All-Hands Meeting II
2023 Summer

Relation-aware Language-Graph Transformer for Question Answering

Presenter: Minseo Yoon
(cooki0615@korea.ac.kr)

Introduction

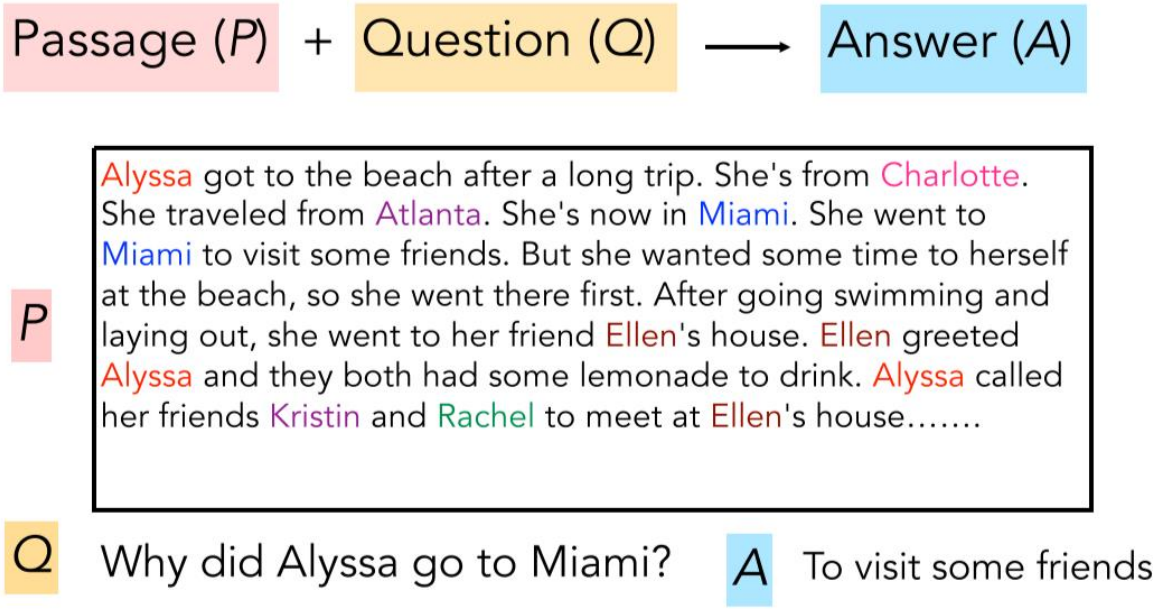
Relation-aware Language-Graph Transformer for Question Answering

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

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

Introduction

- Task
  - Natural Language Processing
  - Question Answering



Slide credit: Stanford University cs224n

Introduction

- Idea
  - Question Answering Transformer (QAT)
    - LM + KG
    - Without an explicit KG encoder, e.g., GNN
  - Meta-Path token
    - Encode KG information based on diverse relations
  - Cross-Modal Relative Position Bias
    - More flexible information exchange

Preliminaries

- Meta-Paths
  - $v_1 \xrightarrow{r_1} v_2 \xrightarrow{r_2} \ldots \xrightarrow{r_l} v_{l+1}, \quad r_l \in \mathcal{R}$
  - $\mathbf{r} = r_1 \circ r_2 \circ \ldots \circ r_l$
    - Define relationship between v_1 and v_{l+1}
  - Multi-hop relation encoding

Question Answering Transformer

- Meta-Path Token Embeddings
  - QAT learns embeddings
    - relationship between each node pair
  - edge type: one-hot vector
  - node type $\mathcal{T}_v = \{\mathbf{Z}, \mathbf{Q}, \mathbf{A}, \mathbf{O}\}$
    - Z: connect all Q and A (cls token), O: “other” type
  - Relation-centric tokens
    - Information on the structural and semantic relations
  - Small difference can induce a substantial change
    - Node-level: same token set except different token

Question Answering Transformer

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

translation $\delta_{h,t} = \boldsymbol{f}_t - \boldsymbol{f}_h$

- $g_{\theta_1}(\cdot)$: MLP $\quad \phi(\cdot) : \mathcal{V} \rightarrow \mathcal{T}_v$: one-hot encoder

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

- Meta-path Token set $\mathcal{H}_{\mathrm{KG}} = \bigcup_{k=1}^{K} \{\boldsymbol{h}_\psi | \psi \in \boldsymbol{\Psi}_k\}$
- Drop-MP: for regularization

Question Answering Transformer

- Relation-Aware Self-Attention
  - Language-Graph Joint Self-Attention

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

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

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

Question Answering Transformer

- Relation-Aware Self-Attention
  - Cross-Modal Relative Position Bias
  - Using Pretrained GloVE, compute similarity

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

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

  - $f(\cdot)$: GloVE-based Embedding

Question Answering Transformer

- Relation-Aware Self-Attention
  - Cross-Modal Relative Position Bias

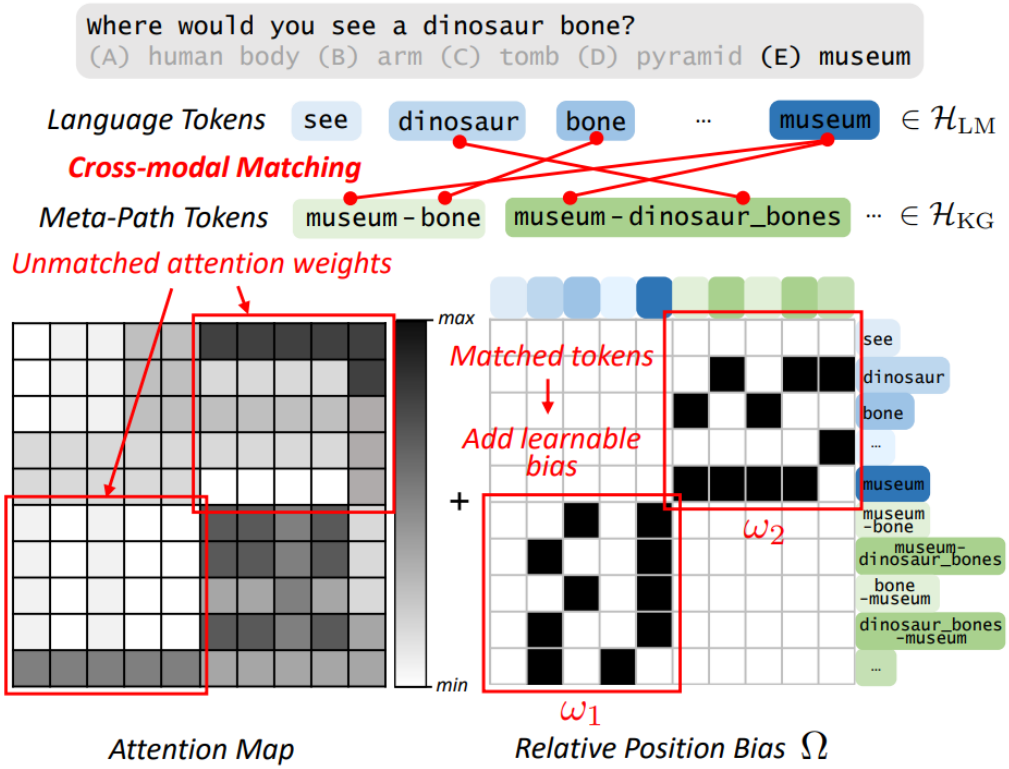
$$\Omega[m,n] = \begin{cases} \omega_1 & \text{, if } \mathbf{p}_{mn} = 1, \mathbf{q}_{mn} = 0, \\ \omega_2 & \text{, if } \mathbf{p}_{mn} = 0, \mathbf{q}_{mn} = 1, \\ 0 & \text{, otherwise} \end{cases}$$

  - m, n: token indices of $\boldsymbol{X} = [\mathcal{H}_{\mathrm{LM}}; \mathcal{H}_{\mathrm{KG}}]$
  - Ω is separately defined for each attention head

Question Answering Transformer

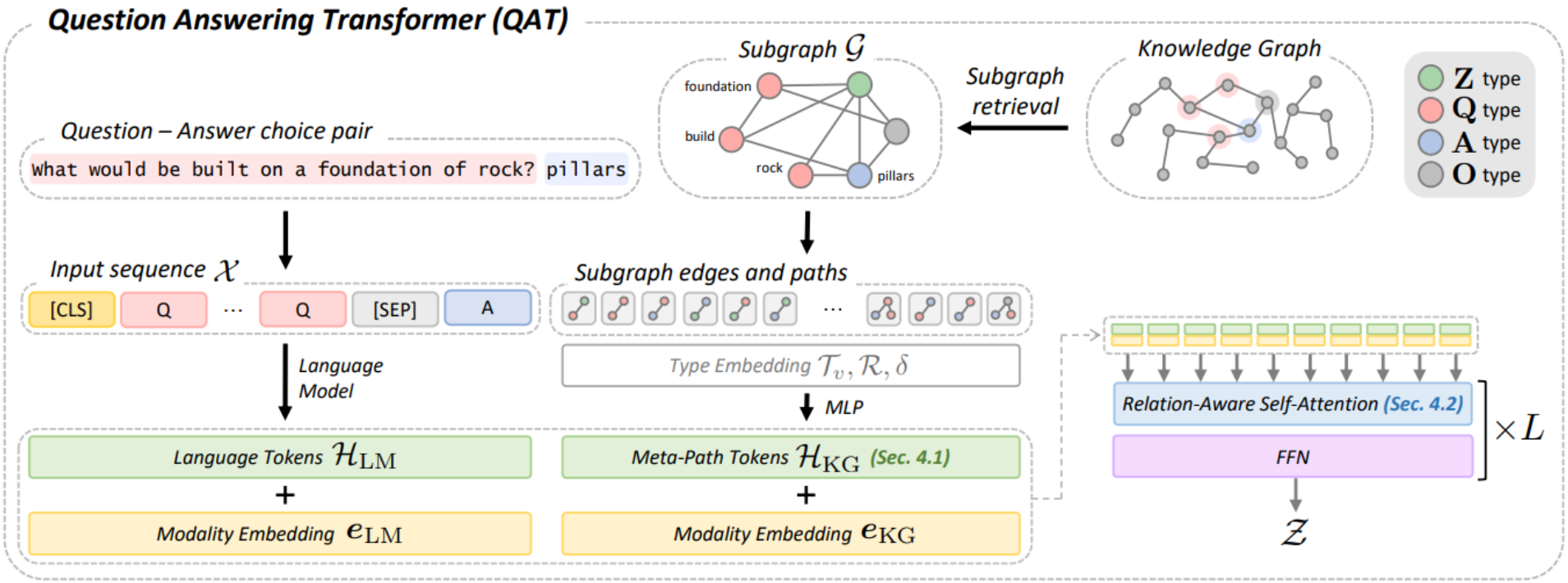
- Cross-Modal Relative Position Bias

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



Question Answering Transformer

- Structure



Loss: $\mathcal{L} = \mathcal{L}_{\mathrm{CE}} - \lambda \sum_{l=1}^{L} \sum_{h=1}^{H} \sigma(\omega^{(hl)})$

$$\hat{\mathcal{Z}} = \mathrm{RASA}(\mathrm{LN}(\boldsymbol{X})) + \boldsymbol{X}$$

$$\mathcal{Z} = \mathrm{FFN}(\mathrm{LN}(\hat{\mathcal{Z}})) + \hat{\mathcal{Z}}$$

Experiments

- Dataset
  - CommensenseQA: Commonsense reasoning
  - OpenbookQA: Elementary science knowledge
  - MedQA-USMLE: Biomedical and clinical knowledge
- Baselines
  - QAT with GNN-based methods and RN-based methods
  - All the baselines and QAT use the same LMs

Experiments

Methods	IHdev-Acc.(%)	IHtest-Acc.(%)
RoBERTa-Large (w/o KG)	73.1 (±0.5)	68.7 (±0.6)
RGCN (Schlichtkrull et al. 2018)	72.7 (±0.2)	68.4 (±0.7)
GconAttn (Wang et al. 2019b)	72.6 (±0.4)	68.6 (±1.0)
KagNet (Lin et al. 2019)	73.5 (±0.2)	69.0 (±0.8)
RN (Santoro et al. 2017)	74.6 (±0.9)	69.1 (±0.2)
MHGRN (Feng et al. 2020)	74.5 (±0.1)	71.1 (±0.8)
QA-GNN (Yasunaga et al. 2021)	76.5 (±0.2)	73.4 (±0.9)
CoSe-CO (Bansal et al. 2022)	78.2 (±0.2)	72.9 (±0.3)
GreaseLM (Zhang et al. 2022)	78.5 (±0.5)	74.2 (±0.4)
JointLK (Sun et al. 2022)	77.9 (±0.3)	74.4 (±0.8)
GSC (Wang et al. 2022)	79.1 (±0.2)	74.5 (±0.4)
QAT (ours)	**79.5 (±0.4)**	**75.4 (±0.3)**

Table 1: **Performance comparison on *CommonsenseQA*.**

- QAT's RASA improves joint reasoning capabilities over the LM and KG

Experiments

Methods	RoBERTa-Large	AristoRoBERTa
Fine-tuned LMs (w/o KG)	64.8 (±2.4)	78.4 (±1.6)
RGCN (Schlichtkrull et al. 2018)	62.5 (±1.6)	74.6 (±2.5)
GconAttn (Wang et al. 2019b)	64.8 (±1.5)	71.8 (±1.2)
RN (Santoro et al. 2017)	65.2 (±1.2)	75.4 (±1.4)
MHGRN (Feng et al. 2020)	66.9 (±1.2)	80.6 (±NA)
QA-GNN (Yasunaga et al. 2021)	67.8 (±2.8)	82.8 (±1.6)
GreaseLM (Zhang et al. 2022)	-	84.8 (±NA)
JointLK (Sun et al. 2022)	70.3 (±0.8)	84.9 (±1.1)
GSC (Wang et al. 2022)	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*.**

- QAT successfully reasons over two different modalities
- QAT is adaptable to diverse LMs

Experiments

Methods	Accuracy
Chance	25.0
PMI (Jin et al. 2021)	31.1
IR-ES (Jin et al. 2021)	35.5
IR-Custom (Jin et al. 2021)	36.1
ClinicalBERT-Base (Alsentzer et al. 2019)	32.4
BioRoBERTa-Base (Gururangan et al. 2020)	36.1
BioBERT-Base (Lee et al. 2020)	34.1
BioBERT-Large (Lee et al. 2020)	36.7
SapBERT-Base (w/o KG) (Liu et al. 2021a)	37.2
QA-GNN (Yasunaga et al. 2021)	38.0
GreaseLM (Zhang et al. 2022)	38.5
QAT (ours)	**39.3**

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

- QAT is effective in reasoning in various domains beyond commonsense reasoning

Analysis

- [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?

Analysis

- Ablation Studies

Meta-Path tokens	Cross-Modal 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 (RoBERTa-Large)	69.0 (±1.6)	69.6 (±0.9)	**71.2 (±0.8)**
OBQA (AristoRoBERTa)	85.6 (±1.2)	86.3 (±0.8)	**86.9 (±0.2)**

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.** (above) The test set performances with different KG tokens are compared. (below) The development set performance on different types of questions are compared.

Analysis

- Qualitative Analyses

- Attention map with the correct answer is smooth
- Obtained by Meta-Path token generation

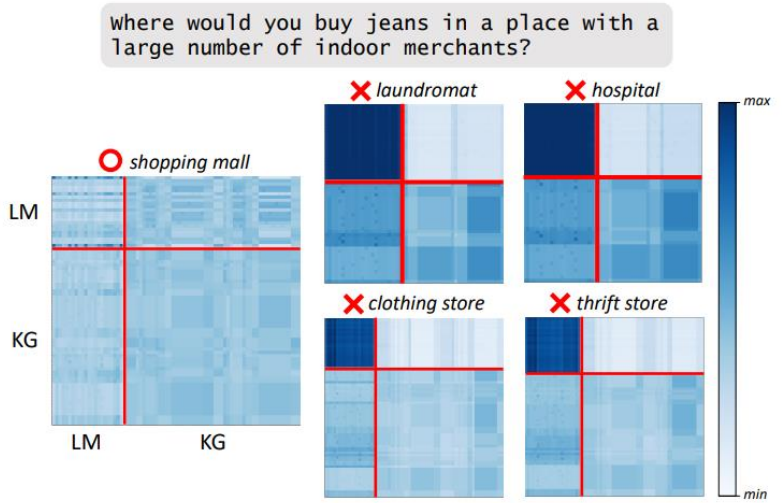


Figure 3: **What and How QAT Selects.** The attention maps for each question-answer pair is plotted. Cross-modal information is actively combined in the selected answer choice.

Analysis

- Qualitative Analyses

- Top: LM tokens
- Bottom: Meta-Path tokens
- RPB augments the cross-modal attention by $\Omega[m,n] = \begin{cases} \omega_1 & \text{, if } \mathbf{p}_{mn} = 1, \mathbf{q}_{mn} = 0, \\ \omega_2 & \text{, if } \mathbf{p}_{mn} = 0, \mathbf{q}_{mn} = 1, \\ 0 & \text{, otherwise} \end{cases}$
- Higher attention and more red arrows are observed 'with RPB'

(1) 'house (LM)' is selected
(2) 'house-entryway (KG)' is selected
(3) 'house (LM)' is an entity with the strongest attention

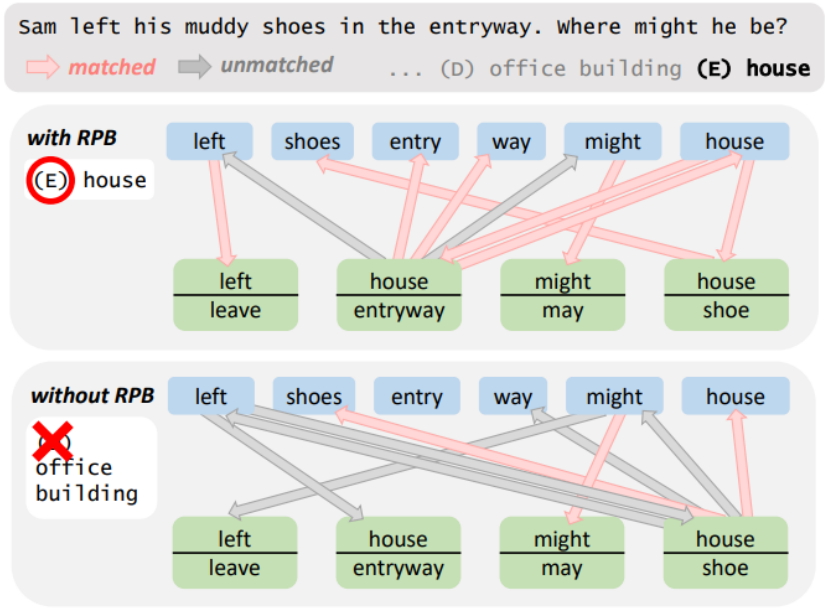


Figure 4: **The Effect of Cross-Modal Relative Position Bias.** The arrows indicate the attention direction, *e.g.*, 'left–leave' gets the most attention from 'left' with RPB, and the arrow colors signify whether the two tokens are matched. By removing our RPB, the attention mapping becomes uninterpretable,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

Reimplementation Result

- CommonsenseQA
  - Dev-Acc: 78.1%
  - Test-Acc: 75.6%
  - 20th epoch
- Colab Pro V100
- 23 epochs / 4 hours

Methods	IHdev-Acc.(%)	IHtest-Acc.(%)
RoBERTa-Large (w/o KG)	73.1 (±0.5)	68.7 (±0.6)
RGCN (Schlichtkrull et al. 2018)	72.7 (±0.2)	68.4 (±0.7)
GconAttn (Wang et al. 2019b)	72.6 (±0.4)	68.6 (±1.0)
KagNet (Lin et al. 2019)	73.5 (±0.2)	69.0 (±0.8)
RN (Santoro et al. 2017)	74.6 (±0.9)	69.1 (±0.2)
MHGRN (Feng et al. 2020)	74.5 (±0.1)	71.1 (±0.8)
QA-GNN (Yasunaga et al. 2021)	76.5 (±0.2)	73.4 (±0.9)
CoSe-CO (Bansal et al. 2022)	78.2 (±0.2)	72.9 (±0.3)
GreaseLM (Zhang et al. 2022)	78.5 (±0.5)	74.2 (±0.4)
JointLK (Sun et al. 2022)	77.9 (±0.3)	74.4 (±0.8)
GSC (Wang et al. 2022)	79.1 (±0.2)	74.5 (±0.4)
QAT (ours)	**79.5 (±0.4)**	**75.4 (±0.3)**

Table 1: **Performance comparison on *CommonsenseQA*.**

Proposed Experiments 1 (If I afford)

- Relative Position Bias
- Another Inductive Bias
  - Attention Guidance

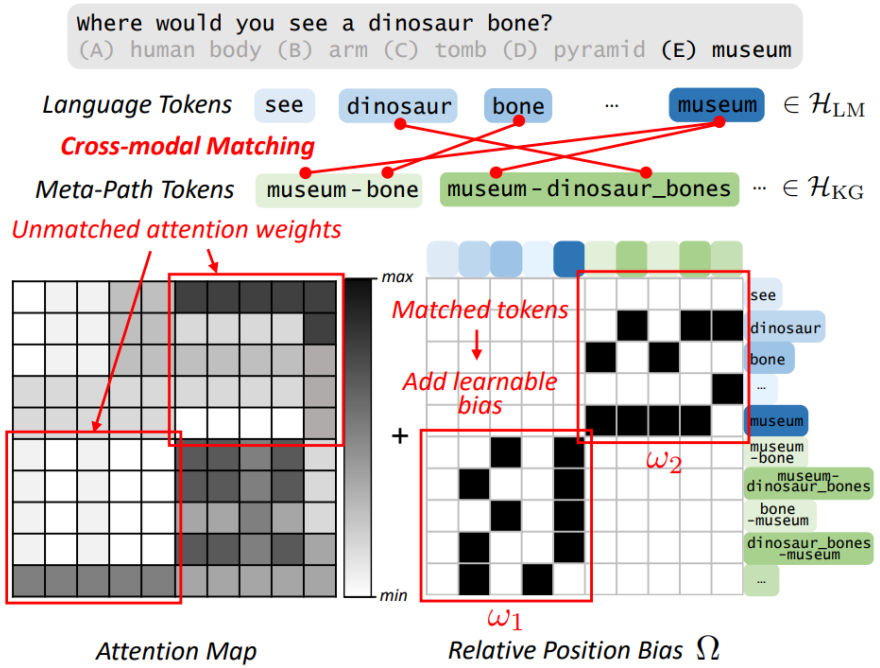
1. Guidance on the relation that helps to come up often as the answer
2. Guidance for relation about the most common correct answer to the wrong answered question

→ 6 Additional Experiment

① With 1 ② 2 ③ 1 + 2

④ 1 + RPB ⑤ 2 + RPB

⑥ 1 + 2 + RPB



Proposed Experiments 2 (Main Idea)

- Language Model
  - RoBERTa → GPT
- Answer Selection → Answer Generation

```
Where would you see a dinosaur bone?
(A) human body (B) arm (C) tomb (D) pyramid (E) museum
```

Where would you see a dinosaur bone?
(A) human body (B) arm (C) tomb (D) pyramid (E) museum

You would most likely see a dinosaur bone in a (E) museum. Museums often have fossil displays and exhibitions, including dinosaur bones, where visitors can learn about prehistoric creatures and their history. Fossilized dinosaur bones are significant artifacts that provide valuable insights into the Earth's past and the evolution of life on our planet.

Questions?