ar^n)
31:     $L^n =$ **Serialize**($q^n.atoms$)
32:     $(q^n, l_g^n) \leftarrow M_\phi(L^n)$
33:     $p_g^n \leftarrow l_d \cdot l_g^n$
34: T.append($(q^0, 1, 0)$)   ▷ directly use q^0 as root
35: $T \leftarrow$ **ReIndexByBFS**(T)
36: **return** T

---

and linking. For every entity in the KB, we recall all the paragraphs in Wikipedia titled by it, then take the entity as subject and other relevant entities appeared in these paragraphs as objects. The sentences that contain the objects are selected as the relation texts. The recall of answer is 51%, i.e, for 51% questions, there exist a complete path from the topic entity to the answer in the relation graph, and this is a upper bound for the performance of TransferNet.

Musique For each question in Musique, we utilize the 20 given paragraphs to build individual relation graph. Specifically, we first identify entities mentioned in these paragraphs via Spacy (Honnibal et al., 2020) and exact match of surface forms with Wikidata entities. Then we take the co-occuring

Operation	Argument	Input → Output	Example
Verify	Value, > / < / = / !=	(Value) → (Bool)	[Verify] [2005] [<] #3 (1998) → (yes)
SelectBetween	greater / smaller	(Value, Ent) (Value, Ent) → (ENT)	[SelectBetween] [smaller] #3 #4 (6670 km, Nile River) (6440 km, Amazon River) → (Amazon River)
SelectAmong	largest / smallest	[(Value, Ent)] → (Ent)	[SelectAmong] [largest] #1 [(8848m, Everest) (8611m, K2) (8516m, Makalu)] → (Everest)
Count	/	[(Ent)] → (Value)	[Count] #2 [(Bronny James) (Bryce James) (Zhuri James)] → (3)
Intersection	/	[(Ent)] [(Ent)] → [(Ent)]	[Intersection] #1 #2 [(apple) (orange) (peach)] [(orange)] → [(orange)]
Union	/	[(Ent)] [(Ent)] → [(Ent)]	[Union] #1 #2 [(apple) (orange)] [(orange) (peach)] → [(apple) (orange) (peach)]

Table 4: Atomic operations proposed for QDKT, along with the corresponding examples. Ent, Value, Pred, and Bool denote entity, attribute, predicate, and boolean variable, respectively. () means tuple and [] means list.

sentences of two entities as the text-form, and take the triples in Wikidata whose subject or object is one of these entities as the label-form. The recall of answer is 72%.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☑ A2. Did you discuss any potential risks of your work?
Section 8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We follow the same license

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The data used in this work are drawn from publicly published encyclopedias, knowledge bases and datasets.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.