Lab Seminar

GreaseLM

(Paper Review)

HTET ARKAR
Junior, Undergraduate
School of Computer Science and Engineering
Chung-Ang University

Graph REASoning Enhanced Language Models for Question Answering

(GreaseLM)

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren,
Percy Liang, Christopher D. Manning, Jure Leskovec
Stanford University

ICLR 2022

Contents

- Introduction
- Previous Works
- Proposed Method
- Experiments
- Future Works
- Conclusion

Introduction

Question Answering

- A challenging task that requires complex reasoning over both
  - Explicit constraints described in the textual context of the question
  - Unstated, relevant knowledge about the world (knowledge about the domain interest)

Complex Reasoning (Zhang, Xikun, et al.)

- The number of prepositional phrases in the questions
  - To easier burrow, after prey, in a what
- The presence of negation terms
  - e.g. no, never
- The presence of hedging terms
  - Terms indicating uncertainty (e.g. sometimes; maybe)

CommonsenseQA	A weasel has a thin body and short legs to easier burrow after prey in a what? (A) tree (B) mulberry bush (C) chicken coop (D) viking ship **(E) rabbit warren**
OpenbookQA	Which of these would let the most heat travel through? (A) a new pair of jeans **(B) a steel spoon in a cafeteria** (C) a cotton candy at a store (D) a calvin klein cotton hat

If it is not used for **hair**, a **round brush** is an example of what?

A. hair brush B. bathroom C. **art supplies***
D. shower E. hair salon

QA context

Foundation of Most Modern QA Systems

Large Pretrained Language Models

- Fine-tuned on QA datasets for Question Answering Task
- After pre-training, learn to implicitly encode broad knowledge about the world
  - Able to leverage when fine-tuned on a domain-specific downstream task
- Struggle when given examples that are distributionally different from examples seen during fine-tuned
  - Learned behavior often relies on simple patterns to offer shortcuts to an answer
- Not perform well on **structured reasoning** (e.g. handling negation)

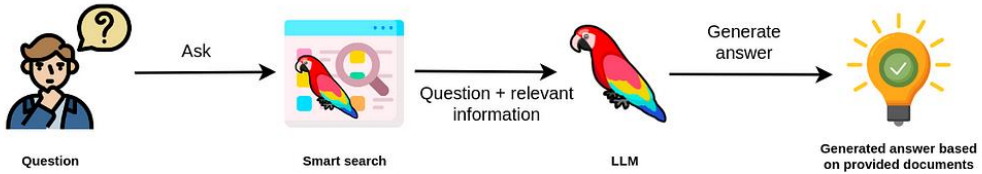


Image credit: neo4j

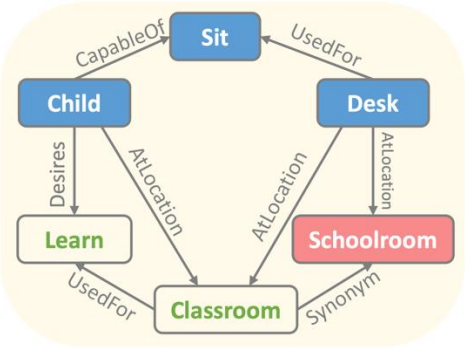
Foundation of Most Modern QA Systems (Cont'd)

Massive Knowledge Graphs (KGs)

- Capture external knowledge explicitly using triples
- Enable explainable predictions e.g. by providing reasoning paths (KagNet)
- Play in structured reasoning and query answering

Where does a child likely sit at a desk?

A. Schoolroom * B. Furniture store C. Patio
D. Office building E. Library



- Relational path:
  - **CHILD** -> AtLocation -> CLASSROOM -> Synonym -> **SCHOOLROOM**

Lin, Bill Yuchen, et al. "Kagnet: Knowledge-aware graph networks for commonsense reasoning." *arXiv preprint arXiv:1909.02151* (2019).
Feng, Yanlin, et al. "Scalable multi-hop relational reasoning for knowledge-aware question answering." arXiv preprint arXiv:2005.00646 (2020).

Foundation of Most Modern QA Systems (Cont'd)

Extending KG to general QA

- Questions and answers are expressed in natural language
  - Not easily mapped to strict logical queries
- Require finding the right integration
  - Knowledge from KG with information and constraints provided by QA example

MedQA-USMLE	A 57-year-old man presents to his primary care physician with a 2-month history of right upper and lower extremity weakness. He noticed the weakness when he started falling far more frequently while running errands. Since then, he has had increasing difficulty with walking and lifting objects. His past medical history is significant only for well-controlled hypertension, but he says that some members of his family have had musculoskeletal problems. His right upper extremity shows forearm atrophy and depressed reflexes while his right lower extremity is hypertonic with a positive Babinski sign. Which of the following is most likely associated with the cause of this patients symptoms? (A) HLA-B8 haplotype (B) HLA-DR2 haplotype **(C) Mutation in SOD1** (D) Mutation in SMN1

Previous Works

Integrating LMs and KGs

- Two challenges: given a QA context, methods need to
  - Identify informative knowledge from a large KG
  - Capture the nuance of the QA context and the structure of the KGs to perform joint reasoning over these two sources of information

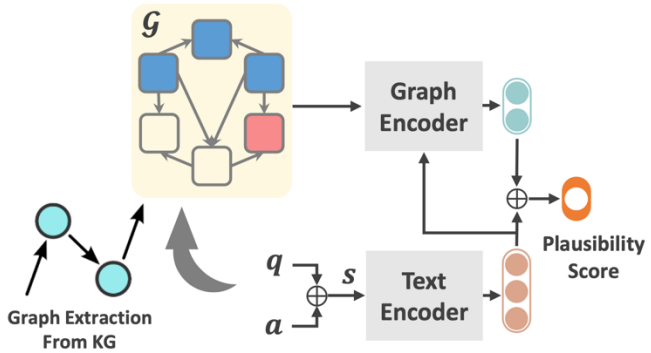


Figure 3: **Overview of the knowledge-aware QA framework.**

Image credit: MHGRN

Integrating LMs and KGs

Multi-Hop Graph Relation Network (MHGRN)

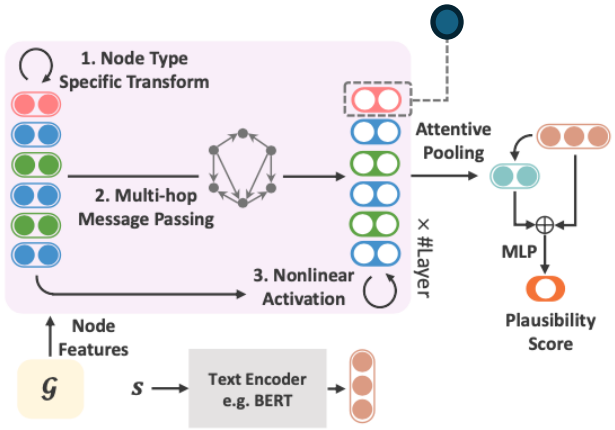
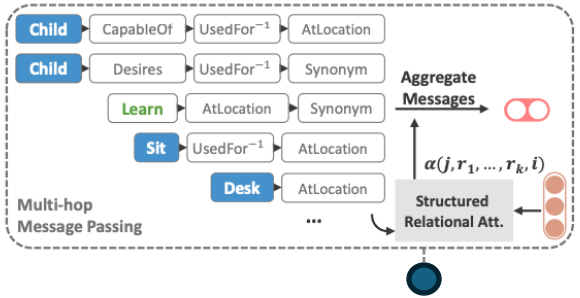
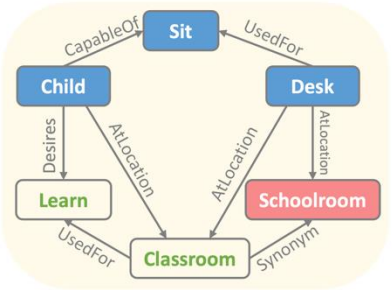


Figure 4: **Our proposed MHGRN architecture for relational reasoning.** MHGRN takes a multi-relational graph $\mathcal{G}$ and a (question-answer) statement vector $\boldsymbol{s}$ as input, and outputs a scalar that represent the plausibility score of this statement.



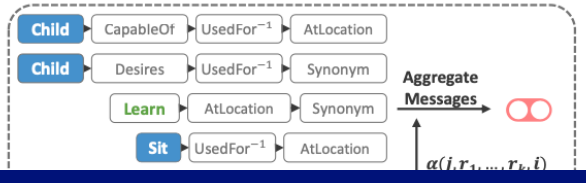


Where does a child likely sit at a desk?

A. Schoolroom * B. Furniture store C. Patio
D. Office building E. Library

Integrating LMs and KGs

❊ Multi-Hop Graph Relation Network (MHGRN)



Encoding both modalities separately and fusing them at the output for a prediction

relational graph $\mathcal{G}$ and a (question-answer) statement vector $\boldsymbol{s}$ as input, and outputs a scalar that represent the plausibility score of this statement.

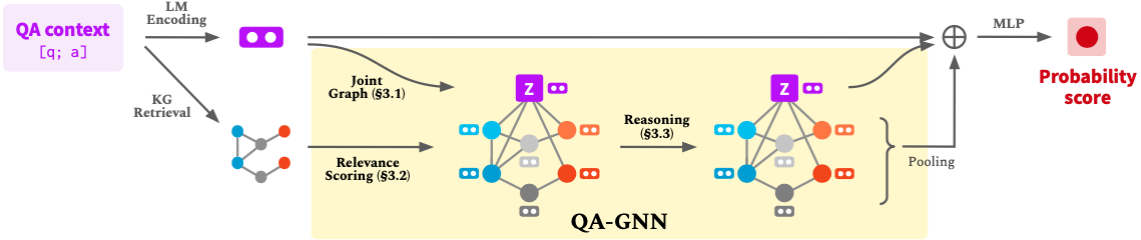
Where does a child likely sit at a desk?

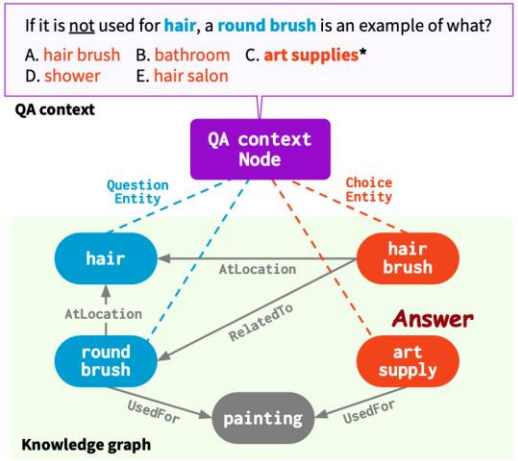
A. Schoolroom * B. Furniture store C. Patio
D. Office building E. Library

Integrating LMs and KGs

QA-GNN

- Retrieving subgraph from KG with relevance score
  - Taking topic entities (KG entities mentioned in the given QA context) and their few hop neighbors
  - Using LM to estimate the importance of KG nodes relative to given QA context
- Joint reasoning
  - Connecting QA context and KG to form a joint graph
  - Mutually updating their representations through GNNs





QA-GNN

❊ KG node relevance scoring

- ❑ Irrelevant nodes may result in overfitting or introduced unnecessary difficulty in reasoning, especially sub-graph is large

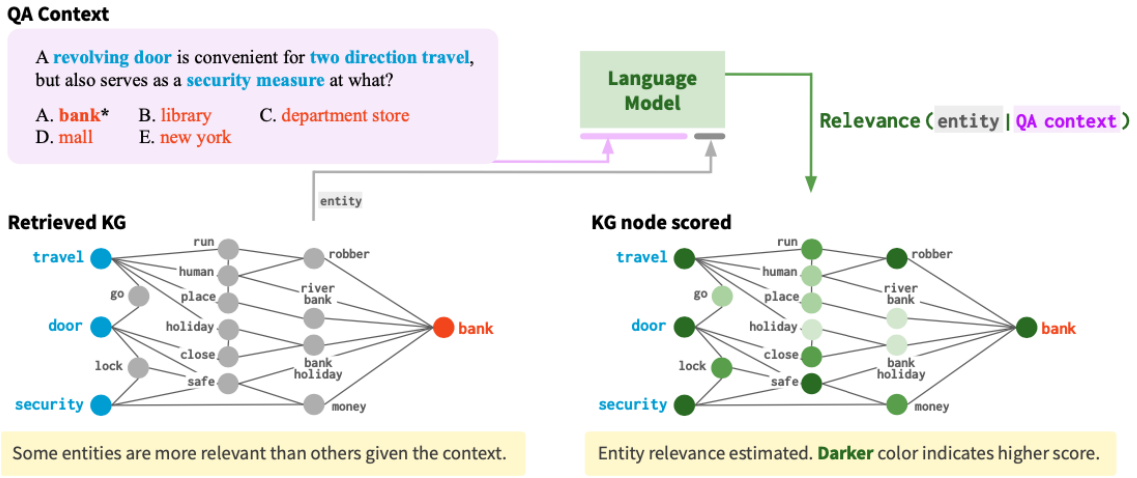
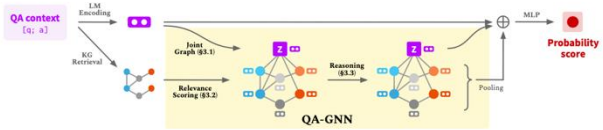


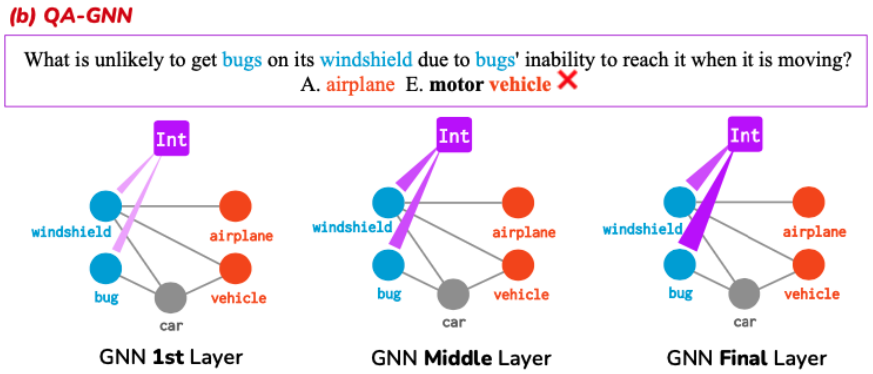
Figure 3: Relevance scoring of the retrieved KG: we use a pre-trained LM to calculate the relevance of each KG entity node conditioned on the QA context (§3.2).

Problems

QA-GNN

- Jointly updating the LM and GNN representation via message passing
- However, using a single pooled representation of the LM
  - To seed the textual component of this joint structure
- Limiting the updates that can be made to the textual representation





Problems

Solving QA tasks

- Existing KG + LM models typically fuse the two modalities **in a shallow and non-interactive manner**
- Previous methods – **restricted capacity to exchange** useful information between two modalities
- Effectively fusing the KG and LM representations is the key – **interacting in a non-shallow way**

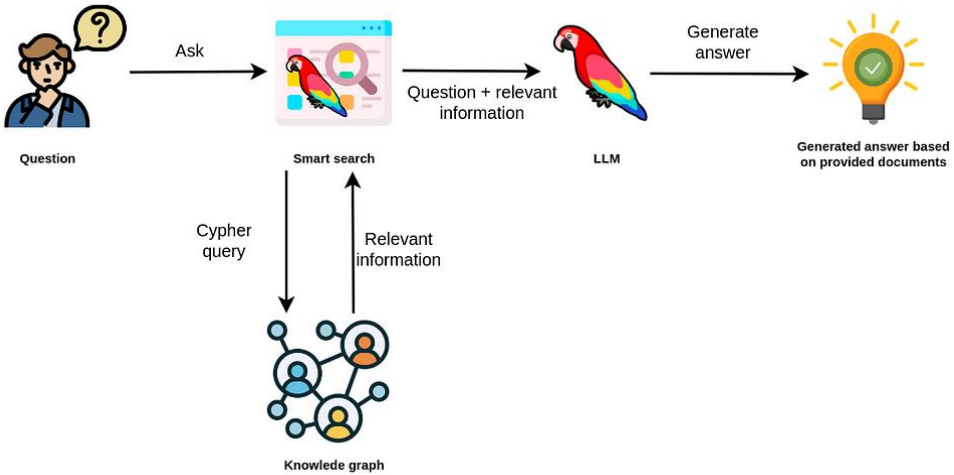
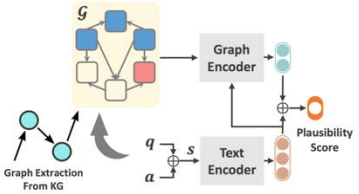


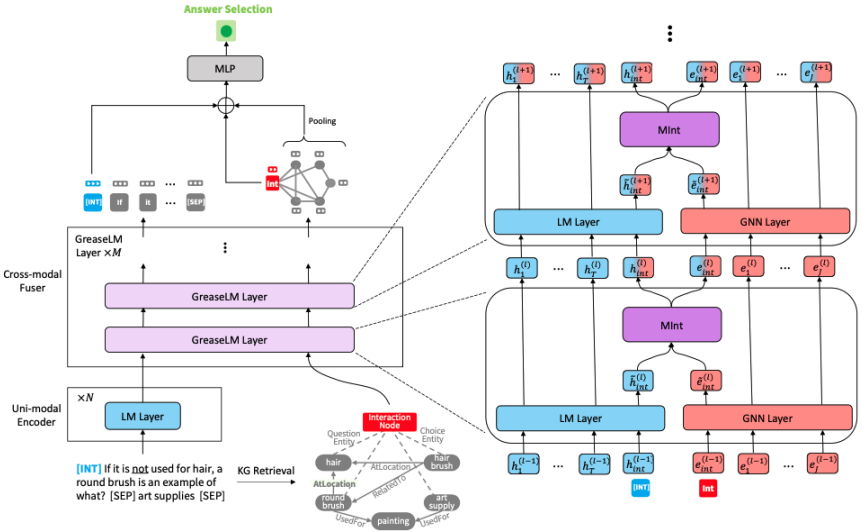
Image credit: neo4j

Proposed Method

GreaseLM

- Pre-trained Language Models + Knowledge Graphs
- Fusion each encoded representation in multiple layers of modality interaction operations
- Exchange information from both modalities
- Use benchmarks:
  - CommonsenseQA, OpenbookQA, MedQA-USMLE

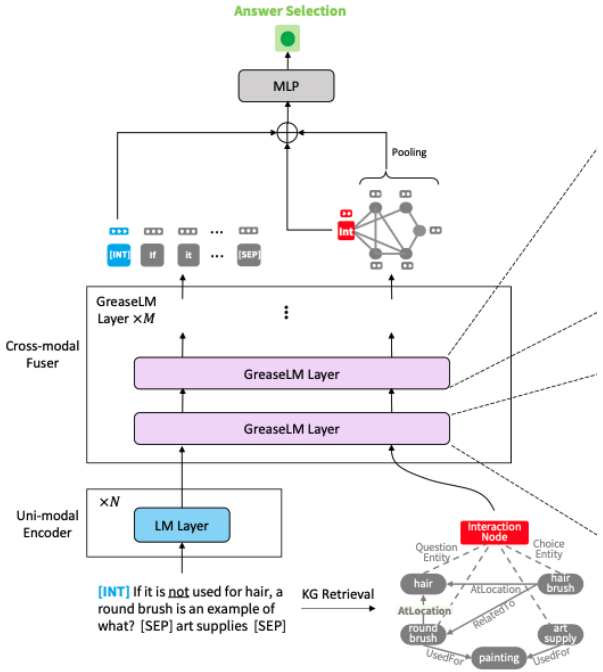




Proposed Method: GreaseLM

❊ Architecture

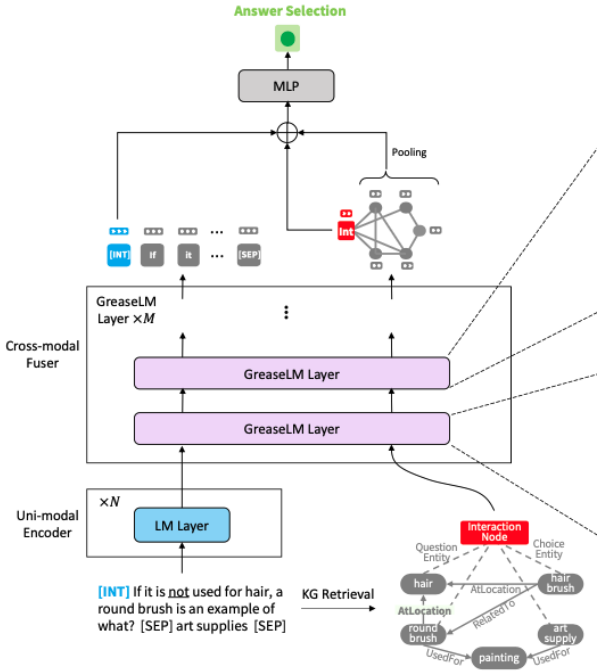
- GreaseLM consists of two stacked components
  - Unimodal LM Layers (N) ➠ initial representation of input tokens
  - Cross-modal GreaseLM Layers (M) ➠LM layers + GNN layers
- Total Layers = N + M



Proposed Method: GreaseLM

Architecture

- Input Representation
  - Concatenate context paragraph c, question q and candidate answer a with separator tokens $[c; q; a]$
  - Tokenize into $\{w_1,\ldots,w_r\}$
  - Use Input sequence to retrieve subgraph of KG
    - Set of nodes in subgraph = $\{e_1, \ldots, e_j\}$
    - Each node in subgraph is assigned a type as {context, question, answer, neighbors}
  - Special interaction token w_{int} ➠ prepend to token sequence
  - Special interaction node e_{int} ➠ connect to all linked nodes



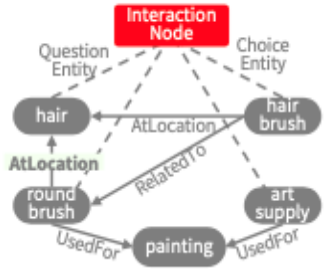
KG Retrieval

Retrieving Method

- A set of topic entities V_{linked} via entity liking to match the tokens in QA context to the entities in KG
- Add any bridge entities that are in a 2-hop path between any paired of linked entities
  - To get a set of retrieved entities
- Prune the set of retrieved nodes using relevance score computed for each node (QA-GNN)
  - Retaining the top 200 scores nodes and removing the remaining ones
- Retrieve all the edges that connect any two nodes in V_{sub}
- Form the retrieved subgraph G_{sub}

[INT] If it is not used for hair, a round brush is an example of what? [SEP] art supplies [SEP]

KG Retrieval



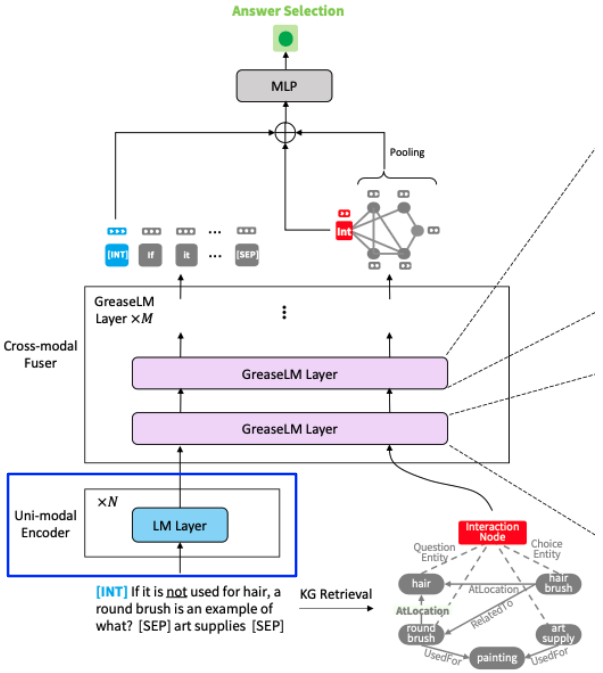
Proposed Method: GreaseLM

❋ Architecture

- ❑ Language Pre-encoding (Unimodal Layer)
  - ◆ $h_{int}^0, h_1^0, \ldots, h_r^0$ = token + segment + positional embeddings

$$\{\boldsymbol{h}_{int}^{(\ell)}, \boldsymbol{h}_1^{(\ell)}, \ldots, \boldsymbol{h}_T^{(\ell)}\} = \text{LM-Layer}(\{\boldsymbol{h}_{int}^{(\ell-1)}, \boldsymbol{h}_1^{(\ell-1)}, \ldots, \boldsymbol{h}_T^{(\ell-1)}\}) \quad \text{for } \ell = 1, \ldots, N$$

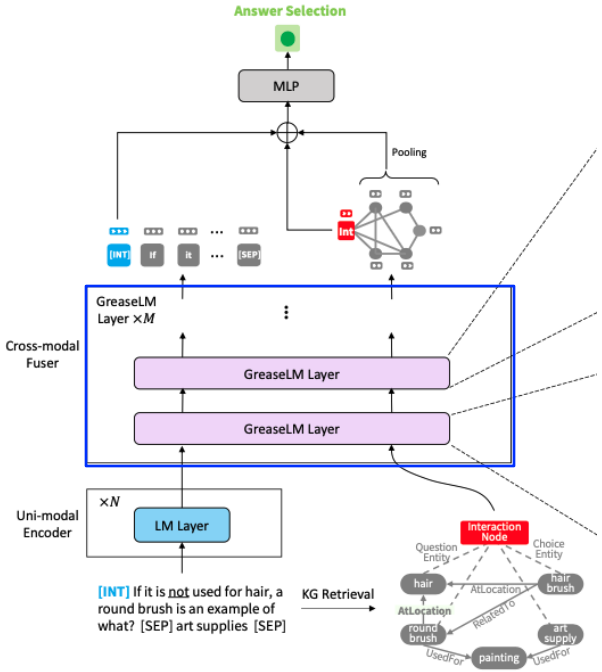
  - ◆ LM-Layer = a single LM encoder layer



Proposed Method: GreaseLM

❊ Architecture

- ❑ GreaseLM Layer (Cross-Modal Layers)
  - ◆ What does
    - ○ Separately encode information from both modalities
    - ○ Fuse both representations using the special interaction token and node
  - ◆ Three components
    - ○ A transformer LM encoder block ➠ Continue to encode language text
    - ○ A GNN layer ➠ Reason over KG entities and relations
    - ○ A Modality interaction layer ➠ TAKE unimodal rep. of interaction token and node, and EXCHANGE information through them

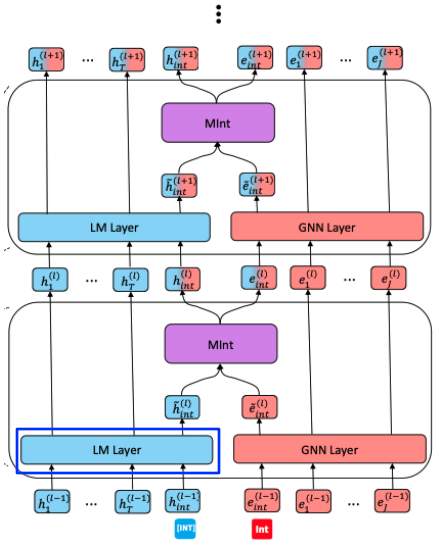


Proposed Method: GreaseLM

❋ Architecture

- ❑ Language Representation (A transformer LM encoder block)
  - ◆ Continue to encode language text in l-th GreaseLM layer

$$\{\tilde{\boldsymbol{h}}_{int}^{(N+\ell)}, \tilde{\boldsymbol{h}}_1^{(N+\ell)}, \ldots, \tilde{\boldsymbol{h}}_T^{(N+\ell)}\} = \text{LM-Layer}(\{\boldsymbol{h}_{int}^{(N+\ell-1)}, \boldsymbol{h}_1^{(N+\ell-1)}, \ldots, \boldsymbol{h}_T^{(N+\ell-1)}\}) \quad \text{for } \ell = 1, \ldots, M$$



Proposed Method: GreaseLM

❋ Architecture

- ❑ Graph Representation (A GNN Layer)
  - ◆ Reasoning over KG entities and relations
  - ◆ Current node embeddings are fed into the layer to perform a round of information propagation between nodes in the graph
  - ◆ e_{int}^0: initialized randomly

$$\{\tilde{\boldsymbol{e}}_{int}^{(\ell)}, \tilde{\boldsymbol{e}}_1^{(\ell)}, \ldots, \tilde{\boldsymbol{e}}_J^{(\ell)}\} = \text{GNN}(\{\boldsymbol{e}_{int}^{(\ell-1)}, \boldsymbol{e}_1^{(\ell-1)}, \ldots, \boldsymbol{e}_J^{(\ell-1)}\}) \quad \text{for } \ell = 1, \ldots, M$$

$$\tilde{\boldsymbol{e}}_j^{(\ell)} = f_n\left(\sum_{e_s \in \mathcal{N}_{e_j} \cup \{e_j\}} \alpha_{sj} \boldsymbol{m}_{sj}\right) + \boldsymbol{e}_j^{(\ell-1)}$$

$$\boldsymbol{r}_{sj} = f_r(\tilde{\boldsymbol{r}}_{sj}, \boldsymbol{u}_s, \boldsymbol{u}_j)$$

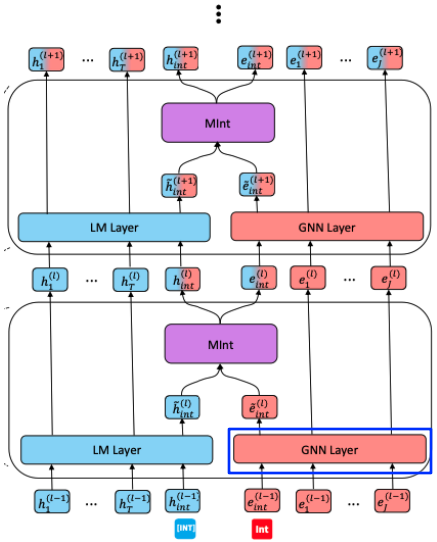
$$\boldsymbol{m}_{sj} = f_m(\boldsymbol{e}_s^{(\ell-1)}, \boldsymbol{u}_s, \boldsymbol{r}_{sj})$$

$$\boldsymbol{q}_s = f_q(\boldsymbol{e}_s^{(\ell-1)}, \boldsymbol{u}_s)$$

$$\boldsymbol{k}_j = f_k(\boldsymbol{e}_j^{(\ell-1)}, \boldsymbol{u}_j, \boldsymbol{r}_{sj})$$

$$\gamma_{sj} = \frac{\boldsymbol{q}_s^\top \boldsymbol{k}_j}{\sqrt{D}}$$

$$\alpha_{sj} = \frac{\exp(\gamma_{sj})}{\sum_{e_s \in \mathcal{N}_{e_j} \cup \{e_j\}} \exp(\gamma_{sj})}$$



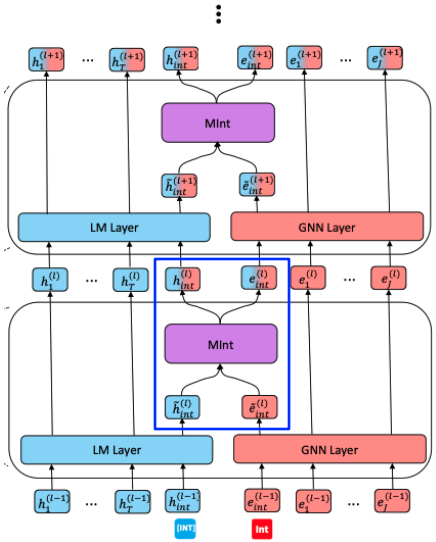
Proposed Method: GreaseLM

Architecture

- A modality interaction layer (MInt)
  - Concatenate pre-fused embeddings of ***int*** token and ***int*** node
  - Pass the joint representation through a mixing operation (MInt)
  - Split the output post-fused embeddings into ***int*** token and ***int*** node

$$[\boldsymbol{h}_{int}^{(\ell)}; \boldsymbol{e}_{int}^{(\ell)}] = \text{MInt}([\tilde{\boldsymbol{h}}_{int}^{(\ell)}; \tilde{\boldsymbol{e}}_{int}^{(\ell)}]),$$

    - Mint(.) : A two-layer MLP



Proposed Method: GreaseLM

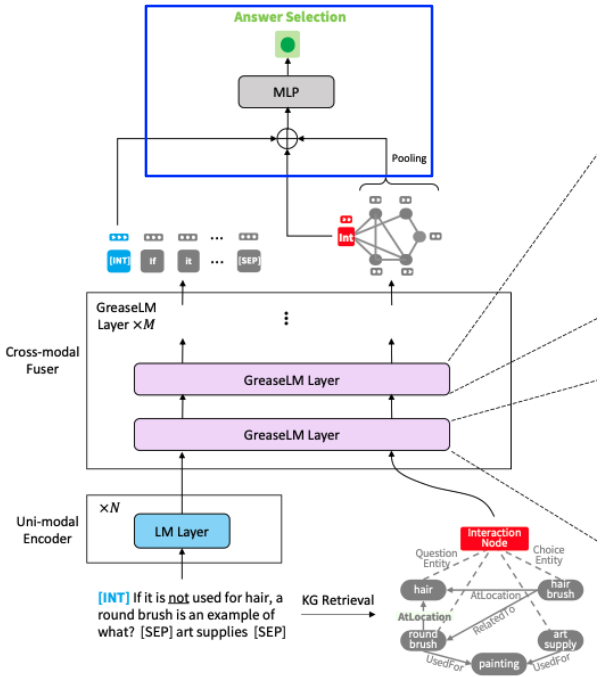
Architecture

- Learning and Inference
  - Compute the probability of a being correct answer

$$p(a \mid q, c) \propto \exp(\text{MLP}(\boldsymbol{h}_{int}^{(N+M)}, \boldsymbol{e}_{int}^{(M)}, \boldsymbol{g}))$$

    - g : attention-based pooling
  - Loss function : Cross Entropy Loss
  - Inference time : predict the most plausible answer

$$\arg\max_{a \in \mathcal{A}} p(a \mid q, c).$$



Experiments

Setup

- Three datasets with two cross domains:
  - CommonsenseQA, OpenBookQA, and MedQA-USMLE
- Knowledge Graphs:
  - ConceptNet and self-constructed KG (Diseased Datavase portion of UMLS and DrugBank) for MedQA-USMLE
- Fine-tuned Language Models
  - RoBERTa-Large (CommonsenseQA), AristoRoBERTa (OpenBookQA), and SapBERT (MedQA-USMLE)
  - To allow model to better understand entity knowledge

Experiments

❇ GreaseLM Results

- ❑ GreaseLM's multi-layer fusion is more expressive than others LM+KG methods

Table 2: **Performance comparison on *CommonsenseQA* 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). Experiments are controlled using same seed LM.

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., 2019)	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)
GREASELM (**Ours**)	**78.5** (±0.5)	**74.2** (±0.4)

Table 3: **Test Accuracy comparison on *OpenBookQA***. Experiments are controlled using the same seed LM for all LM+KG methods.

Model	Acc.
AristoRoBERTa (no KG)	78.4
+ RGCN	74.6
+ GconAttn	71.8
+ RN	75.4
+ MHGRN	80.6
+ QA-GNN	82.8
GREASELM (**Ours**)	**84.8**

Table 4: **Test accuracy comparison to public *OpenBookQA* model implementations**. *UnifiedQA (11B params) and T5 (3B) are 30x and 8x larger than our model.

Model	Acc.	# Params
ALBERT (Lan et al., 2020) + KB	81.0	∼235M
HGN (Yan et al., 2020)	81.4	≥355M
AMR-SG (Xu et al., 2021)	81.6	∼361M
ALBERT + KPG (Wang et al., 2020)	81.8	≥235M
QA-GNN (Yasunaga et al., 2021)	82.8	∼360M
T5* (Raffel et al., 2020)	83.2	**∼3B**
T5 + KB (Pirtoaca)	85.4	**≥11B**
UnifiedQA* (Khashabi et al., 2020)	**87.2**	**∼11B**
GREASELM (**Ours**)	84.8	∼359M

Experiments

GreaseLM Results

- An effective augmentation of pretrained LMs for different domains and KGs

Table 6: Performance on *MedQA-USMLE*

Methods	**Acc.** (%)
Baselines (Jin et al., 2021)	
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
Baselines (Our implementation)	
SapBERT-Base (w/o KG)	37.2
QA-GNN	38.0
GREASELM (**Ours**)	**38.5**

Experiments

GreaseLM Results

- Quantitative Analysis
  - Perform better than the baselines across all questions with prepositional phrases
  - Perform comparably on questions with no prepositional phrases (QA-GNN)
  - Increasing complexity of questions requires deeper cross-modal fusion between language and knowledge representations

Table 5: Performance of GREASELM on the *CommonsenseQA* IH-dev set on complex questions with semantic nuance such as prepositional phrases, negation terms, and hedge terms.

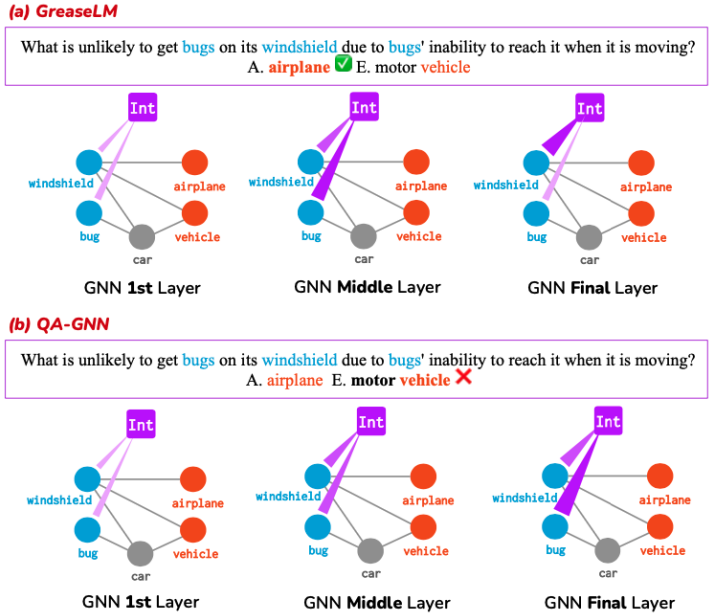
Model	# Prepositional Phrases					Negation Term	Hedge Term
	0	1	2	3	4		
n	210	429	316	171	59	83	167
RoBERTa-Large	66.7	72.3	76.3	74.3	69.5	63.8	70.7
QA-GNN	**76.7**	76.2	79.1	74.9	81.4	66.2	76.0
GREASELM (**Ours**)	75.7	**79.3**	**80.4**	**77.2**	**84.7**	**69.9**	**78.4**

Experiments

❊ GreaseLM Results

- ❑ Qualitative Analysis
  - ◆ Examine node-to-node attention weights induced by GNN layers
  - ◆ Analyze whether reflecting more expressive reasoning

As “bug” is mentioned multiple times in the context, it may be well-represented in QA-GNN’s context node initialization, which is never reformulated by language representations, unlike in GreaseLM.



Future Works

Ethics Statement

- GreaseLM:
  - A method to fuse language rep. and KG rep. for effective reasoning about textual situations
  - Reflect many of the same biases and toxic behaviors exhibited by LMs and KGs that are used to initialize it
  - e.g. biases about race, gender, and other demographic attributes

- KG: encode stereotypes rather than completely clean commonsense knowledge
  - Unethical relationships in its knowledge resource to arrive at conclusions
  - E.g. ConceptNet encoding "women" as "emotional" or "men" as "strong."
    - Question: *"Who is likely to cook dinner?"*
    - Biased Answer: *"A woman."*

Conclusion

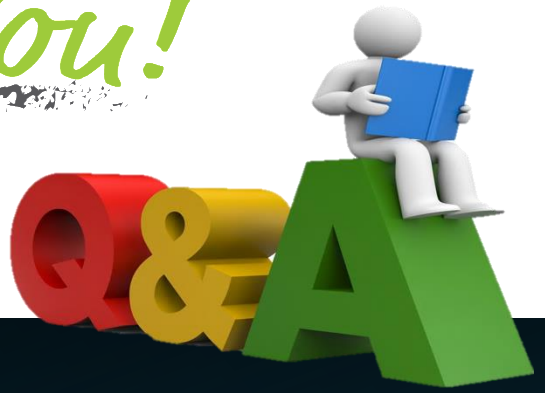
- **GreaseLM:** a new model that enables interactive fusion through joint information exchange between knowledge from language models and knowledge graphs

- Experimental results demonstrate superior performance compared to prior KG+LM and LM-only baselines

- Improved capability modeling questions exhibiting textual nuances, such as negation and hedging

- Discussion about Ethical Concerns
  - Reflect many of the same biases and toxic behaviors exhibited by LMs and KGs

Thank You!



HTET ARKAR
hak3601@cau.ac.kr