Relation-Aware Language-Graph Transformer for Question Answering

Jinyoung Park[1*], Hyeong Kyu Choi[1*], Juyeon Ko[1*], Hyeonjin Park[2], Ji-Hoon Kim[2,3,4], Jisu Jeong[2,3,4], Kyungmin Kim[2,3,4], Hyunwoo J. Kim[1†]

[1]Korea University
[2] NAVER
[3] NAVER Cloud
[4] NAVER AI Lab
{lpmn678, imhgchoi, juyon98, hyunwoojkim}@korea.ac.kr
{hyeonjin.park.ml, genesis.kim, jisu.jeong, kyungmin.kim.ml}@navercorp.com

Abstract

Question Answering (QA) is a task that entails reasoning over natural language contexts, and many relevant works augment language models (LMs) with graph neural networks (GNNs) to encode the Knowledge Graph (KG) information. However, most existing GNN-based modules for QA do not take advantage of rich relational information of KGs and depend on limited information interaction between the LM and the KG. To address these issues, we propose Question Answering Transformer (QAT), which is designed to jointly reason over language and graphs with respect to entity relations in a unified manner. Specifically, QAT constructs Meta-Path tokens, which learn relation-centric embeddings based on diverse structural and semantic relations. Then, our Relation-Aware Self-Attention module comprehensively integrates different modalities via the Cross-Modal Relative Position Bias, which guides information exchange between relevant entites of different modalities. We validate the effectiveness of QAT on commonsense question answering datasets like CommonsenseQA and OpenBookQA, and on a medical question answering dataset, MedQA-USMLE. On all the datasets, our method achieves state-of-the-art performance. Our code is available at http://github.com/mlvlab/QAT.

1 Introduction

Question Answering (QA) is a challenging task that entails complex reasoning over natural language contexts. The large-scale pre-trained language models (LMs) have proven to be capable of processing such contexts, by encoding the implicit knowledge in textual data. To deal with various QA tasks, both generic and domain-specific LMs (Liu et al. 2019; Lee et al. 2020; Liu et al. 2021a) have been employed. But the QA systems only with LMs often struggle with structured reasoning due to the lack of explicit and structured knowledge. Thus, many works (Mihaylov and Frank 2018; Lin et al. 2019; Feng et al. 2020; Wang et al. 2022; Yasunaga et al. 2021; Zhang et al. 2022) in the literature have studied to effectively incorporate Knowledge Graphs (KGs) into the question-answering system.

Most existing efforts for fusing LM and KG utilized a separate graph neural network (GNN) module to process the KG and integrated it with LM in the 'final' prediction step (Feng et al. 2020; Yasunaga et al. 2021; Wang et al. 2022). However, recent works pinpointed the shortcomings of the shallow fusion, and have proposed methods that fuse LM and KG early in the intermediate layers. They aimed to find more appropriate ways to exchange information between the two modalities, via special tokens and nodes (Zhang et al. 2022) or cross-attention (Sun et al. 2022). But the approaches have a modality-specific encoder GNN and the information exchange occurs only on fusion layers resulting in modest information exchange. In addition, the GNNs in previous works are *not* suited for fully exploiting KG data, which comprise triplets of two entities and their relations. To process the rich relational information, a more *relation-centric* KG processor is necessary, in place of the *node-centric* GNNs that learn node features with the message-passing scheme.

To this end, we propose the **Question Answering Transformer (QAT)**, specifically designed to jointly reason over the LM and KG in a unified manner, without an explicit KG encoder, *e.g.*, GNN. We encode relation-centric information of the KG in the form of Meta-Path tokens. Along with LM outputs, the Meta-Path tokens are input into our Relation-Aware Self-Attention (RASA) module, whose cross-modal information integration is facilitated by our Cross-Modal Relative Position Bias. The diverse structural and semantic relations along the meta-paths between question-answer entities can be easily integrated with the LM context via Meta-Path Tokens. Also, the Meta-Path token sets tend to be diverse, bringing enhanced expressivity in self-attention. Moreover, we present a method to exchange information in a unified manner, guided by the Cross-Modal Relative Position Bias. This provides a more flexible means to fuse LM and KG, encouraging information exchange between relevant entities across different modalities.

Then, our **contributions** are as follows:

- We introduce the Meta-Path token, a novel embedding to encode KG information based on diverse relations along meta-paths.
- We present Cross-Modal Relative Position Bias, which

[*]These authors contributed equally.
[†]is the corresponding author

enables more flexible information exchange encouraging interactions between relevant entities across different modalities: LM and KG.

- We propose Question Answering Transformer, a relation-aware language-graph transformer for question answering by joint reasoning over LM and KG.
- We achieve state-of-the-art performances on commonsense question answering: CommonsenseQA and OpenBookQA; and on medical question answering: MedQA-USMLE, thereby demonstrating our method's joint reasoning capability.

2 Related Works

Knowledge Graph based QA. KGs are structured data containing relational information on concepts (*e.g.*, ConceptNet (Speer, Chin, and Havasi 2017), Freebase (Bollacker et al. 2008), and Wikidata (Vrandečić and Krötzsch 2014)), and they have been widely explored in the QA domain. Recent studies on KG based QA process the KG in an end-to-end manner with graph neural networks (Feng et al. 2020; Yasunaga et al. 2021; Wang et al. 2022) or language models (Wang et al. 2020; Bansal et al. 2022) for knowledge module. Most of them focus on adequately processing the KG first, while independently training the LM in parallel. The two modalities are then fused in the final step to render a QA prediction. But this shallow fusion does not facilitate interactions between the two modalities, and several methods (Sun et al. 2022; Zhang et al. 2022) to fuse LM and KG in the earlier layers have recently been introduced as well. The commonality of these works is that the KG nodes are processed with the GNN message passing scheme, which are then fused with LM embeddings. Contrary to this line of works, our method manages the joint reasoning process by explicitly learning the multi-hop relationships, and facilitating active cross-modal attention.

Multi-relational Graph Encoder. GNNs have shown superior performance on various type of graphs in general (Kipf and Welling 2017; Veličković et al. 2018). However, additional operations may be required to encode the rich relational information of multi-relational graphs like the KGs. The KG is a directed graph whose edges are multi-relational, and several relevant works (Schlichtkrull et al. 2018; Wang et al. 2019) have been introduced to process such graph types. Unlike homogeneous graph neural networks such as the GCN (Kipf and Welling 2017), exploiting rich semantic information is essential to encoding multi-relational graphs like KGs. For instance, R-GCN (Schlichtkrull et al. 2018) performs graph convolution for each relation to consider the different interactions between entities. HAN (Wang et al. 2019) adopted meta-paths to enrich semantic information between nodes by defining path-based neighborhoods. Moreover, (Hwang et al. 2020) has proposed meta-path prediction as a self-supervised learning task to encourage GNNs to comprehend relational information.

3 Preliminaries

Question Answering with Knowledge Graphs. In this paper, we consider the question answering task via joint reasoning over the LM and KG. The objective of the task is to understand the context of the given question by jointly reasoning on the language and structured relational data. Specifically, we use masked LMs (*e.g.*, RoBERTa (Liu et al. 2019), SapBERT (Liu et al. 2021a)) and multi-relational KGs (*e.g.*, ConceptNet).

In multiple-choice question answering (MCQA), word entities in question q and answer choice $a \in \mathcal{C}$ are concatenated to form an input sequence $\mathcal{X} = \left[[\text{cls}], \boldsymbol{x}_{q_1}, \ldots, \boldsymbol{x}_{q_{|q|}}, [\text{sep}], \boldsymbol{x}_{a_1}, \ldots, \boldsymbol{x}_{a_{|a|}}\right]$. Then, a pre-trained language model g_{LM} computes the context tokens $\mathcal{H}_{\text{LM}}$ as $\mathcal{H}_{\text{LM}} = g_{\text{LM}}(\mathcal{X})$. To enable joint reasoning with external knowledge, we leverage the KG containing structured relational data. Following prior works (Lin et al. 2019; Yasunaga et al. 2021; Wang et al. 2022), we extract a multi-relational subgraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from the full KG for each question and answer choice pair. We denote $\mathcal{V}$ as a set of nodes (entities) and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$ as the set of subgraph edges where $\mathcal{R}$ indicates the edge types between entities. Also, we respectively denote $\mathcal{V}_q \subseteq \mathcal{V}$ and $\mathcal{V}_a \subseteq \mathcal{V}$ as the set of nodes mentioned in the given question q and answer choice a. Given the multi-relational subgraph, a graph encoder g_{KG} encodes $\mathcal{G}$ as $\mathcal{H}_{\text{KG}} = g_{\text{KG}}(\mathcal{G})$. Given multimodal information $\mathcal{H}_{\text{LM}}$ and $\mathcal{H}_{\text{KG}}$, module $\mathcal{F}$ can be applied to output a joint representation $\mathcal{Z}_{a \in \mathcal{C}} = \mathcal{F}(\mathcal{H}_{\text{LM}}, \mathcal{H}_{\text{KG}})$. Here, $\mathcal{F}$ can be as simple as feature concatenation or as complex as the Transformer module (Vaswani et al. 2017).

Meta-Paths. The meta-path (Sun et al. 2011) is defined as a path on a graph $\mathcal{G}$ whose edges are multi-relational edges, *i.e.*, $v_1 \xrightarrow{r_1} v_2 \xrightarrow{r_2} \ldots \xrightarrow{r_l} v_{l+1}$, where $r_l \in \mathcal{R}$ refers to the edge type of the l^{th} edge within the meta-path. Then, we may view the meta-path as a composite relation $\mathbf{r} = r_1 \circ r_2 \circ \ldots \circ r_l$ that defines the relationship between nodes v_1 and v_{l+1}. The meta-path can be naturally extended to the multi-hop relation encoding, providing a useful means to define the relationship between an arbitrary entity pair from the KG subgraph. For instance, a meta-path (jeans – *related_to* $\rightarrow$ merchants – *at_location* $\rightarrow$ shopping_mall) can be defined to describe the relationship between 'jeans' and 'shopping mall' for the question in Figure 3.

4 Question Answering Transformer

In this section, we introduce Question Answering Transformer (QAT). QAT is a relation-aware language-graph Transformer for knowledge-based question answering, which performs joint reasoning over LM and KG. We take a novel approach of encoding the *meta-path* between KG entities to jointly self-attend to KG and LM representations. This is enabled by our Meta-Path tokens (Section 4.1) and our Relation-Aware Self-Attention (Section 4.2) module.

4.1 Meta-Path Token Embeddings

The gist of question answering is to understand the relationship between entities. Accordingly, QAT learns embed-

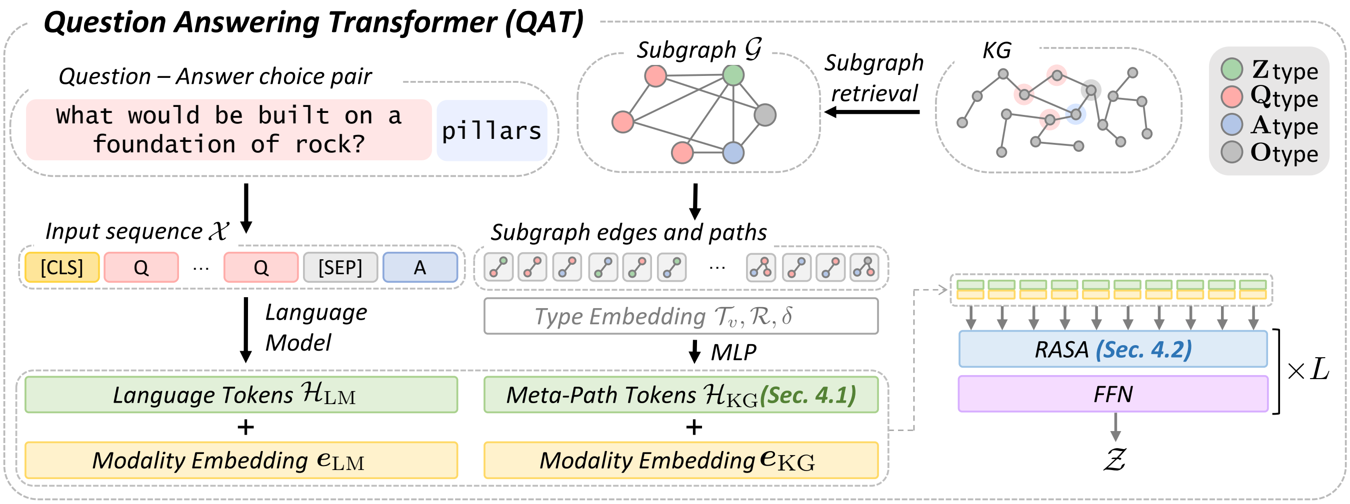

Figure 1: Overall architecture of QAT. The Question Answering Transformer (QAT) is a relation-aware language-graph transformer for knowledge-based question answering tasks. Given a question-answer choice pair, a relevant subgraph $\mathcal{G}$ is retrieved from the knowledge graph. Then, the Meta-Path token is encoded for each n-hop meta-path with type embeddings based on structural and semantic relationships between nodes (Section 4.1). Then, the set of Meta-Path tokens $\mathcal{H}_{\text{KG}}$ are concatenated with the LM outputs $\mathcal{H}_{\text{LM}}$, to be fed into our Relation-Aware Self-Attention module. The Relation-Aware Self-Attention module effectively aggregates multimodal information via the Cross-Modal Relative Position Bias (Section 4.2). After L attention blocks, our QAT renders $\mathcal{Z}$ for each answer choice, which is later passed through a prediction head to output a logit value.

dings that represent the relationship between each node pair, dubbed Meta-Path tokens.

When encoding the edge feature, the edge type $r \in \mathcal{R}$ is utilized in the form of a one-hot vector. The node types are also provided as input, but the node features are not directly employed when learning the embeddings. Following Yasunaga et al., a node type among $\mathcal{T}_v = \{\mathbf{Z}, \mathbf{Q}, \mathbf{A}, \mathbf{O}\}$ is selected based on the node's source. To elaborate, $\mathbf{Z}$ is a node which is deliberately inserted to connect all question and answer entities for context aggregation (similar to the *cls* token in Transformers), $\mathbf{Q}$ and $\mathbf{A}$ each denote the entities from the question and answer choice, respectively. $\mathbf{O}$ is the "other" type of node that does not fall in the three categories, but constitutes the extracted subgraph. We further input the feature translation (Bordes et al. 2013) of the head (h) and tail (t) nodes that comprise an edge $h \xrightarrow{r} t$. We compute the translation $\delta_{h,t}$ as $\delta_{h,t} = \boldsymbol{f}_t - \boldsymbol{f}_h$, where $\boldsymbol{f}_t, \boldsymbol{f}_h$ are the tail and head node features, respectively. Then, the embedding for an edge is computed as

$$\boldsymbol{h}_{(h,r,t)\in\mathcal{E}} = g_{\theta_1}([\phi(h), r, \phi(t), \delta_{h,t}]), \tag{1}$$

where $g_{\theta_1}(\cdot)$ is a MLP and $\phi(\cdot) : \mathcal{V} \rightarrow \mathcal{T}_v$ is a one-hot encoder that maps a node $v \in \mathcal{V}$ to a node type in $\mathcal{T}_v$.

In addition, we aim to diversify the embedding set by incorporating multi-hop paths. Specifically, we learn the embedding for each path ψ between a question node and an answer node; either one is the head or tail node. We consider paths of length n, and utilize a separate network for each n-hop path embeddings, where n-hop path ψ from h to t is featurized as $[\phi(h), r_1, \phi(v_1), \ldots, \phi(v_{n-1}), r_n, \phi(t), \delta_{h,t}]$. Then, for instance, the Meta-Path token of a 2-hop path is

$$\boldsymbol{h}_{\psi\in\boldsymbol{\Psi}_2} = g_{\theta_2}([\phi(h), r_1, \phi(v_1), r_2, \phi(t), \delta_{h,t}]), \tag{2}$$

where g_{θ_2} denotes the 2-hop path encoder and v_1 is an intermediate node. $\boldsymbol{\Psi}_2$ refers to the set of all 2-hop paths that connects h and t. Since an edge can be interpreted as a 1-hop path, we may concisely define the Meta-Path token set as

$$\mathcal{H}_{\text{KG}} = \bigcup_{k=1}^{K} \{\boldsymbol{h}_\psi | \psi \in \boldsymbol{\Psi}_k\}, \tag{3}$$

where we set K to 2 in our experiments.

Compared to QAT, many previous works have employed graph neural networks to encode the KG (Yasunaga et al. 2021; Wang et al. 2022; Lin et al. 2019). Generally adopting the message-passing mechanism, the neighboring nodes and edge features are aggregated by propagating a layer, and the *nodes* are finally pooled for prediction. Although the nodes consequently incorporate relational information, the objective of message-passing is to learn node features rather than relations, characterizing GNNs to be innately *node-centric*. On the other hand, our QAT explicitly encodes the relational information into Meta-Path tokens. By utilizing Meta-Path tokens for joint reasoning, the model takes advantage of the *relation-centric* tokens which contain information on the structural and semantic relations within the subgraph, see Section 4.3 for further discussions.

Drop-MP. Inspired by (Rong et al. 2019), we selectively apply a simple tactic of stochastically dropping meta-paths for regularization. Considering that each meta-path corresponds to one Meta-Path token, we simply drop a certain portion of Meta-Path tokens during training.

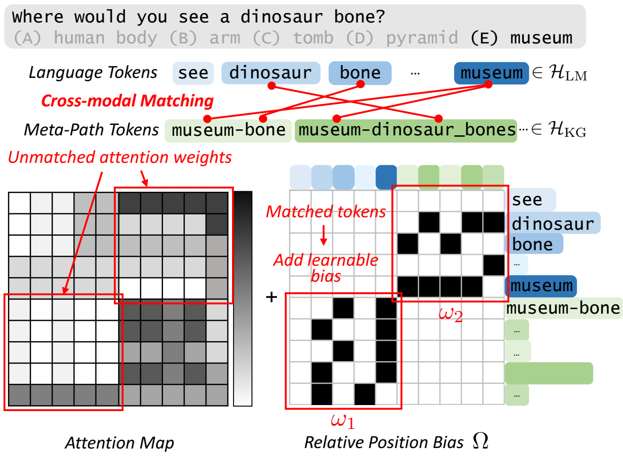

Figure 2: Illustration of Cross-Modal Relative Position Bias. Each Meta-Path token incorporates at least one head h and tail t node. The two nodes are each matched with an LM token, whose corresponding position is parameterized in our relative position bias Ω. Then, Ω is added to the attention map (Eq. (8)).

4.2 Relation-Aware Self-Attention

After retrieving the LM token embeddings $\mathcal{H}_{\text{LM}}$ and Meta-Path tokens $\mathcal{H}_{\text{KG}}$ (Eq. (3)), we now jointly reason over the concatenated token set to compute a logit value for each question and answer choice pair.

Language-Graph Joint Self-Attention. The multi-head self-attention module in the Transformer model is well known for its learning capacity and flexibility. By concatenating the two modalities and applying self-attention, each token aggregates features based on inter- and intra-modal relations. We also add learnable modality embeddings, $\boldsymbol{e}_{\text{LM}}$ and $\boldsymbol{e}_{\text{KG}}$, to the query and key tokens in order to encode the modality source into feature vectors. Then, the self-attention output $\hat{\mathcal{Z}}$ is computed by

$$\boldsymbol{X} = [\mathcal{H}_{\text{LM}}; \mathcal{H}_{\text{KG}}], \quad \tilde{\boldsymbol{X}} = [\mathcal{H}_{\text{LM}} + \boldsymbol{e}_{\text{LM}}; \mathcal{H}_{\text{KG}} + \boldsymbol{e}_{\text{KG}}] \tag{4}$$

$$\boldsymbol{Q} = \tilde{\boldsymbol{X}} W_{\boldsymbol{Q}}, \quad \boldsymbol{K} = \tilde{\boldsymbol{X}} W_{\boldsymbol{K}}, \quad \boldsymbol{V} = \boldsymbol{X} W_{\boldsymbol{V}} \tag{5}$$

$$\hat{\mathcal{Z}} = \text{Softmax}(\boldsymbol{Q}\boldsymbol{K}^\top / \sqrt{d})\ \boldsymbol{V}, \tag{6}$$

where $W_{\boldsymbol{Q}}$, $W_{\boldsymbol{K}}$, and $W_{\boldsymbol{V}}$ refers to the linear projection weights, and d is the key embedding dimension. Note, we described a single-headed case to avoid clutter.

Cross-Modal Relative Position Bias. Although the Transformer architecture alone is powerful in itself, providing guidance on learning *how* to aggregate information further enhances performance (Liu et al. 2021b; Yang et al. 2021; Wu et al. 2021). To encourage joint reasoning, we utilize a learnable relative position bias Ω that controls the cross-modal attention weights between relevant tokens.

The interrelation of a token pair from the two modalities is determined by whether the natural language representation of each entity is similar to each other. We utilize the pretrained GloVE (Pennington, Socher, and Manning 2014) word embeddings to vectorize each entity. In the case of KG, several node entities contain a series of words joined with underbars (*e.g.*, jewelry_store). For those entities, we take the average of the discretized word's GloVE embedding as its representation vector (*e.g.*, [GloVE("jewelry") + GloVE("store")] / 2). Then, for each relation token in $\mathcal{H}_{\text{KG}}$, we take its pertinent head and tail entity to compute their similarity with LM entities using cosine similarity, respectively. The head and tail nodes of a Meta-Path token are mapped to an LM token in $\mathcal{H}_{\text{LM}}$ by selecting the LM token with the highest similarity. That is, each Meta-Path token embedding is matched with two LM tokens. Based on this matching, our Cross-Modal Relative Position Bias (RPB) $\Omega \in \mathbf{R}^{|\boldsymbol{X}|\times|\boldsymbol{X}|}$ is defined. With abuse of notation, i, j refer to the i^{th}, j^{th} token:

Let

$$\mathbf{p}_{ij} = \mathbb{1}\left[i = \underset{i' \in \mathcal{H}_{\text{KG}}}{\text{argmax}} f(i'_v) \cdot f(j), \exists v \in \{h, t\}\right] \cdot \mathbb{1}\left[j \in \mathcal{H}_{\text{LM}}\right],$$

$$\mathbf{q}_{ij} = \mathbb{1}\left[i = \underset{i' \in \mathcal{H}_{\text{LM}}}{\text{argmax}} f(i') \cdot f(j_v), \exists v \in \{h, t\}\right] \cdot \mathbb{1}\left[j \in \mathcal{H}_{\text{KG}}\right].$$

Then,

$$\Omega[m, n] = \begin{cases} \omega_1 & \text{, if } \ \mathbf{p}_{mn} = 1, \mathbf{q}_{mn} = 0, \\ \omega_2 & \text{, if } \ \mathbf{p}_{mn} = 0, \mathbf{q}_{mn} = 1, \\ 0 & \text{, otherwise} \end{cases} \tag{7}$$

where $f(\cdot)$ is the GloVE-based embedding, i_v is the text representation of node v from token i, and $\mathbb{1}[\cdot]$ is the indicator function. ω_1, ω_2 are learnable scalar parameters, and m and n are token indices of $\boldsymbol{X}$. Also note that Ω is separately defined for each attention head, see Figure 2.

Then, our Relation-Aware Self-Attention (RASA) can be finally written as

$$\hat{\mathcal{Z}} = \text{Softmax}(\boldsymbol{Q}\boldsymbol{K}^\top / \sqrt{d} + \Omega)\ \boldsymbol{V}, \tag{8}$$

which is followed by an FFN layer that aggregates multiple heads. To enlist our transformer layer computation :

$$\begin{aligned} \hat{\mathcal{Z}} &= \text{RASA}(\text{LN}(\boldsymbol{X})) + \boldsymbol{X} \\ \mathcal{Z} &= \text{FFN}(\text{LN}(\hat{\mathcal{Z}})) + \hat{\mathcal{Z}}, \end{aligned} \tag{9}$$

where LN is the layer norm, and RASA refers to our Relation-Aware Self-Attention module (Eq. (8)).

Furthermore, to encourage a positive bias, we use a regularizer term in the final loss as

$$\mathcal{L} = \mathcal{L}_{\text{CE}} - \lambda \sum_{l=1}^{L} \sum_{h=1}^{H} \sigma(\omega^{(hl)}) \tag{10}$$

where $\mathcal{L}_{\text{CE}}$ is the cross entropy loss for question answering, and σ can be a nonlinear function (*e.g.*, logarithm, tanh). $\omega^{(hl)}$ refer to the relative position bias terms from the h^{th} attention head of layer l, and λ is a constant.

4.3 Comparisons with Existing Methods

In this section, we discuss QAT with GNN-based methods and RN-based methods for question answering problems.

GNN-based methods. GNNs learn node representations with the message-passing scheme, which recursively aggregates the representations of neighbors as follows (Gilmer et al. 2017; Kipf and Welling 2017; Yasunaga et al. 2021):

$$\boldsymbol{h}_i^{(l+1)} = \text{UPDATE}\left(\sum_{v_j \in \mathcal{N}_i} \alpha_{i,j}^{(l+1)} \boldsymbol{m}_{i,j}^{(l+1)}\right), \tag{11}$$

where $\boldsymbol{m}_{i,j}$ is the message that node v_j sends to v_i, $\alpha_{i,j}$ denotes the attention weight, and $\text{UPDATE}(\cdot)$ indicates a function that updates the node embedding by considering messages from neighborhoods.

Based on the learned node features, a representation for a graph $\mathcal{G}$ can be obtained with a set function f_{gnn} (*i.e.*, attentive pooling) as follows:

$$\text{GNN}(\mathcal{G}) = f_{\text{gnn}}\left(\left\{\boldsymbol{h}'_1, \boldsymbol{h}'_2, \ldots, \boldsymbol{h}'_n\right\}\right), \tag{12}$$

where $\boldsymbol{h}'_i$ is the embedding of node v_i from the final layer. Like Eq. (12), GNN-based models with message-passing perform reasoning at the node level; so, they are not readily available for path-based encoding and interpretation. On the other hand, QAT performs reasoning at the (meta-)path level, considering a graph as a set of relations. By tokenizing the meta-path relations, its reasoning ability is enhanced. We verify this effect by comparing the method by which the KG is encoded, in Section 6.

RN-based methods. The Relation Network (RN) (Santoro et al. 2017) is a neural network architecture that represents an object set $\mathcal{O}$ based on the relations between the objects within the set. A general form of RN is

$$\text{RN}(\mathcal{O}) = f\left(\left\{g\left(o_i, o_j\right) | o_i, o_j \in \mathcal{O}\right\}\right), \tag{13}$$

where function g encodes the relation of the pair of objects in $\mathcal{O}$, and f learns to represent the set of encoded relations. Considering that knowledge-based QA problems also entail representing the set of relations between concept entities, the QA problem can naturally be regarded an extension of RN. That is, given a pair of question and answer choice, a QA problem solver encodes the set of entity relations as written in (Feng et al. 2020) as

$$\text{RN}(\mathcal{G}) = f\left(\left\{g\left(v_i, r, v_j\right) | v_i \in \mathcal{V}_q, v_j \in \mathcal{V}_a, \left(v_i, r, v_j\right) \in \mathcal{E}\right\}\right), \tag{14}$$

where function g (*e.g.*, MLP) outputs the relations between question and answer entities, and function f (*e.g.*, pooling function) aggregates the relations. $\mathcal{V}_q$ and $\mathcal{V}_a$ refer to the entities from the question and answer choice, respectively.

Our QAT is a more flexible and powerful architecture that generalizes the relation network. Specifically, QAT can be formulated as follows:

$$\begin{aligned} &\text{QAT}\left(\mathcal{G}, \mathcal{H}^{\text{LM}}\right) \\ &= f(\{g\left(v_i, r_1, \ldots, r_k, v_j\right) | \left(v_i, r_1, \ldots, r_k, v_j\right) \in \boldsymbol{\Psi}_k, \\ &\quad k = 1, \ldots, K\} \cup \left\{\boldsymbol{h}_{cls}^{\text{LM}}, \boldsymbol{h}_1^{\text{LM}}, \ldots, \boldsymbol{h}_N^{\text{LM}}\right\}), \end{aligned} \tag{15}$$

where $\boldsymbol{\Psi}_k$ indicates a set of meta-paths with length k. In contrast to RN (Eq. (14)), our QAT captures the multi-hop relations (*i.e.*, meta-paths) beyond mere one-hop relations (*i.e.*, edges), via Meta-Path token embeddings. Moreover, QAT extends RN to integrate both the LMs and KGs with our relation-aware self-attention.

5 Experiments

5.1 Datasets

We evaluate our method on three question-answering datasets: CommonsenseQA (Talmor et al. 2019), OpenBookQA (Mihaylov et al. 2018), and MedQA-USMLE (Jin et al. 2021). Dataset statistics are in the supplement.

CommonsenseQA. This dataset contains questions that require commonsense reasoning. Since the official test set labels are not publicly available, we mainly report performance on the in-house development (IHdev) and test (IHtest) sets following (Lin et al. 2019). For CommonsenseQA, we adopt *ConceptNet* (Speer, Chin, and Havasi 2017) as the structured knowledge source.

OpenBookQA. This dataset requires reasoning with elementary science knowledge. Here, *ConceptNet* is again adopted as the knowledge graph, and we experiment on the official data split from (Mihaylov and Frank 2018).

MedQA-USMLE. This dataset originates from the United States Medical License Exam (USMLE) practice sets, requiring biomedical and clinical knowledge. Thus, we utilize a knowledge graph provided by (Jin et al. 2021). The graph is constructed via integrating the Disease Database portion of the Unified Medical Language System (Bodenreider 2004) and the DrugBank (Wishart et al. 2018) knowledge graph. We use the same data split as (Jin et al. 2021).

5.2 Baselines and Experimental Setup

We conduct experiments to compare our QAT with GNN-based methods and RN-based methods. For a fair comparison, all the baselines and our model use the same LMs in all experiments. The details on the backbone LMs and technical details are in the supplement.

5.3 Experimental Results

CommonsenseQA. Here, we report QAT and its baseline performances evaluated on CommonsenseQA in Table 1. Our QAT outperforms all the methods and achieved a new SOTA in-house test accuracy (IHtest-Acc.) of 75.4%. The proposed model provides a 6.7% performance boost compared to the vanilla RoBERTa language model (RoBERTa-L (w/o KG)) that does not utilize the KG. Specifically, the increase over GreaseLM and JointLK suggests that our QAT's Relation-Aware Self-Attention improves joint reasoning capabilities over the LM and KG. Also, the performance gaps compared to GSC, QA-GNN *etc.* indicates the superiority of our relation-centric Meta-Path tokens.

OpenBookQA. In Table 2, we report QAT performance on OpenBookQA. QAT with the RoBERTa-Large renders 71.2% mean test accuracy, which is a significant improvement over previous state-of-the-art by 0.9% and a boost of 6.4% compared to the vanilla RoBERTa-large LM. Following prior works (Yasunaga et al. 2021), we further experiment using AristoRoBERTa as the pre-trained LM, utilizing additional text data. By adopting AristoRoBERTa, we again

Methods	IHdev-Acc.(%)	IHtest-Acc.(%)
RoBERTa-L (w/o KG)	73.1 (±0.5)	68.7 (±0.6)
RGCN	72.7 (±0.2)	68.4 (±0.7)
GconAttn	72.6 (±0.4)	68.6 (±1.0)
KagNet	73.5 (±0.2)	69.0 (±0.8)
RN	74.6 (±0.9)	69.1 (±0.2)
MHGRN	74.5 (±0.1)	71.1 (±0.8)
QA-GNN	76.5 (±0.2)	73.4 (±0.9)
CoSe-CO	78.2 (±0.2)	72.9 (±0.3)
GreaseLM	78.5 (±0.5)	74.2 (±0.4)
JointLK	77.9 (±0.3)	74.4 (±0.8)
GSC	79.1 (±0.2)	74.5 (±0.4)
QAT (ours)	**79.5 (±0.4)**	**75.4 (±0.3)**

Table 1: Performance comparison on *CommonsenseQA*.

Methods	RoBERTa-Large	AristoRoBERTa
LMs (w/o KG)	64.8 (±2.4)	78.4 (±1.6)
RGCN	62.5 (±1.6)	74.6 (±2.5)
GconAttn	64.8 (±1.5)	71.8 (±1.2)
RN	65.2 (±1.2)	75.4 (±1.4)
MHGRN	66.9 (±1.2)	80.6 (±NA)
QA-GNN	67.8 (±2.8)	82.8 (±1.6)
GreaseLM	-	84.8 (±NA)
JointLK	70.3 (±0.8)	84.9 (±1.1)
GSC	70.3 (±0.8)	86.7 (±0.5)
QAT (ours)	**71.2 (±0.8)**	**86.9 (±0.2)**

Table 2: Test accuracy comparison on *OpenBookQA*.

achieved a state-of-the-art performance of 86.9% mean accuracy. This indicates that QAT successfully reasons over two different modalities and is adaptable to diverse LMs.

MedQA-USMLE. We additionally conduct an experiment on MedQA-USMLE (Jin et al. 2021) to evaluate the domain generality of our QAT in Table 3. The table shows that QAT achieves the state-of-the-art performance of 39.3% and improves performance over various medical LMs. This result indicates that our QAT is effective in reasoning in various domains beyond commonsense reasoning.

6 Analysis

In this section, we analyze our QAT to answer the following research questions: **[Q1]** Does each component in QAT boost performance? **[Q2]** Are relation-centric Meta-Path tokens really better than node-centric embeddings? **[Q3]** How does Relation-Aware Self-Attention utilize the language-graph relations when answering questions?

Ablation Studies. To answer the research question **[Q1]**, we conduct a series of ablation studies to verify the effect of each component of QAT. By omitting Drop-MP, all the Meta-Path tokens (MP tokens) are used in the training phase.

Methods	Accuracy
Chance	25.0
PMI	31.1
IR-ES	35.5
IR-Custom	36.1
ClinicalBERT-Base	32.4
BioRoBERTa-Base	36.1
BioBERT-Base	34.1
BioBERT-Large	36.7
SapBERT-Base (w/o KG)	37.2
QA-GNN	38.0
GreaseLM	38.5
QAT (ours)	**39.3**

Table 3: Test accuracy comparison on *MedQA-USMLE*.

MP tokens	RPB	Drop-MP	IHtest-Acc.(%)
✓	✓	✓	**75.4 (±0.3)**
✓	✓		75.3 (±0.7)
✓			75.0 (±0.5)
			73.8 (±0.4)

Table 4: Ablation study on *CommonsenseQA*.

Dataset	Node	Node+GNN	Meta-Path
CSQA	73.8 (±0.4)	73.9 (±0.2)	**75.4 (±0.3)**
OBQA	69.0 (±1.6)	69.6 (±0.9)	**71.2 (±0.8)**

Question Types	Node	Node+GNN	Meta-Path
Full question set	77.5	77.9	**79.8 (↑1.9)**
Question w/ negation	75.2	75.9	**79.0 (↑3.1)**
Question w/ entities ≤ 7	76.2	76.2	**79.9 (↑3.7)**
Question w/ entities > 7	79.1	79.5	**79.7 (↑0.2)**

Table 5: Meta-Path token and Node token Comparison. The test set performances with different KG tokens are compared (above). The development set performance on different types of questions are compared (below).

Ablating Cross-Modal RPB is equivalent to having no relative postion bias, and by removing MP tokens, we substitute the KG tokens with node features without additional processing. In Table 4, we present the ablation results where each component is removed one by one. By removing all our components, performance decreased significantly by 1.6%, and the biggest degradation was observed when the MP token was ablated. This proves the effectiveness of encoding the relation-centric information with Meta-Path tokens.

Meta-Path tokens vs. Node tokens. In order to address **[Q2]**, we elucidate the efficacy of Meta-Path tokens (**Meta-Path**) by comparing them with *node-centric* tokens. We extensively compare our *relation-centric* method with raw

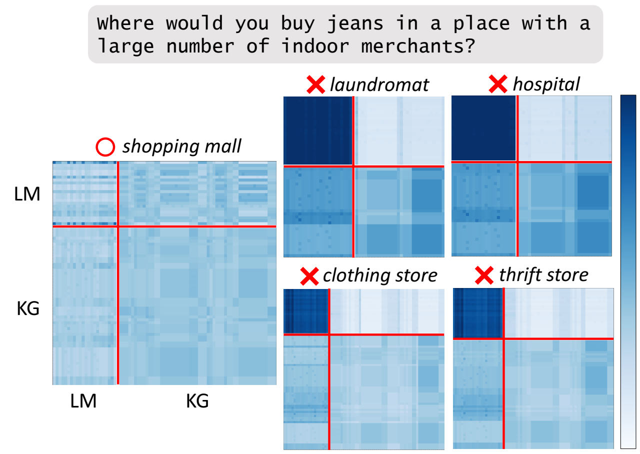

Figure 3: What and How QAT Selects. The attention maps for each question-answer pair is plotted. Cross-modal information is actively combined in the selected answer choice.

node-level tokens **(Node)** as well as the GNN-processed node tokens **(Node+GNN)**, whose GNN architecture is borrowed from (Yasunaga et al. 2021). Their performances are compared on CommonsenseQA and OpenBookQA in Table 5 (above). In all cases, we observed that the Meta-Path token consistently outperforms the node-centric tokens.

To better understand how each reacts to different question traits, we further experimented with diverse question types such as questions with negation, questions with fewer entities ($\leq$ 7), and questions with more entities (>7). In Table 5 (below), our Meta-Path tokens show clear dominance over the GNN-processed node tokens. Especially, it performs well in questions with negations by a significant margin of 3.1%, indicating a superiority in processing logically complex questions. Furthermore, outperformance in questions with less entities (an increase of 3.7%) implies that our method better leverages the KG to reason over external concepts selecting the correct answer.

Qualitative Analyses. We conduct qualitative analyses to provide intuition on how QAT understands the relationship between two modalities, thereby answering **[Q3]**. We first observe that cross-modal attention is crucial to distinguishing a correct answer from the others.

Our Relation-Aware Self-Attention (RASA) exhibits relatively stronger cross-modal attention, especially from KG to LM (right-upper quadrant), when a correct question-answer pair is given. Figure 3 shows that the attention map with the correct answer “shopping mall” is smooth and QAT integrates features from both modalities: LM and KG. On the other hand, for wrong answers (*e.g.*, laundromat, hospital, etc.) the cross-modal attention from KG to LM (right-upper quadrant) has relatively smaller attention weights and the model mainly focuses on the LM tokens (left-upper quadrant). These distinctive attention maps are obtained by our Meta-Path token generation. Note that QAT extracts a subgraph from KG for each question-answer pair and constructs a new set of Meta-Path tokens. Even with one-word answer

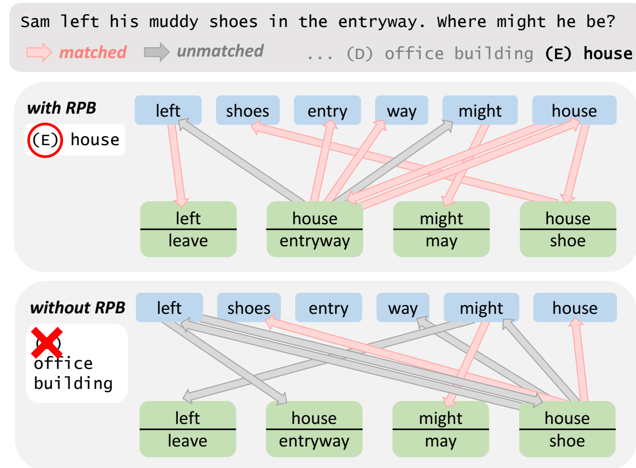

Figure 4: The Effect of Cross-Modal Relative Position Bias. The arrows indicate the attention direction, *e.g.*, ‘left–leave’ gets the most attention from ‘left’ with RPB, and the arrow colors signify whether the two tokens are matched. By removing our RPB, the attention mapping becomes uninterpretable, and leads to wrong answer selection.

choices, Meta-Path token generation results in more diversified KG token sets than node-level token generation.

We provide additional qualitative analysis to understand how our Cross-Modal RPB further enhances cross-modal attention. Figure 4 visualizes the attention mappings between LM and KG tokens. To avoid clutter, we only show the strongest attention between entities, and the arrows indicate the attention direction. For example, ‘left–leave’ gets the most attention from ‘left’, with RPB. The top blue tokens, *e.g.*, ‘left’ and ‘shoes’, are LM tokens, and bottom green tokens, *e.g.*, ‘left–leave’ and ‘house–entryway’, are Meta-Path tokens. Without our RPB (bottom row in Figure 4), we see more gray arrows indicating that cross-modal attentions between unmatched (irrelevant) tokens are prevalent. In opposition, the RPB augments the cross-modal attention by ω_1 $(LM \rightarrow KG)$ and ω_2 $(KG \rightarrow LM)$ for matched (relevant) tokens. Thus, matched tokens have higher attention and more red arrows are observed ‘with RPB’ on the top row. To be specific, (1) ‘house (LM)’ is selected by ‘house–entryway (KG)’ and ‘house–shoe (KG)’, (2) ‘house–entryway (KG)’ is selected by ‘entry (LM)’, ‘way (LM)’, and ‘house (LM)’ as an entity with the strongest attention.

7 Conclusion

We proposed Question Answering Transformer (QAT), which effectively combines LM and KG information via our Relation-Aware Self-Attention module equipped with Cross-Modal Relational Position bias. The key to state-of-the-art performance in various QA datasets such as CommonsenseQA, OpenbookQA, and MedQA-USMLE is our Meta-Path token, which learns *relation-centric* representations for the KG subgraph of a question-answer pair.

Acknowledgments

This work was partly supported by NAVER Corp. and MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program (IITP-2023-2020-0-01819) supervised by the IITP.

References

Bansal, R.; Aggarwal, M.; Bhatia, S.; Kaur, J. N.; and Krishnamurthy, B. 2022. CoSe-Co: Text Conditioned Generative CommonSense Contextualizer. In *NAACL*.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1): D267–D270.

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *NeurIPS*.

Feng, Y.; Chen, X.; Lin, B. Y.; Wang, P.; Yan, J.; and Ren, X. 2020. Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering. In *EMNLP*.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural Message Passing for Quantum Chemistry. In *ICML*.

Hwang, D.; Park, J.; Kwon, S.; Kim, K.; Ha, J.-W.; and Kim, H. J. 2020. Self-supervised Auxiliary Learning with Meta-paths for Heterogeneous Graphs. In *NeurIPS*.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14): 6421.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Lin, B. Y.; Chen, X.; Chen, J.; and Ren, X. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *EMNLP-IJCNLP*.

Liu, F.; Shareghi, E.; Meng, Z.; Basaldella, M.; and Collier, N. 2021a. Self-Alignment Pretraining for Biomedical Entity Representations. In *NAACL-HLT*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *ICCV*.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*.

Mihaylov, T.; and Frank, A. 2018. Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge. In *ACL*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*.

Rong, Y.; Huang, W.; Xu, T.; and Huang, J. 2019. DropEdge: Towards Deep Graph Convolutional Networks on Node Classification. In *ICLR*.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. In *NeurIPS*.

Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Berg, R. v. d.; Titov, I.; and Welling, M. 2018. Modeling Relational Data with Graph Convolutional Networks. In *ESWC*.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*.

Sun, Y.; Han, J.; Yan, X.; Yu, P. S.; and Wu, T. 2011. PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. *VLDB Endowment*, 4(11): 992–1003.

Sun, Y.; Shi, Q.; Qi, L.; and Zhang, Y. 2022. JointLK: Joint Reasoning with Language Models and Knowledge Graphs for Commonsense Question Answering. In *NAACL-HLT*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *ACL*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2018. Graph Attention Networks. In *ICLR*.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. In *CACM*.

Wang, K.; Zhang, Y.; Yang, D.; Song, L.; and Qin, T. 2022. GNN is a Counter? Revisiting GNN for Question Answering. In *ICLR*.

Wang, P.; Peng, N.; Ilievski, F.; Szekely, P.; and Ren, X. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *EMNLP-Findings*.

Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; and Yu, P. S. 2019. Heterogeneous Graph Attention Network. In *WWW*.

Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(Database-Issue): D1074–D1082.

Wu, K.; Peng, H.; Chen, M.; Fu, J.; and Chao, H. 2021. Rethinking and Improving Relative Position Encoding for Vision Transformer. In *ICCV*.

Yang, J.; Li, C.; Zhang, P.; Dai, X.; Xiao, B.; Yuan, L.; and Gao, J. 2021. Focal Attention for Long-Range Interactions in Vision Transformers. In *NeurIPS*.

Yasunaga, M.; Ren, H.; Bosselut, A.; Liang, P.; and Leskovec, J. 2021. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In *NAACL-HLT*.

Zhang, X.; Bosselut, A.; Yasunaga, M.; Ren, H.; Liang, P.; Manning, C. D.; and Leskovec, J. 2022. GreaseLM: Graph REASoning Enhanced Language Models for Question Answering. In *ICLR*.