論文紹介

QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering

Yasunaga, Michihiro, et al. "QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering." *arXiv preprint arXiv:2104.06378* (2021).

NAACL

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

M1, Zhu Yihua(シュ　ゲイカ)

チエック担当者: Momose Oyama (大山百々勢)

下平研究室ゼミ

2021.12.17

Contents

Introduction

QA task period

1, only large language models (LMs)
have broad coverage of knowledge but cannot structure reasoning.

2, only Knowledge graph (KG)
Can suit for structure reasoning but cannot use broad coverage knowledge

	LMs	KG
Broad coverage of Knowledge	●	✕
Structure Reasoning	✕	●

Then how to combine both sources (LM+KG)?

Combine these two sources have two challenges:

1, identify informative knowledge from a large KG.

2, capture the nuance of the QA context and the structure of KGs to perform joint reasoning over these two sources of information (will explain in the approach part)

Structure Reasoning

Enable the model be interpretable and explainable, and it is crucial for making robust predictions

Introduction

Previous works

Some papers separate QA context and KG. Then apply LMs to QA and apply GNNs to KG.

Drawback: cannot perform structured reasoning.

This paper

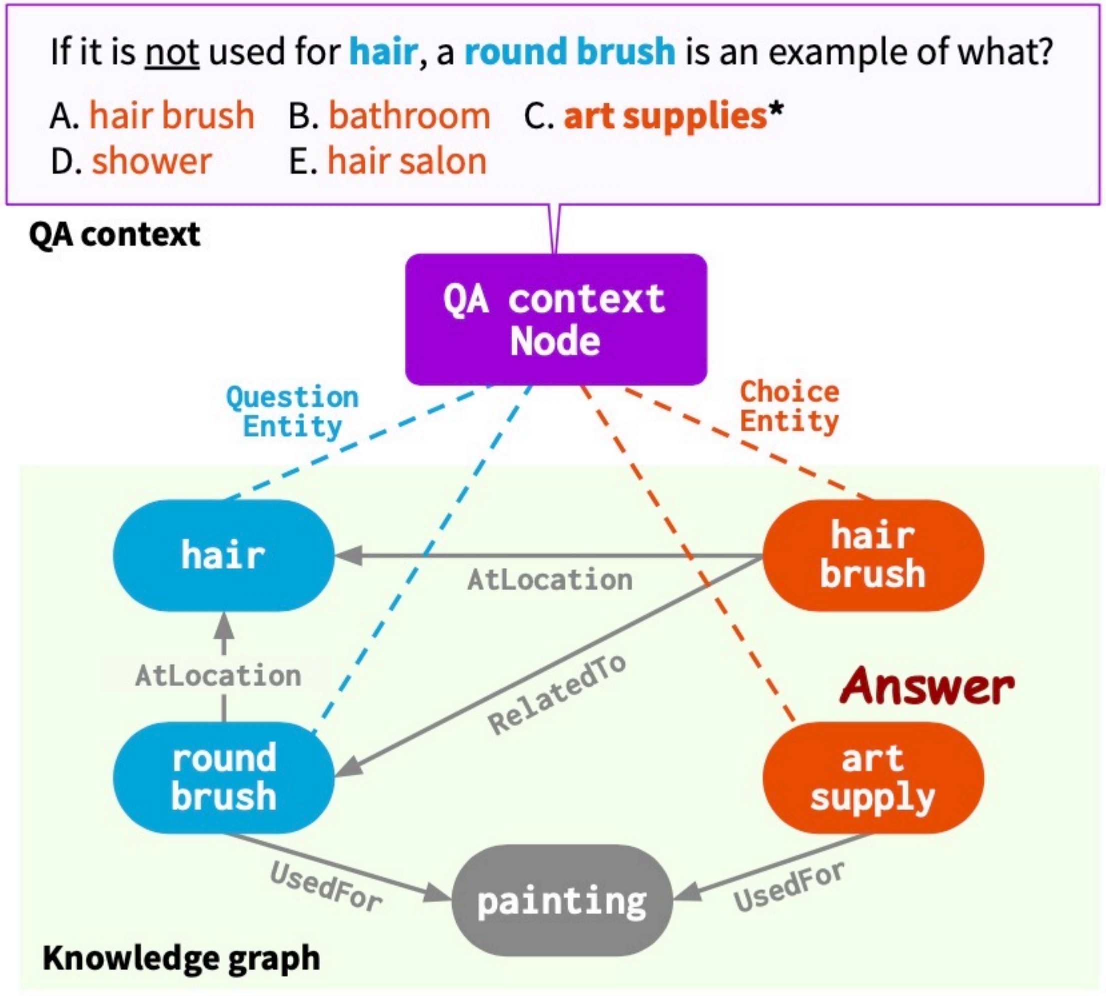
Novel methods applied in this paper:

1, relevance scoring. 2, joint reasoning.   (will explain in the conclusion part)

Experiment result can show:

1, improvements over existing LM and LM+KG models on question answering tasks
2, perform interpretable
3, structured reasoning

Definition



Multi-relation graph $G = (\mathcal{V}, \mathcal{E})$:

$\mathcal{V}$ represents the set of nodes in KG.
$\mathcal{E} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$ represents the edges that connect nodes.
$\mathcal{R}$ represents relation types.
$\mathcal{T}$ represents node types.

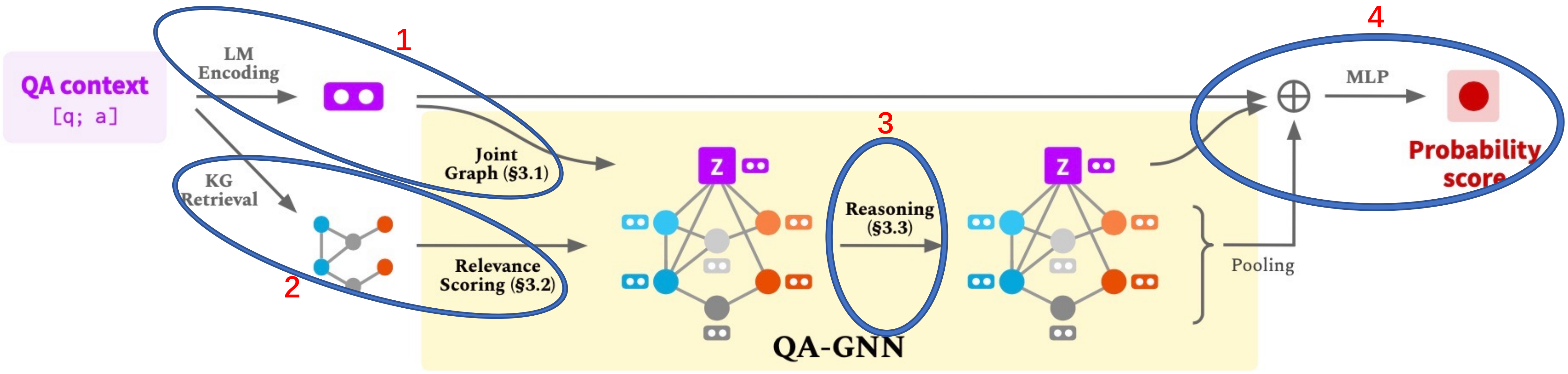
This graph (Working graph):

$G_W = (\mathcal{V}_W, \mathcal{E}_W)$
$\mathcal{V}_W = \mathcal{V}_{sub} \cup \{z\}$
$\mathcal{E}_W = \mathcal{E}_{sub} \cup \{z, r_{z,q}, \mathcal{V}_q | v \in \mathcal{V}_q\} \cup \{z, r_{z,a}, \mathcal{V}_a | v \in \mathcal{V}_a\}$

Edges: $r_{z,q}$: blue line
$r_{z,a}$: orange line
other: gray line

Vertex:
question nodes $\mathcal{V}_q$ (blue)
answer nodes $\mathcal{V}_a$ (orange)
context nodes Z (purple)
KG nodes $\mathcal{V}_{sub}$ (gray)

Approach: QA-GNN



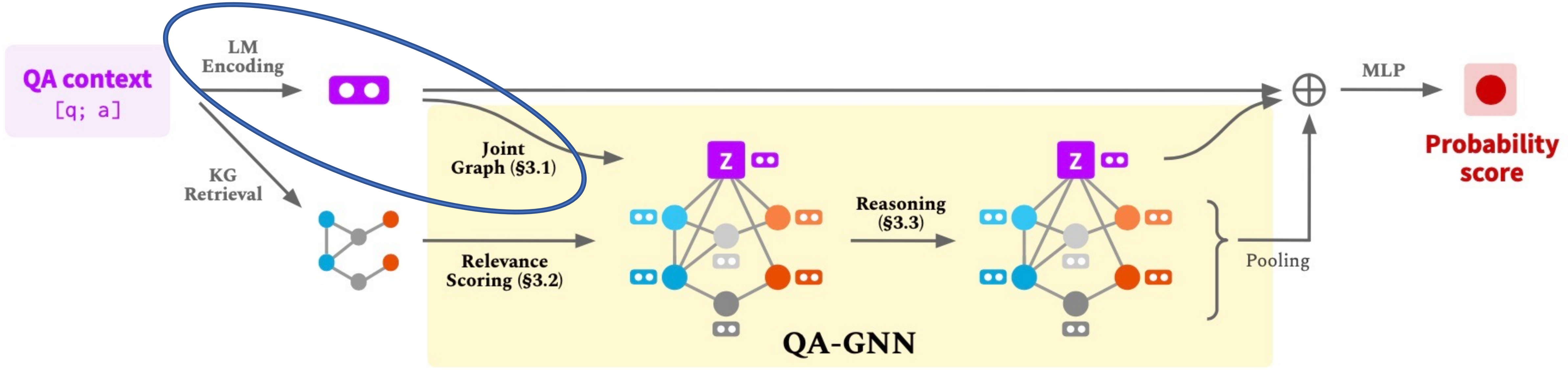
1, LM Encoding and Joint graph

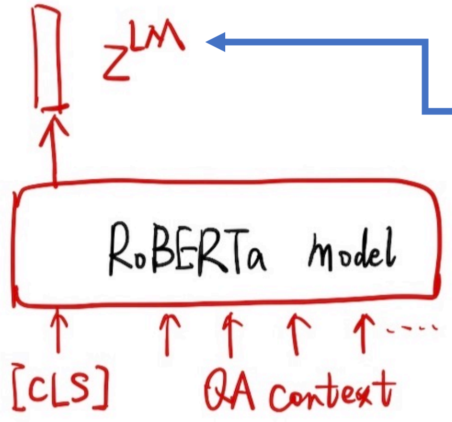
2, KG Retrieval and Relevance Scoring

3,Reasoning

4, Inference and Learning

Approach: QA-GNN (1, LM Encoding and Joint Graph)





LM Encoding:

$$z^{LM} = f_{enc}(text(z))$$

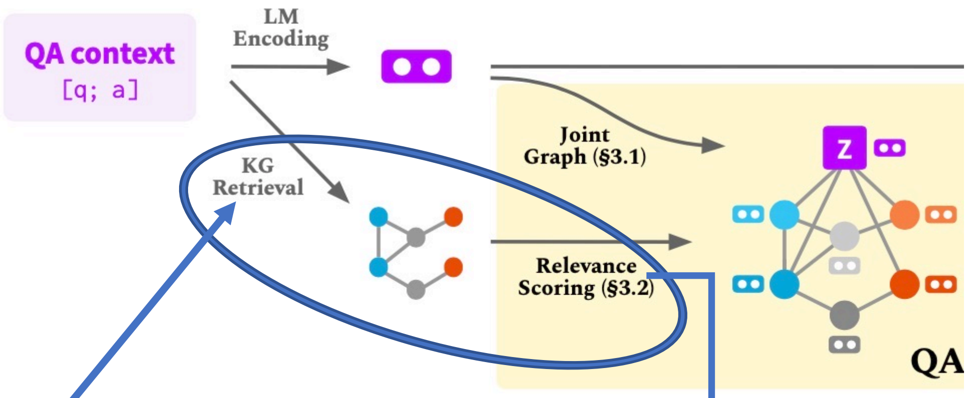
contextualized vector representation

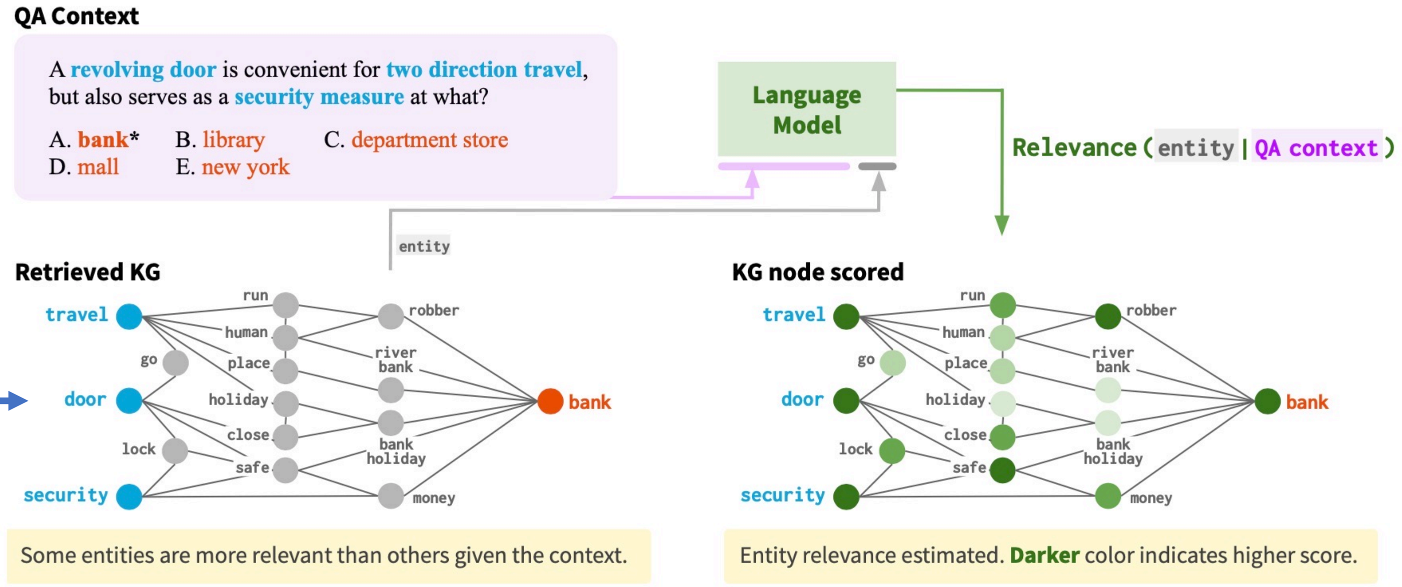
Encoder function: use masked language models (*e.g.*, RoBERTa) as f_{enc}

Joint Graph:

connect Z and KG

Approach: QA-GNN (2, KG Retrieval and Relevance Scoring)





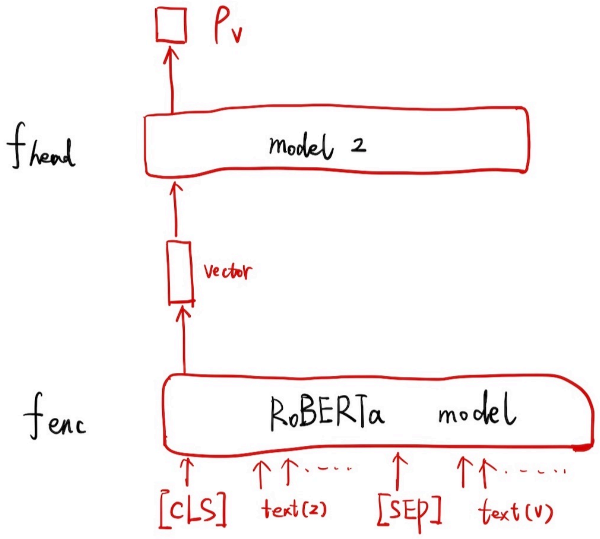
KG Retrieval:

1, choose the same entities appeared in QA context.

2, these entities with few-hop neighbors.

Approximate 200 nodes in a working graph.

Relevance Scoring:

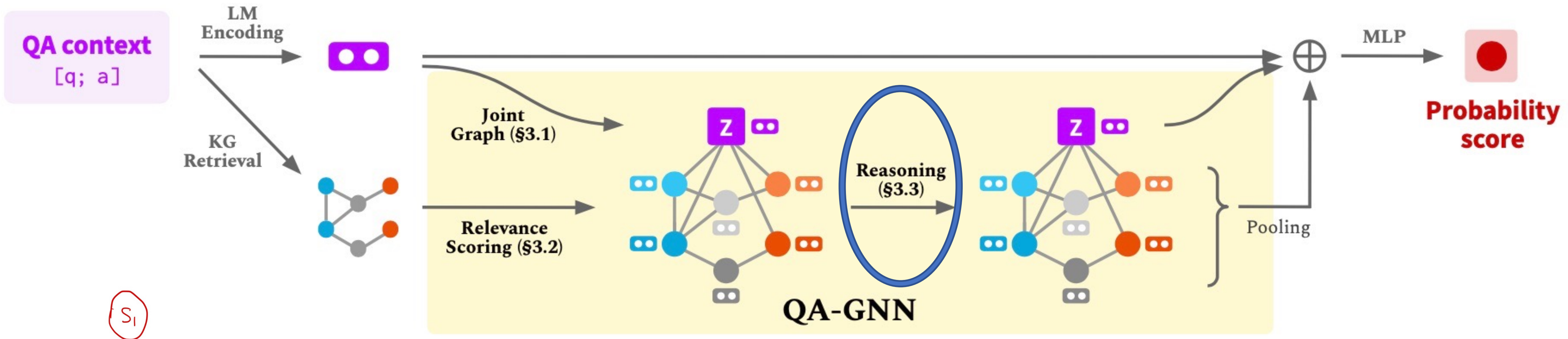


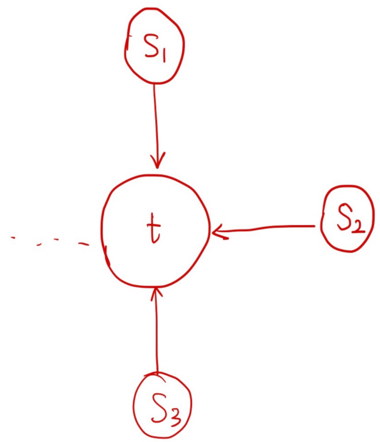
$$\rho_v = f_{head}(f_{enc}([text(z); text(v)]))$$

importance of each KG node relative to the given QA context

Calculate the relevance between entity and QA context by using the vector representation

Approach: QA-GNN (3, Reasoning (GNN architecture))





attention weight that scales each message m_{st} from s to t

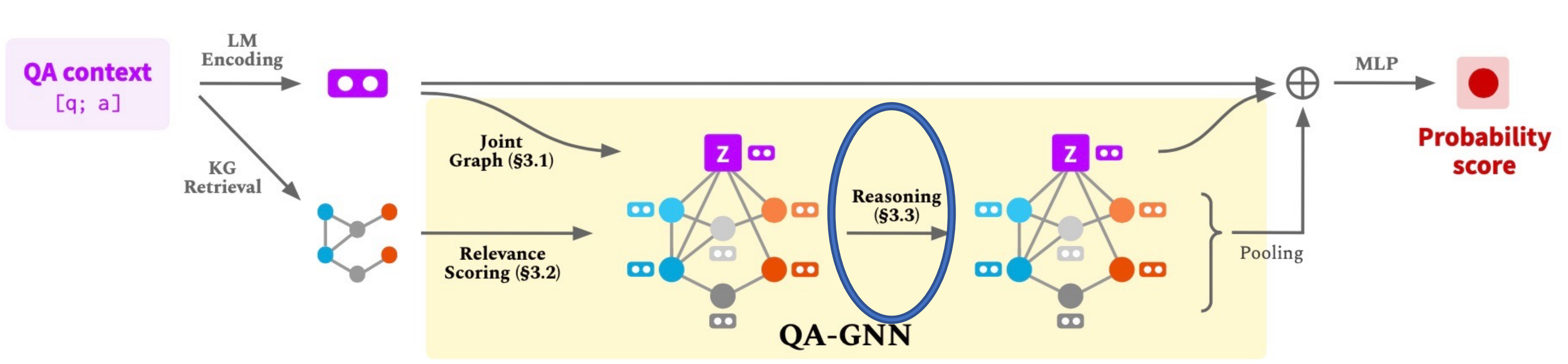
$$h_t^{\ell+1} = f_n\left(\sum_{s\in\mathcal{N}_t\cup\{t\}} \alpha_{st}\cdot m_{st}\right) + h_t^{\ell}$$

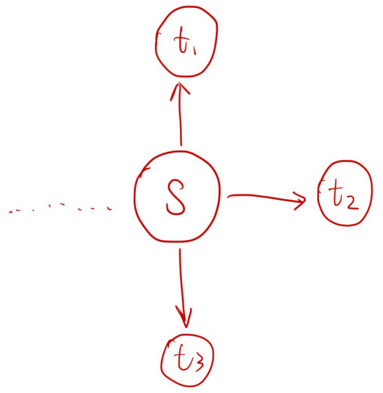
f_n: $\mathbb{R}^D \to \mathbb{R}^D$,2-layer MLP with Batch Normalization

expressive message, $m_{st} \in \mathbb{R}^D$: message from each neighbor node s to t

$\mathcal{N}_t$: the neighborhood of node t

Approach: QA-GNN (3, Reasoning (Node type and relation-aware message))





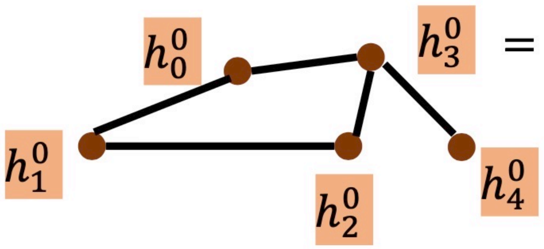
$$\mu_t = f_u(u_t),$$
$$\mu_s = f_u(u_s),$$
$$r_{st} = f_r(e_{st}, u_s, u_t)$$

$$m_{st} = f_m(h_s^{\ell}, \mu_s, r_{st})$$

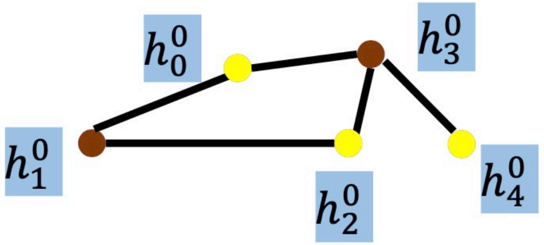
μ_t: type embedding
r_{st}:relation embedding
$u_t, u_s \in \{0,1\}^{|\mathcal{T}|}$:one-hot vectors indicates the node types.
$e_{st} \in \{0,1\}^{|\mathcal{R}|}$:one-hot vectors indicates the relation types.

f_u: $\mathbb{R}^{|\mathcal{T}|} \to \mathbb{R}^{D/2}$:linear transformation
f_r: $\mathbb{R}^{|\mathcal{R}|+2|\mathcal{T}|} \to \mathbb{R}^{D}$: 2-layer MLP
f_m: $\mathbb{R}^{2.5D} \to \mathbb{R}^{D}$: linear transformation

Approach: QA-GNN (3, Reasoning (Graph Attention Framework (GAT)))



$$h_3^0 = e_{3,0} \cdot h_0^0 + e_{3,2} \cdot h_2^0 + e_{3,4} \cdot h_4^0$$



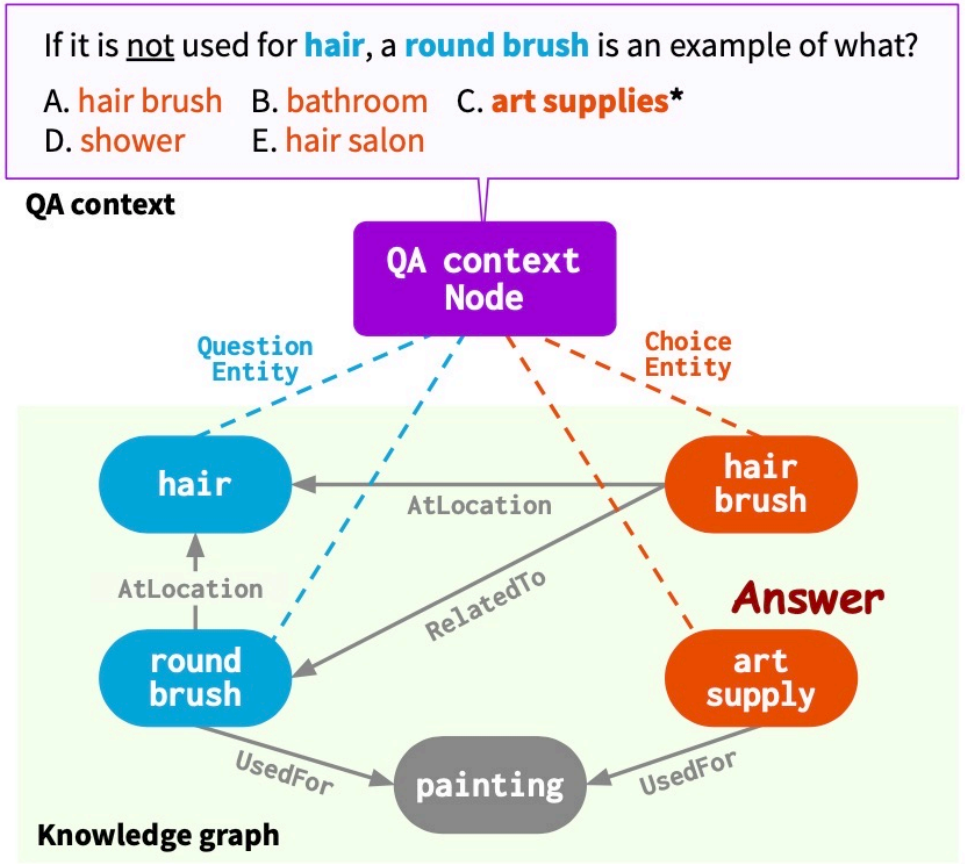
$$f(h_3^0, h_0^0) = e_{3,0}$$

$$f(h_3^0, h_2^0) = e_{3,2}$$

$$f(h_3^0, h_4^0) = e_{3,4}$$

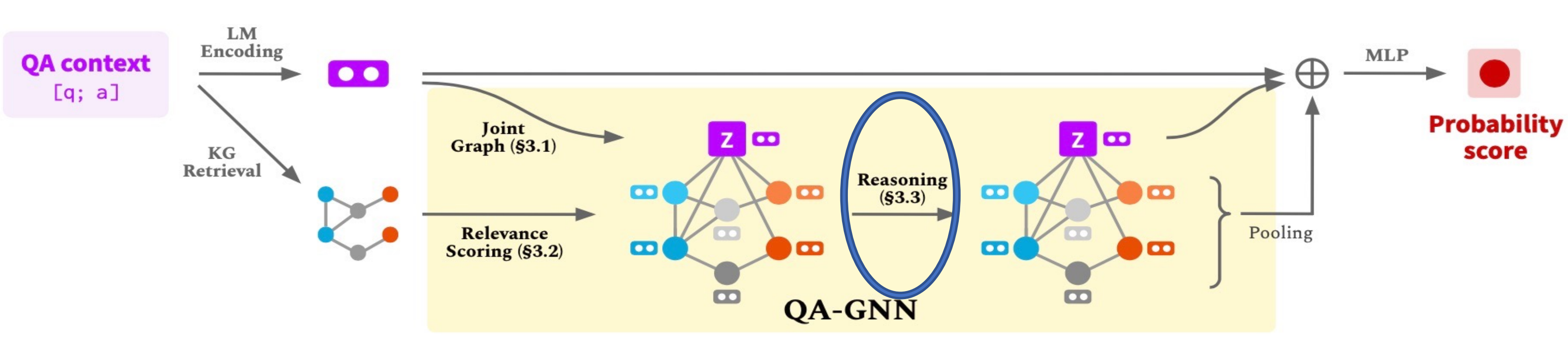
https://speech.ee.ntu.edu.tw/~hylee/ml/ml2021-course-data/cnn_v4.pptx

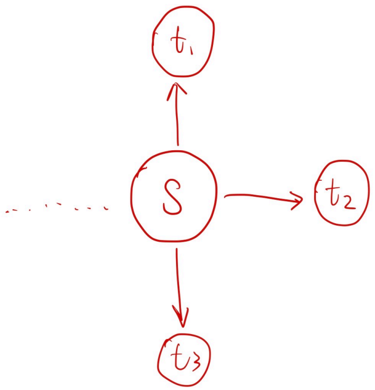
$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\vec{u}^{(l)T}[\vec{W}^{(l)}\mathbf{h}_i^{(l-1)} || \vec{W}^{(l)}\mathbf{h}_j^{(l-1)}]))}{\sum_{v_k \in N(v_i)} \exp(\text{LeakyReLU}(\vec{u}^{(l)T}[\vec{W}^{(l)}\mathbf{h}_i^{(l-1)} || \vec{W}^{(l)}\mathbf{h}_k^{(l-1)}]))}$$



Multiple relations, multiple nodes, directed edges

Approach: QA-GNN (3, Reasoning (Node type, relation, and score-aware attention))





1, embed the relevance score:

$$\boldsymbol{\rho}_t = f_\rho(\rho_t) \ ,$$
$$f_\rho : \mathbb{R}^1 \to \mathbb{R}^{D/2}\text{: MLP}$$

2, query vectors q :

$$q_s = f_q\left(h_s^\ell, \mu_s, \boldsymbol{\rho}_s\right),$$
$$f_q : \mathbb{R}^{2D} \to \mathbb{R}^{D} \text{ : Linear transformation}$$

3, key vectors k :

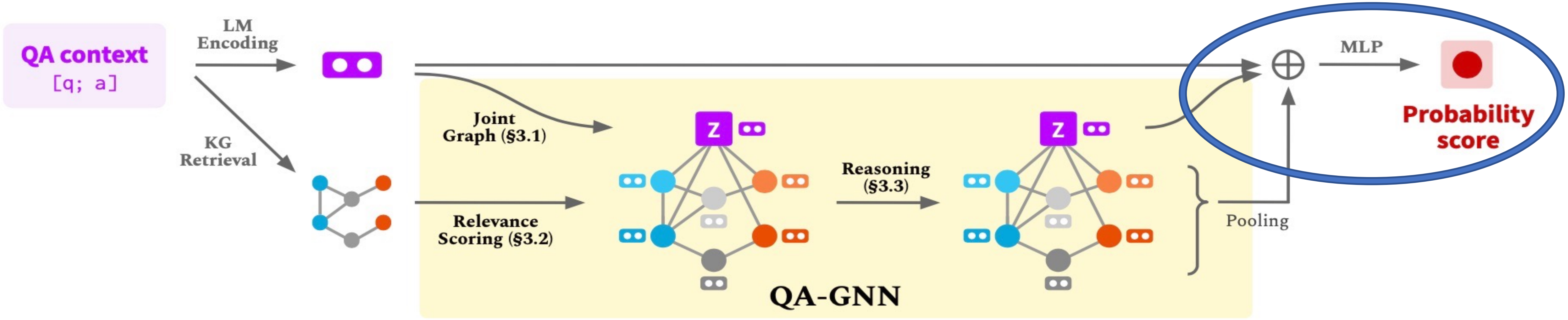
$$k_t = f_k(h_t^\ell, \mu_t, \boldsymbol{\rho}_t, r_{st}),$$
$$f_k : \mathbb{R}^{3D} \to \mathbb{R}^{D}\text{: Linear transformation}$$

4, attention α_{st}:

$$\gamma_{st} = \frac{{q_s}^T \cdot k_t}{\sqrt{D}}$$

$$\alpha_{st} = \frac{\exp(\gamma_{st})}{\sum_{t' \in \mathcal{N}_s \cup \{s\}} \exp(\gamma_{st'})}$$

Approach: QA-GNN (4, Inference and Learning)



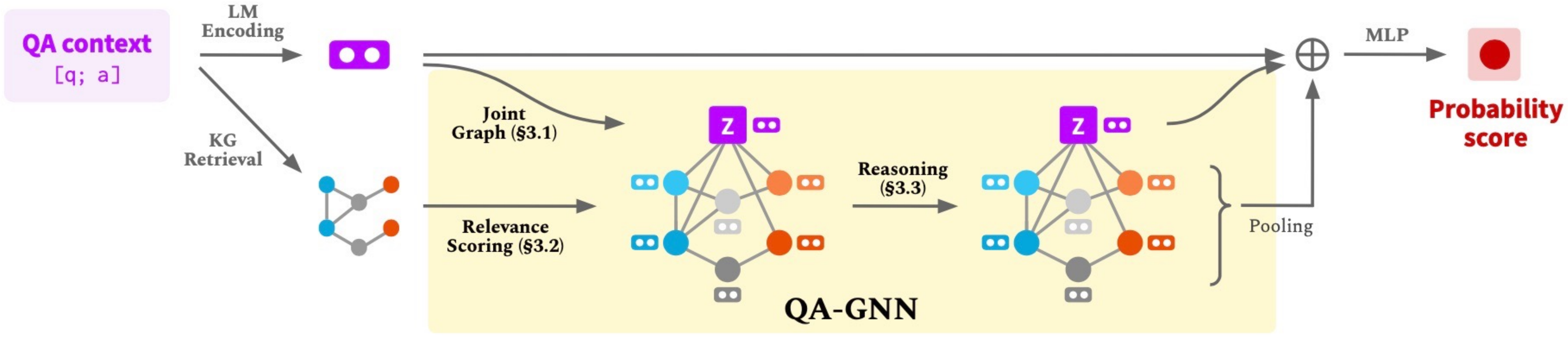
Given a question q and answer a, we can get the probability:

$$P(a|q) \propto \exp(MLP(z^{LM}, z^{GNN}, g))$$

Pooling of $\left\{h_v^{(L)} | v \in \mathcal{V}_{sub}\right\}$

Optimization: cross entropy loss

Approach: QA-GNN



1, LM Encoding and Joint graph

2, KG Retrieval and Relevance Scoring

3, Reasoning

4, Inference and Learning

This is the approach part, please free to ask me if you have questions?

Experiments (Dataset and Baselines)

QA Dataset:

1, CommonsenceQA
5-way multiple choice QA task that requires reasoning with commonsense knowledge, containing 12,102 questions.

https://github.com/jonathanherzig/commonsenseqa

2, OpenBookQA
4-way multiple choice QA task that requires reasoning with elementary science knowledge, containing 5,957 questions.

3, MedQA-USMLE
4-waymultiplechoiceQA task that requires biomedical and clinical knowledge, containing 12723 questions.

Train details:

1, dimension (D = 200)
2, number of layers (L = 5)
3, dropout rate = 0.2
4, RAdam optimizer
5, two GPUs

Knowledge Graph:

ConceptNet
general-domain knowledge graph
799,273 nodes and 2,487,810 edges in total

Baselines:

1, *Fine-tuned LM:*
RoBERTa-large for CommenseceQA
RoBERTa-large and AristoRoBERTa for OpenBookQA
SapBERT for MedQA-USMLE

2, *Existing LM+KG models:*
Relation Network(RN)
RGCN
GconAttn
KagNet
MHGRN (top performance)

Experiments (results: CommenseQA)

Methods	IHdev-Acc. (%)	IHtest-Acc. (%)
RoBERTa-large (w/o KG)	73.07 (±0.45)	68.69 (±0.56)
+ RGCN (Schlichtkrull et al., 2018)	72.69 (±0.19)	68.41 (±0.66)
+ GconAttn (Wang et al., 2019a)	72.61(±0.39)	68.59 (±0.96)
+ KagNet (Lin et al., 2019)	73.47 (±0.22)	69.01 (±0.76)
+ RN (Santoro et al., 2017)	74.57 (±0.91)	69.08 (±0.21)
+ MHGRN (Feng et al., 2020)	74.45 (±0.10)	71.11 (±0.81)
+ QA-GNN (**Ours**)	**76.54** (±0.21)	**73.41** (±0.92)

Table 2: **Performance comparison on *Commonsense QA* in-house split** (controlled experiments). As the official test is hidden, here we report the in-house Dev (IHdev) and Test (IHtest) accuracy, following the data split of Lin et al. (2019).

Methods	Test
RoBERTa (Liu et al., 2019)	72.1
RoBERTa+FreeLB (Zhu et al., 2020) (ensemble)	73.1
RoBERTa+HyKAS (Ma et al., 2019)	73.2
RoBERTa+KE (ensemble)	73.3
RoBERTa+KEDGN (ensemble)	74.4
XLNet+GraphReason (Lv et al., 2020)	75.3
RoBERTa+MHGRN (Feng et al., 2020)	75.4
Albert+PG (Wang et al., 2020b)	75.6
Albert (Lan et al., 2020) (ensemble)	76.5
UnifiedQA* (Khashabi et al., 2020)	**79.1**
RoBERTa + QA-GNN (**Ours**)	76.1

Table 3: **Test accuracy on *CommonsenseQA*'s official leaderboard**. The top system, UnifiedQA (11B parameters) is 30x larger than our model.

Experiments (results: OpenBookQA)

Methods	Test
Careful Selection (Banerjee et al., 2019)	72.0
AristoRoBERTa	77.8
KF + SIR (Banerjee and Baral, 2020)	80.0
AristoRoBERTa + PG (Wang et al., 2020b)	80.2
AristoRoBERTa + MHGRN (Feng et al., 2020)	80.6
Albert + KB	81.0
T5* (Raffel et al., 2020)	83.2
UnifiedQA* (Khashabi et al., 2020)	**87.2**
AristoRoBERTa + QA-GNN (**Ours**)	82.8

Table 5: **Test accuracy on *OpenBookQA* leaderboard.** All listed methods use the provided science facts as an additional input to the language context. The top 2 systems, UnifiedQA (11B params) and T5 (3B params) are 30x and 8x larger than our model.

Methods	RoBERTa-large	AristoRoBERTa
Fine-tuned LMs (w/o KG)	64.80 ($\pm$2.37)	78.40 ($\pm$1.64)
+ RGCN	62.45 ($\pm$1.57)	74.60 ($\pm$2.53)
+ GconAtten	64.75 ($\pm$1.48)	71.80 ($\pm$1.21)
+ RN	65.20 ($\pm$1.18)	75.35 ($\pm$1.39)
+ MHGRN	66.85 ($\pm$1.19)	80.6
+ QA-GNN (**Ours**)	**67.80** ($\pm$2.75)	**82.77** ($\pm$1.56)

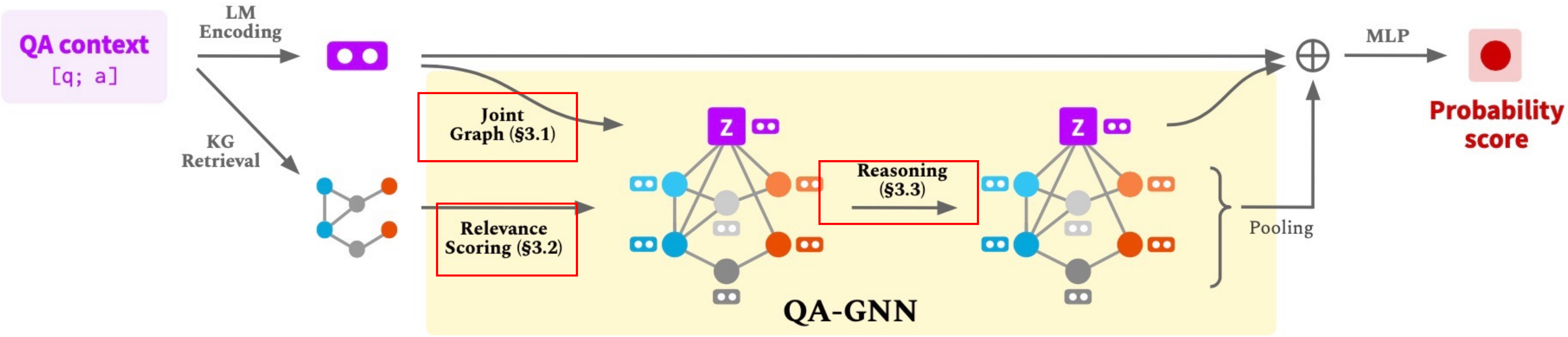
Table 4: **Test accuracy comparison on *OpenBook QA*** (controlled experiments). Methods with AristoRoBERTa use the textual evidence by Clark et al. (2019) as an additional input to the QA context.

Experiments (results: MedQA-USMLE)

Methods	Test
BERT-base (Devlin et al., 2019)	34.3
BioBERT-base (Lee et al., 2020)	34.1
RoBERTa-large (Liu et al., 2019)	35.0
BioBERT-large (Lee et al., 2020)	36.7
SapBERT (Liu et al., 2020a)	37.2
SapBERT + QA-GNN (**Ours**)	**38.0**

Table 6: **Test accuracy on *MedQA-USMLE*.**

Analysis (Ablation studies)



Graph Connection (§3.1)	Dev Acc.
No edge between Z and KG nodes	74.81
Connect Z to all KG nodes	76.38
Connect Z to QA entity nodes (**final**)	**76.54**

Relevance scoring (§3.2)	Dev Acc.
Nothing	75.56
w/ contextual embedding	76.31
w/ relevance score (**final**)	**76.54**
w/ both	76.52

GNN Attention & Message (§3.3)	Dev Acc.
Node type, relation, score-aware (**final**)	**76.54**
- type-aware	75.41
- relation-aware	75.61
- score-aware	75.56

GNN Layers (§3.3)	Dev Acc.
$L = 3$	75.53
$L = 4$	76.34
$L = 5$ (**final**)	**76.54**
$L = 6$	76.21
$L = 7$	75.96

Table 7: **Ablation study** of our model components, using the CommonsenseQA IHdev set.

Analysis (Model interpretability)

Can explain why this answer is right, how can we get this answer, and find more general reasoning structure

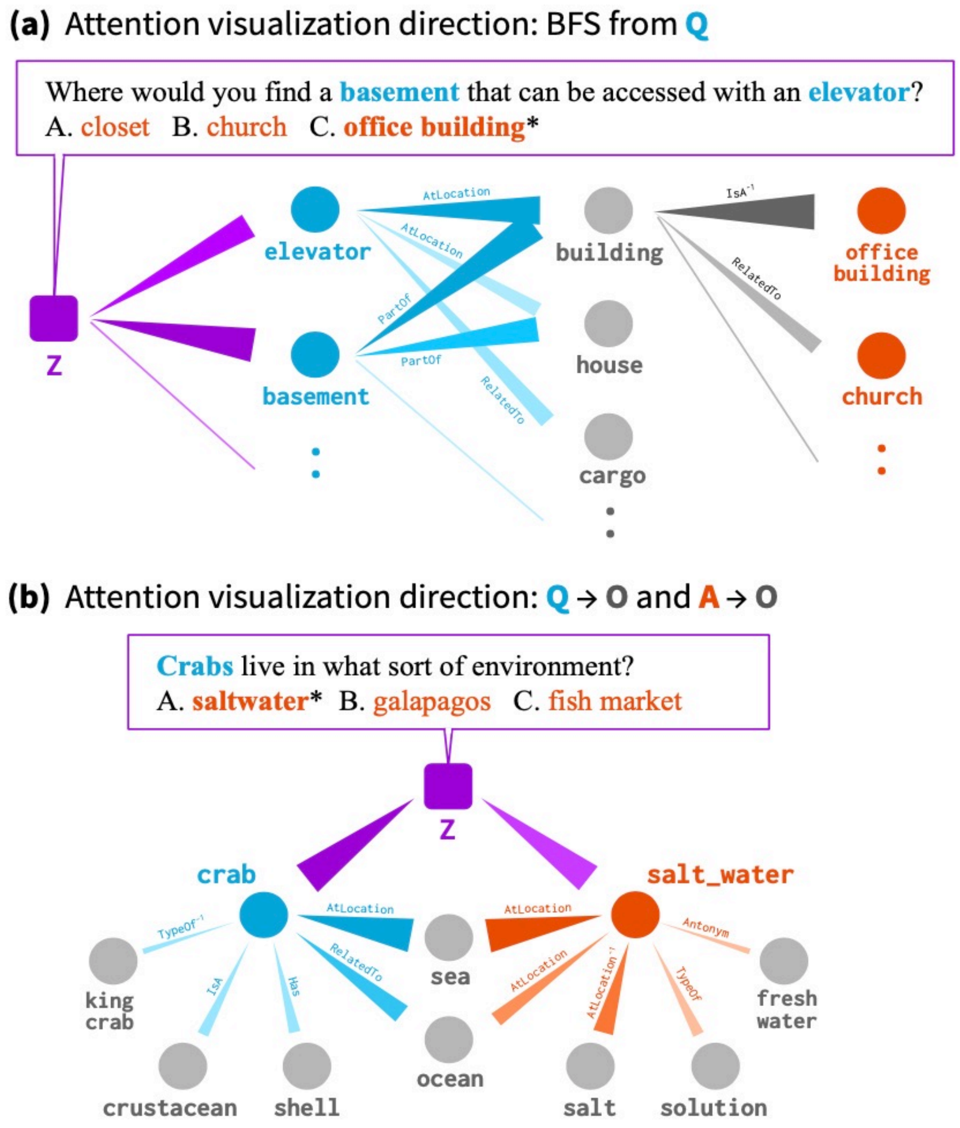


Figure 4: **Interpreting QA-GNN's reasoning process** by analyzing the node-to-node attention weights induced by the GNN. Darker and thicker edges indicate higher attention weights.

Analysis (Structure reasoning)

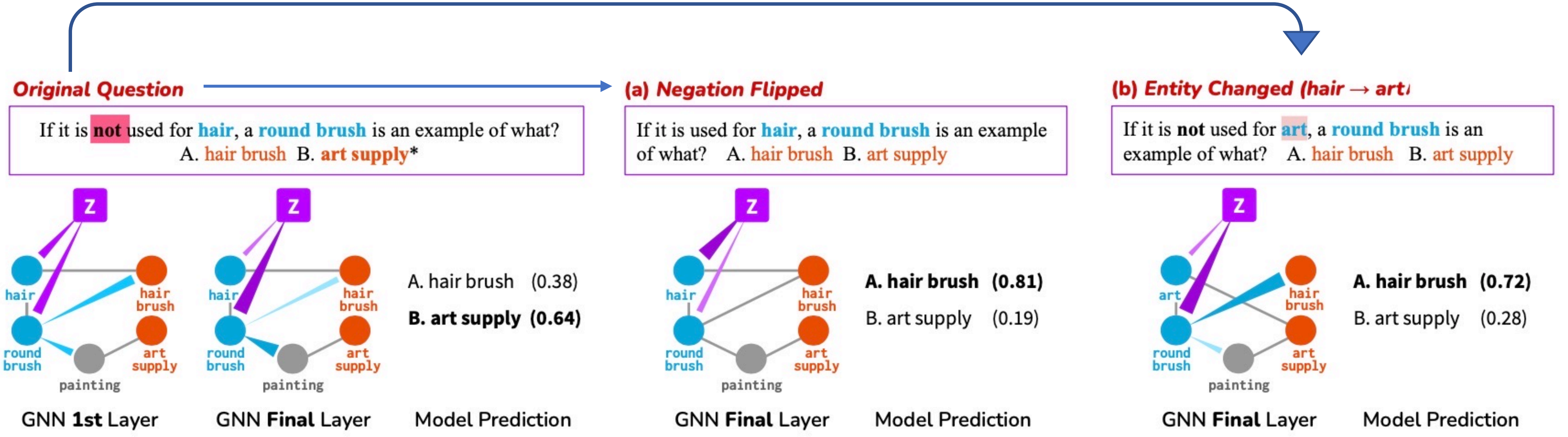


Figure 5: **Analysis of QA-GNN's behavior for structured reasoning**. Given an original question (left), we modify its negation (middle) or topic entity (right): we find that QA-GNN adapts attention weights and final predictions accordingly, suggesting its capability to handle structured reasoning.

Structure reasoning is crucial for making robust predictions

Analysis (Structure reasoning)

Example (Original taken from *CommonsenseQA* Dev)	RoBERTa Prediction	Our Prediction
[Original] If it is **not** used for hair, a round brush is an example of what? A. hair brush B. art supply	A. hair brush (✗)	B. art supply (✓)
[Negation flip] If it is used for hair, a round brush is an example of what?	A. hair brush (✓ just no change?)	A. hair brush (✓)
[Entity change] If it is not used for art a round brush is an example of what?	A. hair brush (✓ just no change?)	A. hair brush (✓)
[Original] If you have to read a book that is very dry you may become what? A. interested B. bored	B. bored (✓)	B. bored (✓)
[Negation ver 1] If you have to read a book that is very dry you may **not** become what?	B. bored (✗)	A. interested (✓)
[Negation ver 2] If you have to read a book that is **not** dry you may become what?	B. bored (✗)	A. interested (✓)
[Double negation] If you have to read a book that is **not** dry you may **not** become what?	B. bored (✓ just no change?)	A. interested (✗)

Table 8: **Case study of structured reasoning**, comparing predictions by RoBERTa and our model (RoBERTa + QA-GNN). Our model correctly handles changes in negation and topic entities.

From results:

QA-GNN adapts predictions to the modifications correctly, and can making robust predictions

Analysis (relevance scoring)

Methods	IHtest-Acc. (Question w/ ≤10 entities)	IHtest-Acc. (Question w/ >10 entities)
RoBERTa-large (w/o KG)	68.4	70.0
+ MHGRN	71.5	70.1
+ QA-GNN (w/o node relevance score)	72.8 (+1.3)	71.5 (+1.4)
+ QA-GNN (w/ node relevance score; **final system**)	73.4 (+1.9)	**73.5 (+3.4)**

Table 10: Performance on **questions with fewer/more entities** in *CommonsenseQA*. () shows the difference with MHGRN (LM+KG baseline). KG node relevance scoring (§3.2) boosts the performance on questions containing more entities (i.e. larger retrieved KG).

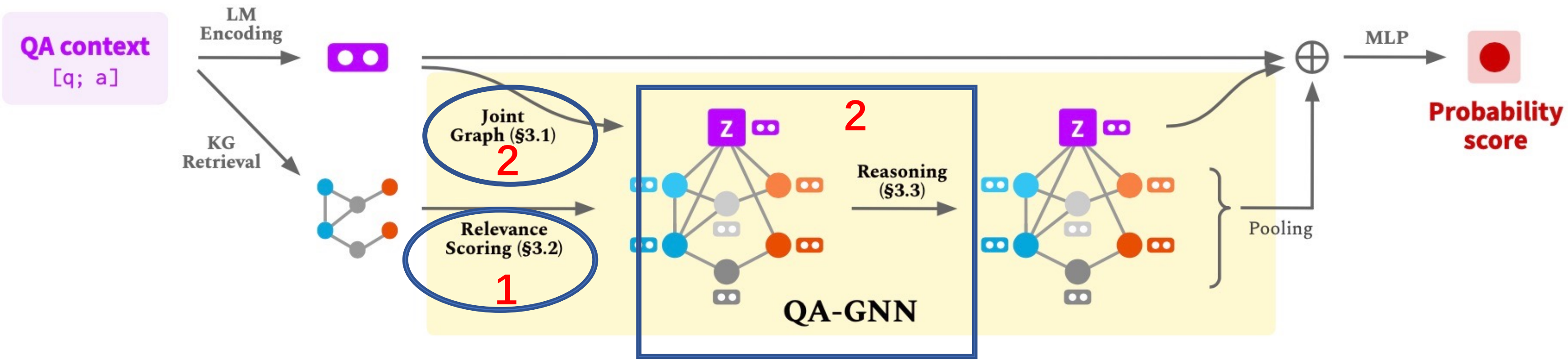
Existing LM+KG models such as MHGRN achieve limited performance on questions with more entities due to **the size and noisiness of retrieved KGs**

KG node relevance scoring is helpful when the retrieved KG is large

conclusion

Key innovations:

1, relevance scoring
2, Joint reasoning



Achievements:

1, improvements over existing LM and LM+KG models on question answering tasks
2, perform interpretable
3, structured reasoning

Reference

Yasunaga, Michihiro, et al. "QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering." *arXiv preprint arXiv:2104.06378* (2021).

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In Empirical Methods in Natural Language Processing (EMNLP).

Peter Clark, Oren Etzioni, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, et al. 2019. From'f'to'a'on the ny regents science exams: An overview of the aristo project. arXiv preprint arXiv:1909.01958.

Petar Velicˇkovic´, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio`, and Yoshua Bengio. 2018. Graph attention networks. In International Conference on Learning Representations (ICLR).