Reasoning with Language Models and Knowledge Graphs

Michihiro Yasunaga

Stanford University

Joint work with H. Ren, A. Bosselut, X. Zhang, C.D. Manning, P. Liang, J. Leskovec

Stanford Artificial Intelligence

Stanford NLP

Reasoning with Knowledge

Q: 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

Common sense

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Encyclopedia

Q: Who are the current presidents of European countries that never held a world cup?

Multi-hop

Q: Predict which drugs interact with proteins that are predicted to associate with a given disease?

Biomedical expertise

Reasoning with Knowledge

Question:

where is the bowling hall of fame located?

Long Answer:

The World Bowling Writers (WBW) International Bowling Hall of Fame was established in 1993 and is located in the International Bowling Museum and Hall of Fame , on the International Bowling Campus in Arlington , Texas.

Short Answer:

Arlington , Texas

Passage Segment

...The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process...

Question

Which governing bodies have veto power?

Rajpurkar+2016, Kwiatkowski+2019

Where is Knowledge?

Knowledge can be stored in:

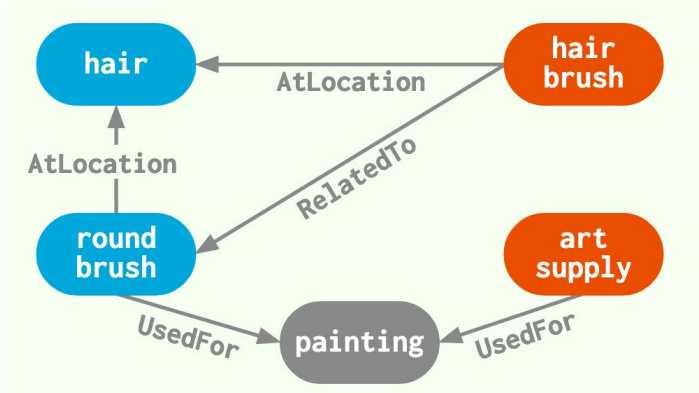
Text & Pretrained Language Model (LM)

[Devlin+2019; Liu+2019; Brown+2020; ….]

Knowledge Graph (KG)

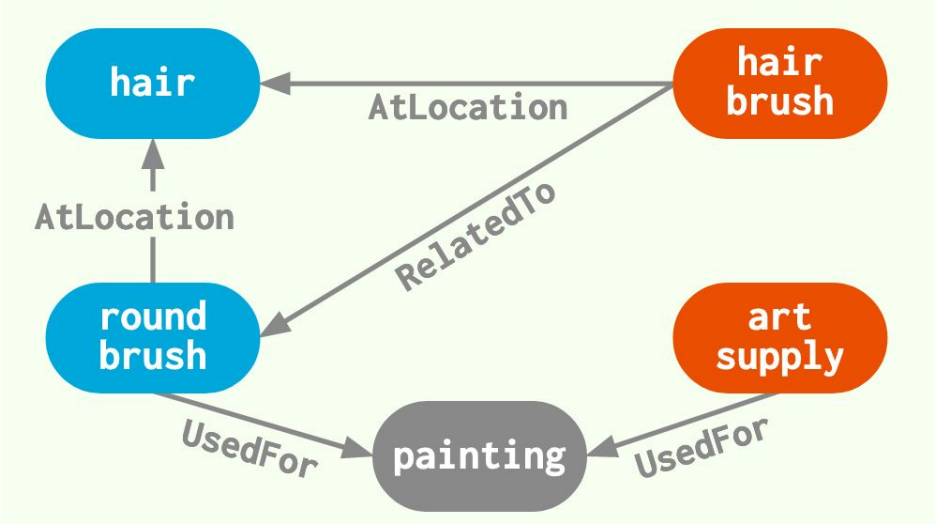
ConceptNet
An open, multilingual knowledge g

Freebase



[Bollacker+2008; Speer+2016]

(1) Knowledge Graphs



ConceptNet
An open, multilingual knowledge g

Freebase

[Bollacker+2008; Speer+2016]

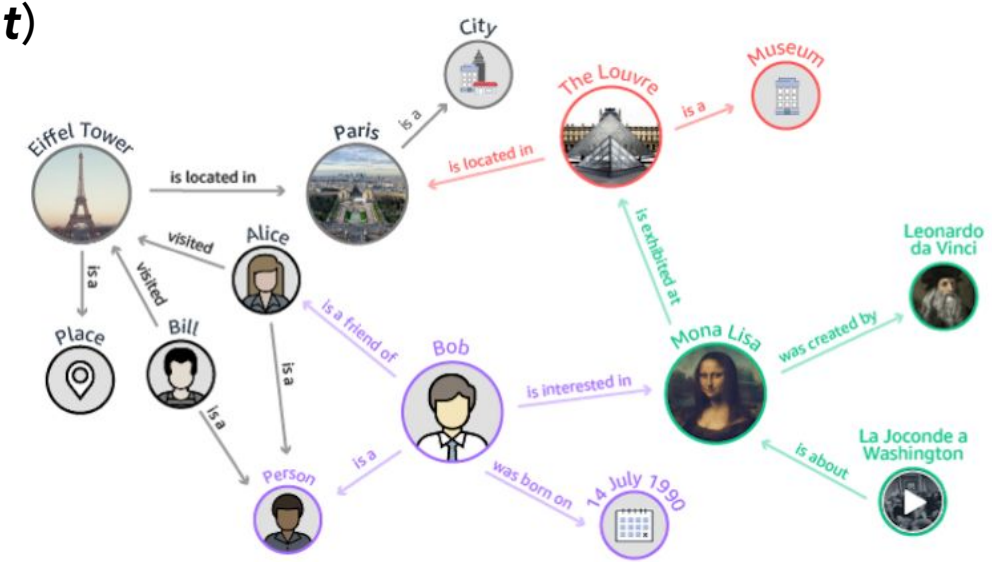
Knowledge Graphs (KGs)

Knowledge Graphs are heterogenous graphs

- Multiple types of entities and relations exist

Facts are represented as triples (***h***, ***r***, ***t***)

- (‘Paris’, ‘is_a’, ‘City’)
- (‘Paris’, ‘pupulation’, ‘2.1m’)
- ...

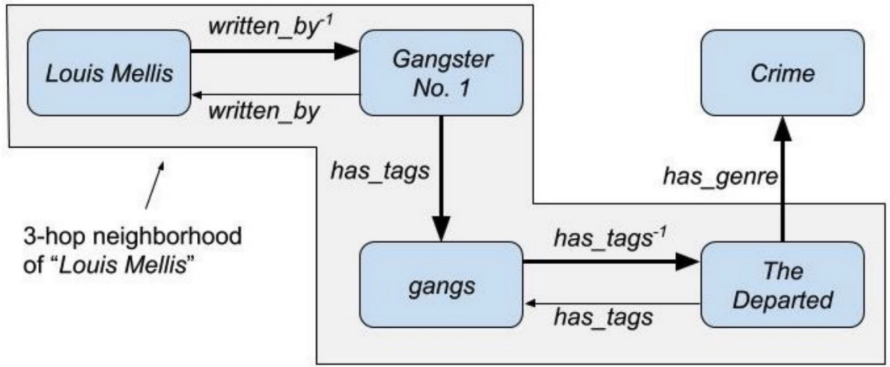


Structured Query

Approach:

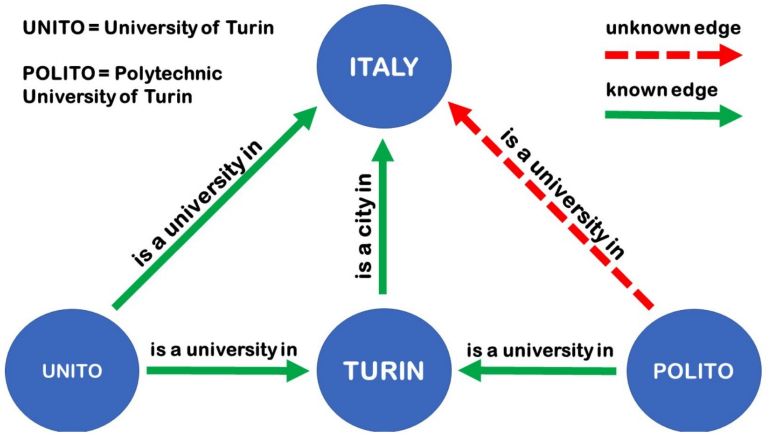
- Translate the question into a structured query (e.g. SQL)
- Execute the query on the knowledge graph
  - Match grounded entities
  - “Traverse” the knowledge graph

Question: *What are the genres of movies written by Louis Mellis?*
Answer : *Crime*



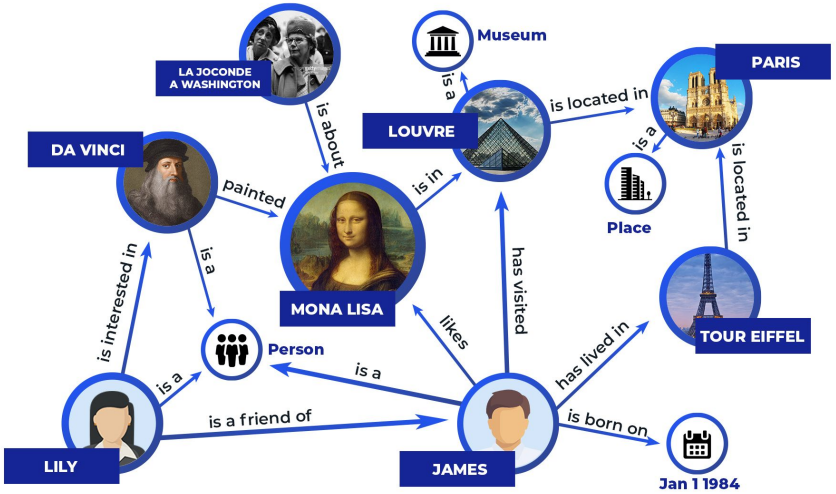
Strengths of KGs

- Explicitly stores knowledge
  - Provides interpretability and provenance
- Easy to update knowledge
- Specifies structure and rules
  - Easy to model multi-hop knowledge and logical reasoning
- (some are) Curated by humans
  - Covers knowledge not commonly stated in text, e.g. commonsense KG
  - Annotates true and accepted facts, e.g. scientific KG



Challenges with KGs

- Incomplete
  - Missing entities and relations
  - Not all facts can be expressed as (h, r, t)
- Brittle
  - Hard to encode real-world complexity and context
  - Some questions can be hard to express as formal queries over KGs
- Needs entity linking / retrieval



(2) Language Models

[Devlin+2019; Liu+2019; Brown+2020; ….]

Language Models (LMs)

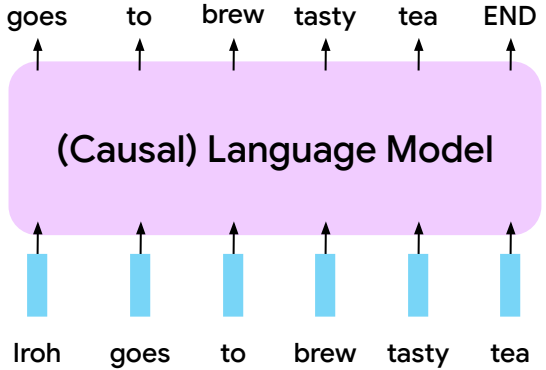
Trained over collections of text

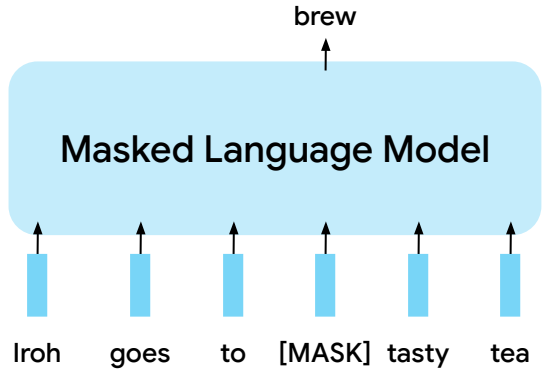
- Wikipedia, books, news, PubMed, GitHub, ...

Trained with self-supervised tasks

- (Causal) language model: predict the next word
- Masked language model: predict masked words

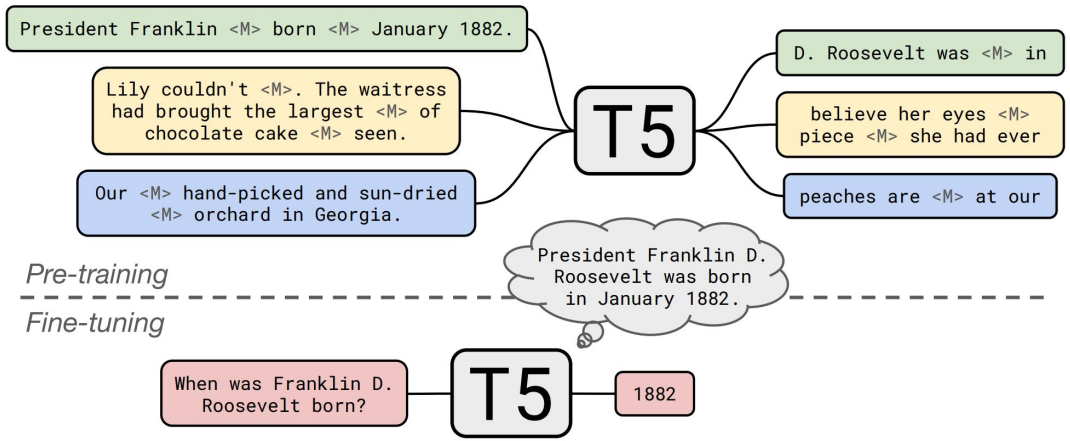
Store knowledge in the parameters of the neural network





(1) Strengths of LMs

- Broad coverage of knowledge
  - Self-supervised over massive amounts of text
- Can **encode** and **decode** anything that can be expressed as words
  - Can take in any question, and decode some textual answer or use the hidden representation to do classification
- Captures context



Raffel+2019

Challenges with LMs

- Hard to interpret or trust
  - Unclear why LM produces this answer ↔ KG has provenance
  - LM may produce a realistic but incorrect answer ↔ KG either returns the correct answer or returns no answer
- Hard to modify
  - Hard to update knowledge in LM ↔ KG is directly editable
- Unclear if it can truly reason
  - e.g. LM fails to handle negation correctly [Kassner+2020]

Goal: Combine strengths of both for reasoning

Broad Coverage

Text &
Pretrained Language Model (LM)

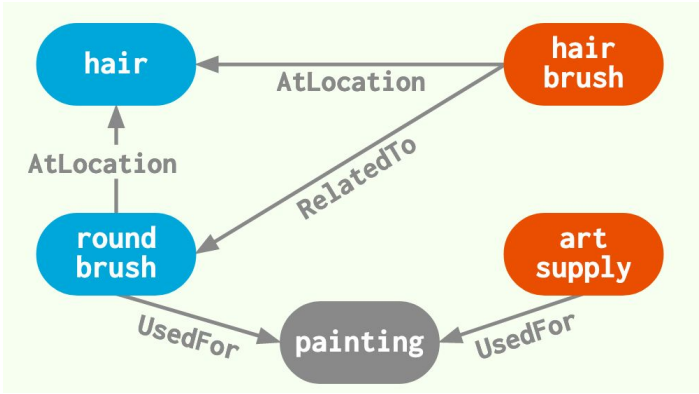
[Devlin+2019; Liu+2019; Brown+2020; ….]

Structured & Interpretable

Knowledge Graph (KG)

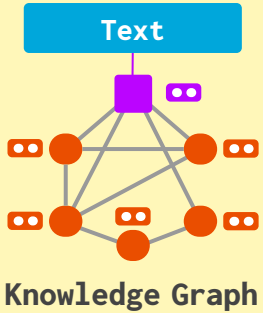
ConceptNet
An open, multilingual knowledge g

Freebase



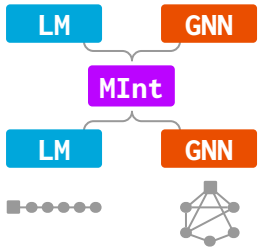
[Bollacker+2008; Speer+2016]

Outline



How to combine LM and KG for reasoning?

- **QAGNN**: Reasoning with Language Models and Knowledge Graphs for Question Answering [NAACL'21]



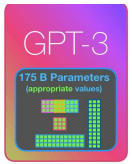
How to perform more expressive reasoning?

- **GreaseLM**: Graph Reasoning Enhanced Language Model for Question Answering [ICLR'22]

Goal: Combine language and KG

Broad Coverage

Text &
Pretrained Language Model (LM)



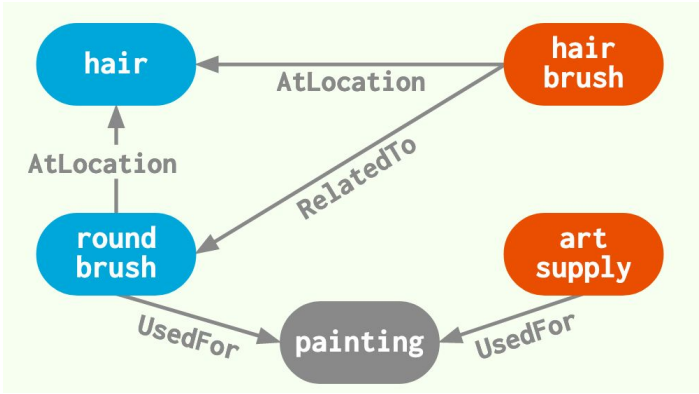
[Devlin+2019; Liu+2019; Brown+2020; ….]

Structured & Interpretable

Knowledge Graph (KG)

ConceptNet
An open, multilingual knowledge g

Freebase



[Bollacker+2008; Speer+2016]

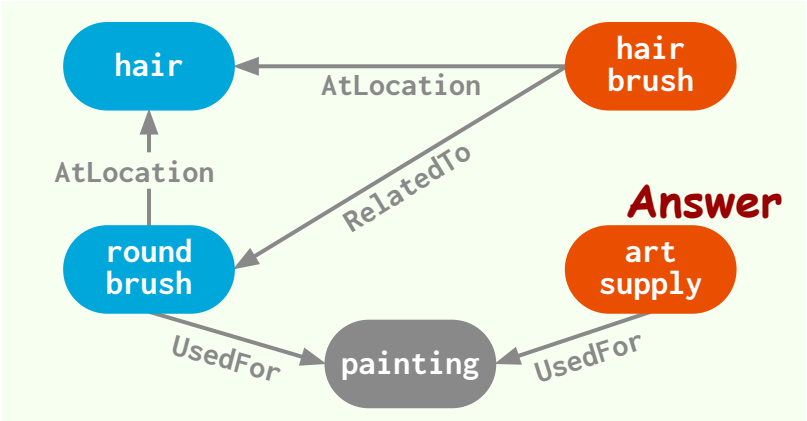
Why Is It Hard?

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

QA Context + LM

Knowledge Graph



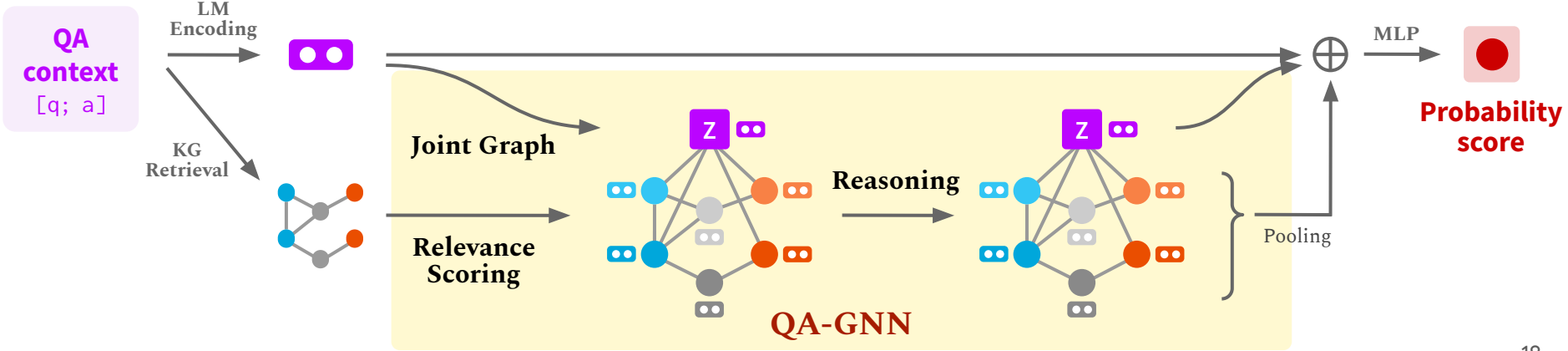
- How to identify relevant KG subset?
- How to **jointly reason** over the text and KG?

Our Idea: QA-GNN

(1) Language-conditioned KG node relevance scoring

(2) Joint Reasoning:

- Connect text and KG to form a joint graph (*working graph*)
- Mutually update their representations via Graph Neural Net (GNN)



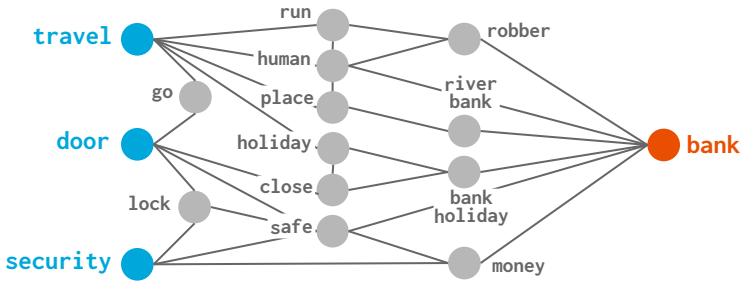
Existing KG Retrieval Method

QA Context

A **revolving door** is convenient for **two direction travel**, but also serves as a **security measure** at what?

A. **bank*** B. library C. department store
D. mall E. new york

Retrieved KG



Identify topic entities in the text:
`travel`, `door`, `security`, bank

Retrieve *k*-hop heighbors/paths in KG

Some entities are irrelevant to the given QA context

- Off-topic - e.g. **`holiday`**
- Polysemy - e.g. **`river_bank`**
- Generic - e.g. **`human`, `place`**

Ours: Score KG nodes by LM

QA Context

A **revolving door** is convenient for **two direction travel**, but also serves as a **security measure** at what?

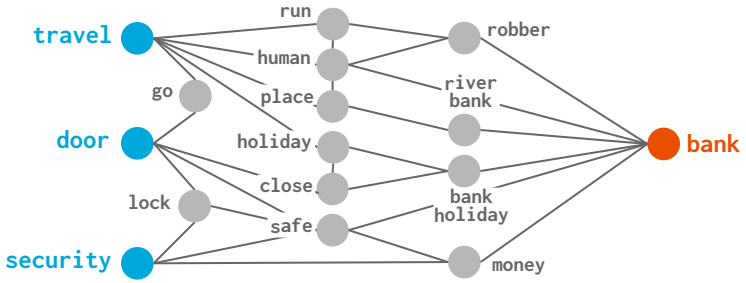
A. **bank*** B. library C. department store
D. mall E. new york

Language Model

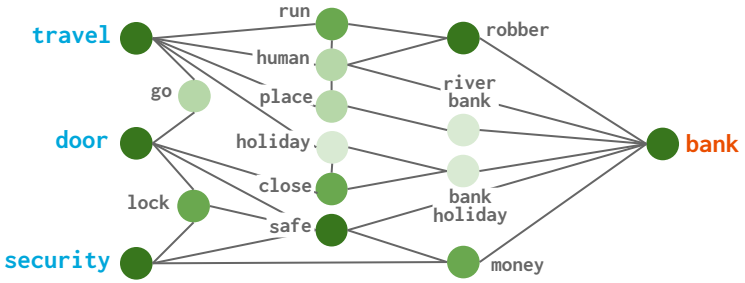
Relevance(entity | context)

entity

Retrieved KG



KG node scored



Some entities are irrelevant to the given QA context!

Entity relevance estimated by LM. **Darker** color indicates higher score.

How to Use the KG Node Scores?

- **Option 1.** Prune KG nodes
  - Reduces noise in KG. Improves model efficiency (time/space)

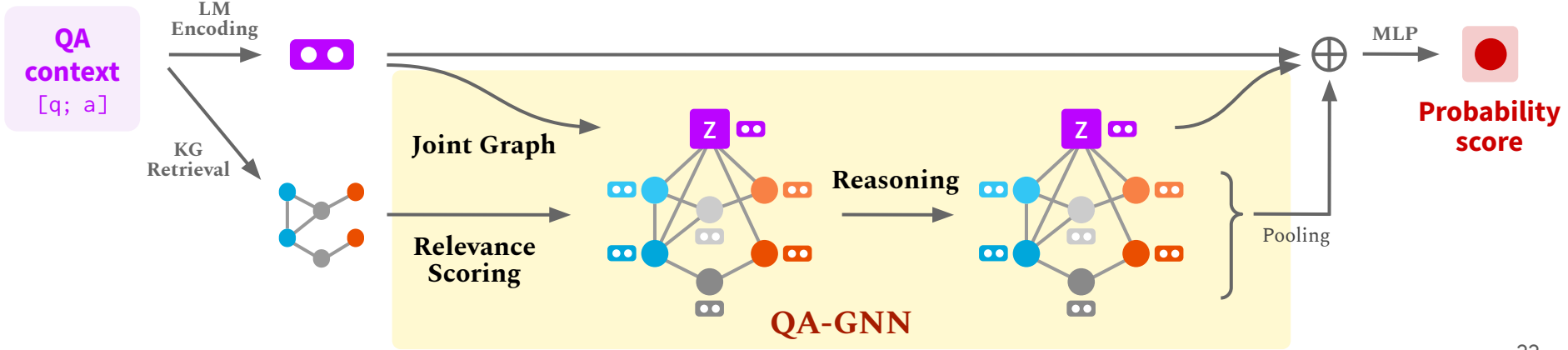
- **Option 2.** Incorporate as auxiliary feature of KG node
  - General way to weight information on KG

Our Idea: QA-GNN

(1) Language-conditioned KG node relevance scoring

(2) Joint Reasoning:

- Connect text and KG to form a joint graph (*working graph*)
- Mutually update their representations via Graph Neural Net (GNN)

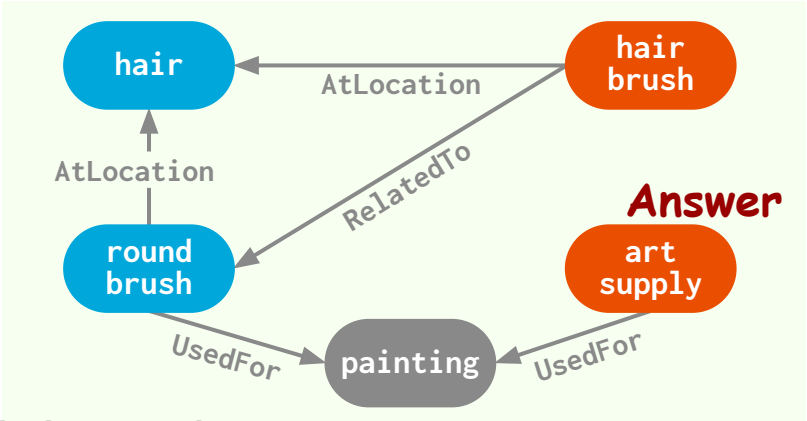


(2) Joint Reasoning

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

QA Context + LM



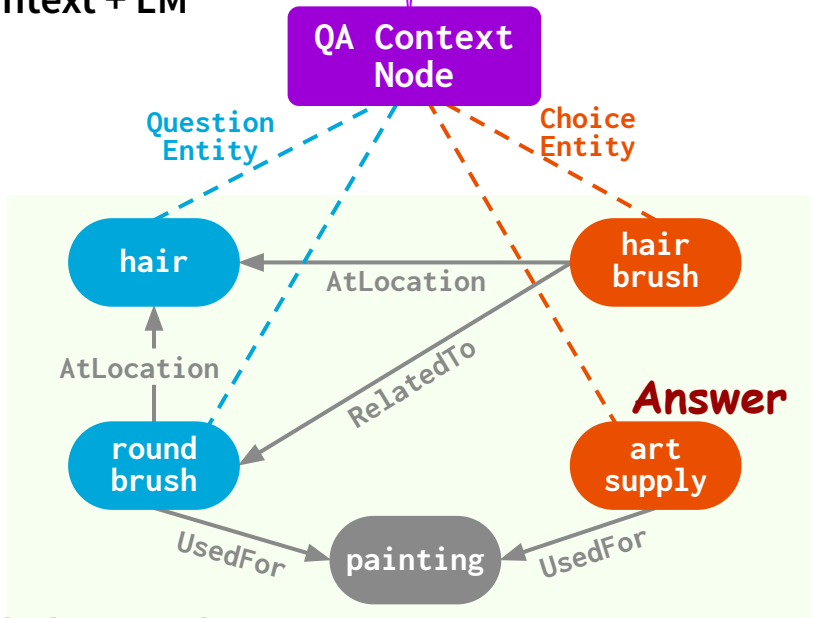
Knowledge Graph

Build Joint Graph (*Working Graph*)

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

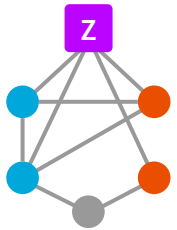
A. hair brush B. bathroom C. **art supplies*** D. shower

QA Context + LM



Knowledge Graph

Joint graph that provides a fused reasoning space for QA context and KG



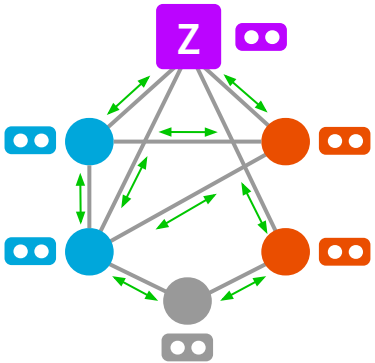
QA-GNN Message Passing

Message Passing

$$\boldsymbol{h}_t^{(\ell+1)} = f_n\left(\sum_{s\in\mathcal{N}_t\cup\{t\}} \alpha_{st}\boldsymbol{m}_{st}\right) + \boldsymbol{h}_t^{(\ell)}$$

Attention (s → t)

Message (s → t)



Node types

- QA Context
- Question entity
- Answer entity
- Other entity

Node type & relation-aware message

$$\boldsymbol{m}_{st} = f_m(\boldsymbol{h}_s^{(\ell)}, \boldsymbol{u}_s, \boldsymbol{r}_{st})$$

Node type, relation, & score-aware attention

$$\boldsymbol{q}_s = f_q(\boldsymbol{h}_s^{(\ell)}, \boldsymbol{u}_s, \boldsymbol{\rho}_s)$$

$$\boldsymbol{k}_t = f_k(\boldsymbol{h}_t^{(\ell)}, \boldsymbol{u}_t, \boldsymbol{\rho}_t, \boldsymbol{r}_{st})$$

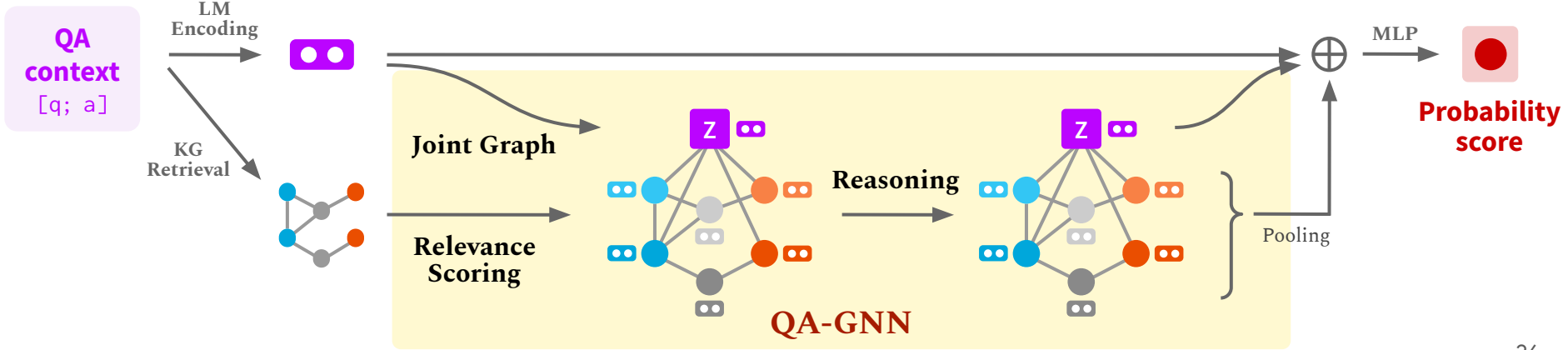
$$\alpha_{st} = \frac{\exp(\gamma_{st})}{\sum_{t'\in\mathcal{N}_s\cup\{s\}}\exp(\gamma_{st'})}, \quad \gamma_{st} = \frac{\boldsymbol{q}_s^\top \boldsymbol{k}_t}{\sqrt{D}}$$

Our Idea: QA-GNN

(1) Language-conditioned KG node relevance scoring

(2) Joint Reasoning:

- Connect text and KG to form a joint graph (*working graph*)
- Mutually update their representations via Graph Neural Net (GNN)



Experiments

QA datasets

- **CommonsenseQA** (reasoning with commonsense knowledge)
  - Train / Dev / Test: 8,500 / 1,221 / 1,241
- **OpenBookQA** (reasoning with elementary science knowledge)
  - Train / Dev / Test: 4,957 / 500 / 500

CommonsenseQA [Talmor+2018]
What do people typically do while playing guitar?

(A) cry
(B) hear sounds
(C) singing
(D) arthritis
(E) making music

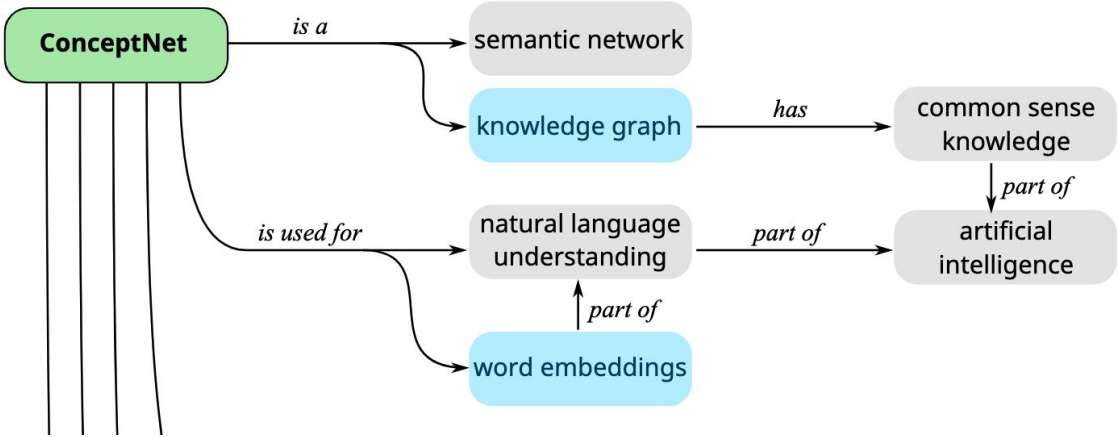
OpenBookQA [Mihaylov+2018]
Which of these would let the most heat travel through?

(A) a new pair of jeans
(B) a steel spoon in a cafeteria
(C) a cotton candy at a store
(D) a calvi klein cotton hat

Experiments

KG

- **ConceptNet (English)**
  - ~800,000 nodes
  - 17 relation types



Liu+2004, Speer+2016

Experiments

Baselines

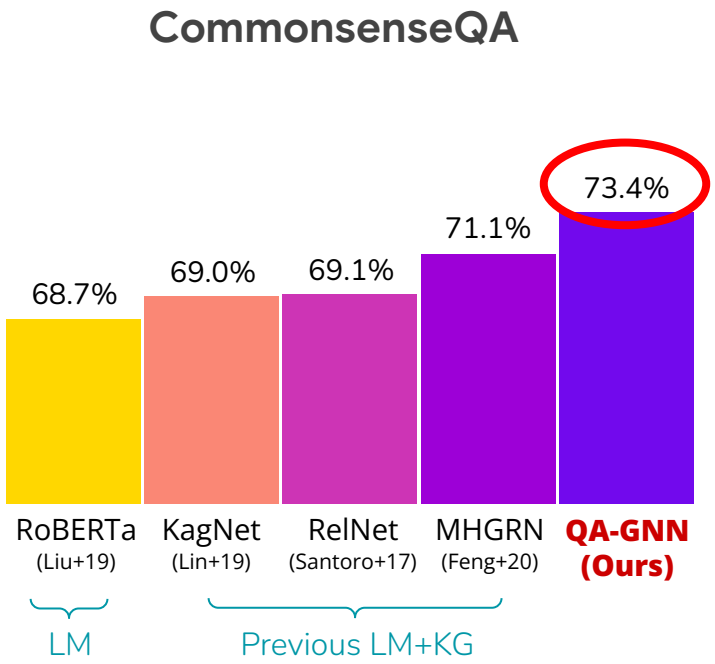
- Fine-tuned LM
  - RoBERTa [Liu+2019]
- LM+KG
  - KagNet [Lin+2019]
  - MHGRN [Feng+2020]

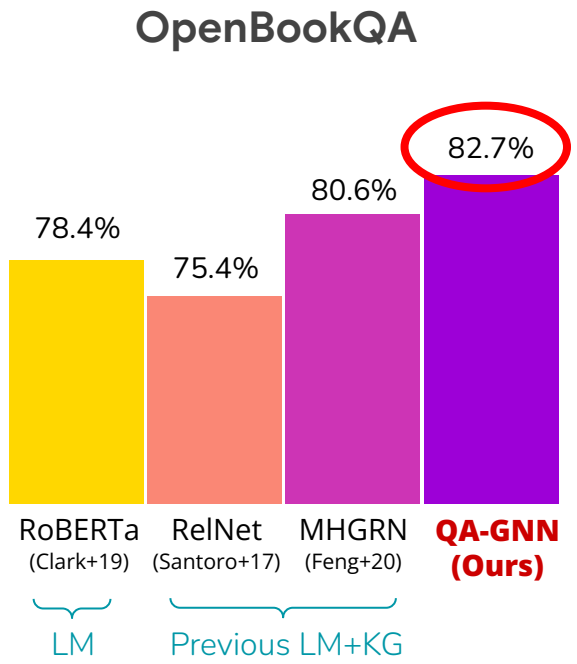
Innovations of QA-GNN:

1. Use KG node score (relevance of a node given the question)
2. Mutually update the LM and KG representations via a Graph Neural Network

Performance

Improved performance on two QA tasks





Ablation study

Performance drops to the score of previous LM+KG models

Graph Connection	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**

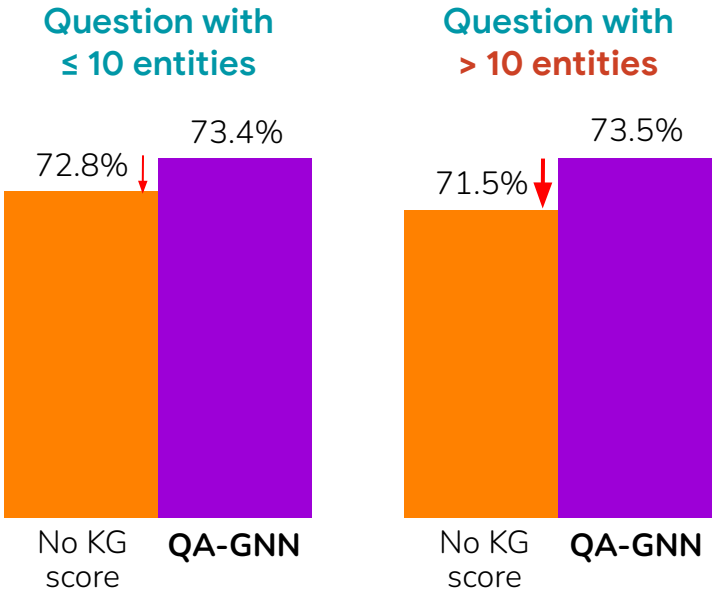
Relevance scoring	Dev Acc.
Nothing	75.56
w/ contextual embedding	76.31
w/ relevance score (**final**)	**76.54**
w/ both	76.52

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

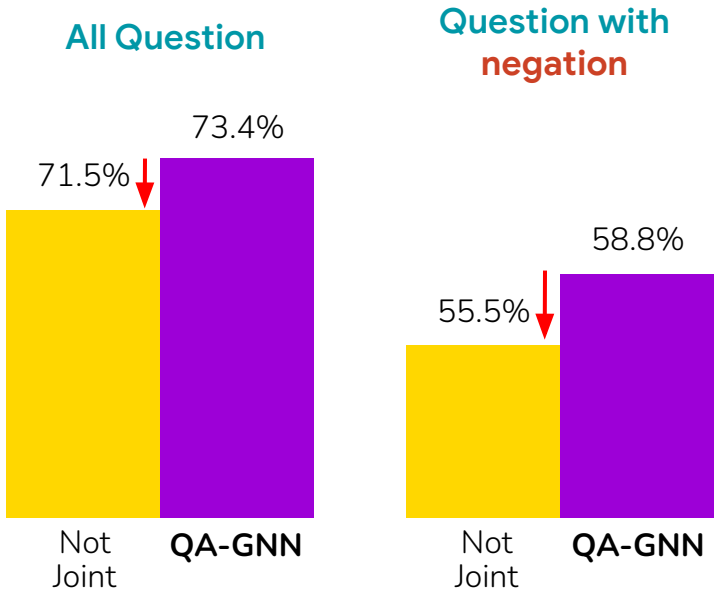
GNN Layers (§3.3)	Dev Acc.
$L = 3$	75.53
$L = 4$	76.34
$L = 5$ (**final**)	**76.54**
$L = 6$	76.21
$L = 7$	75.96

Analysis: Node Scoring & Joint Graph

- Node scoring helps when retrieved KG is big

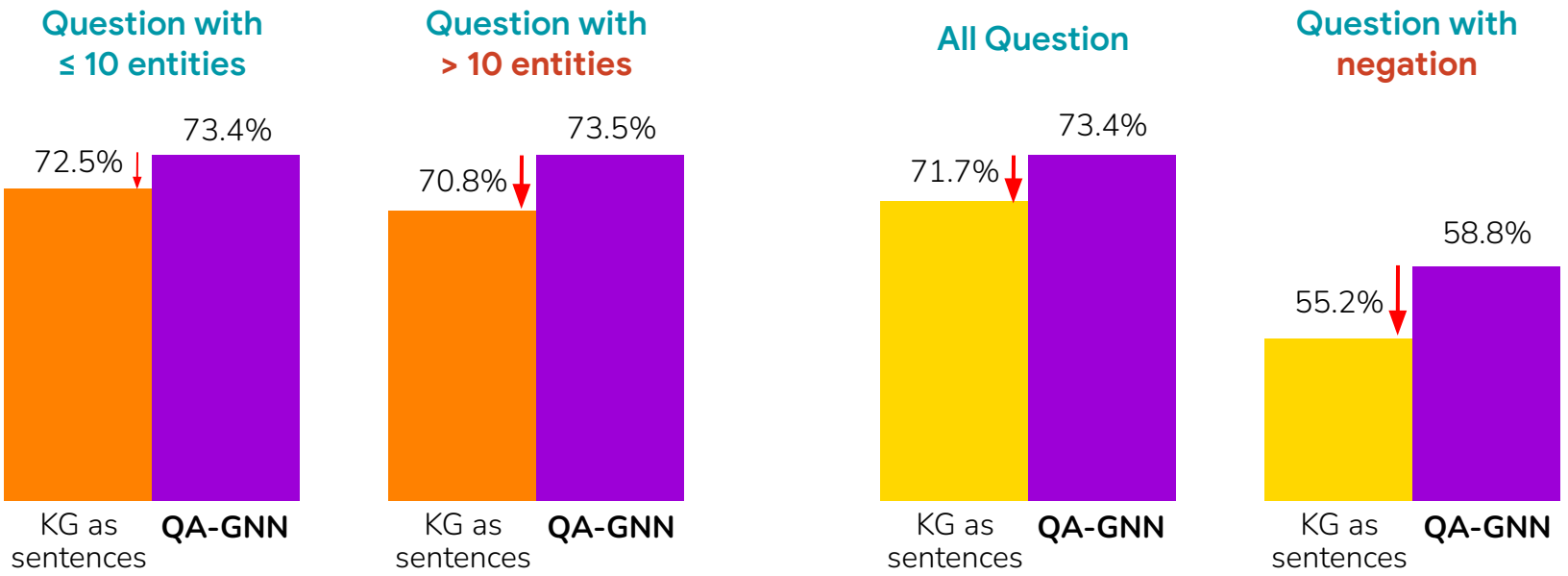


- Joint graph helps when question has negation

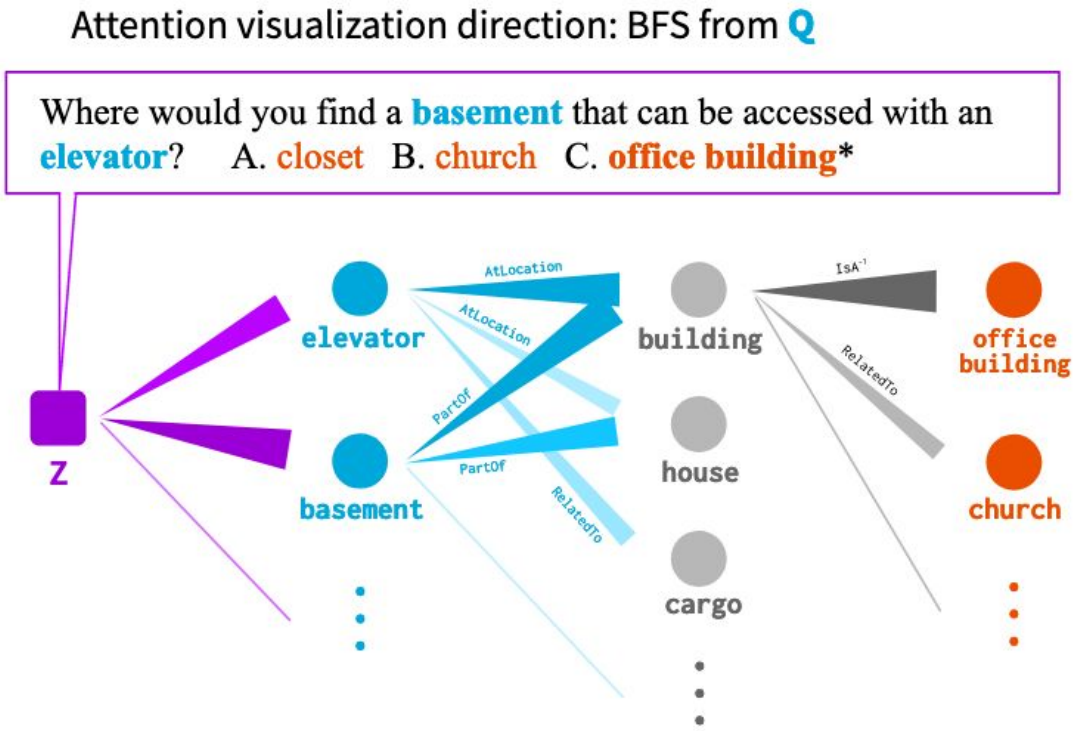


Analysis: Does KG graph structure matter?

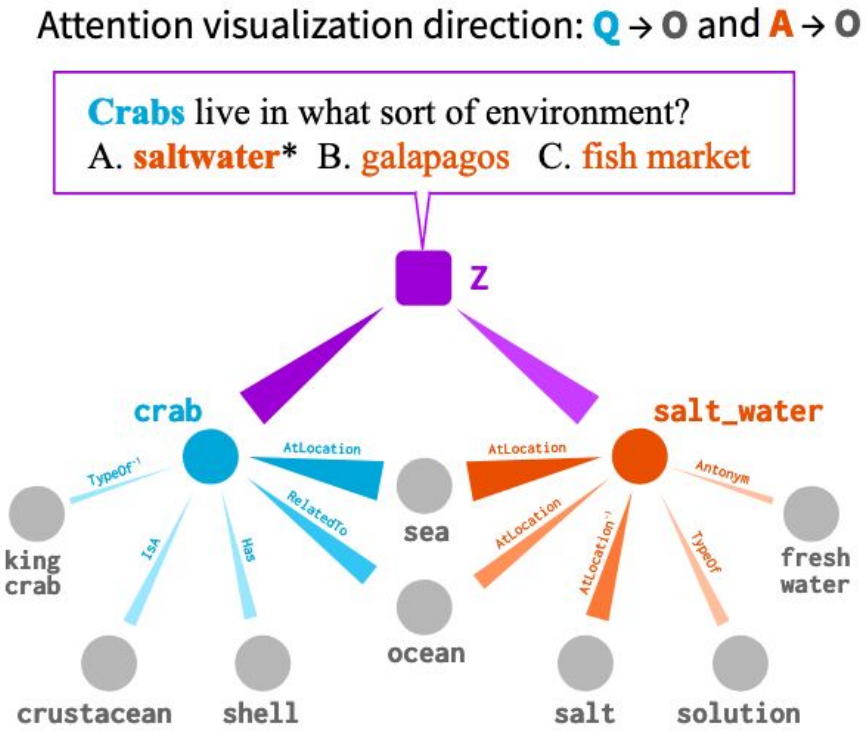
Using KG as a graph outperforms converting KG as sentences, especially on complex questions



Benefit 1: Interpretability



Benefit 1: Interpretability

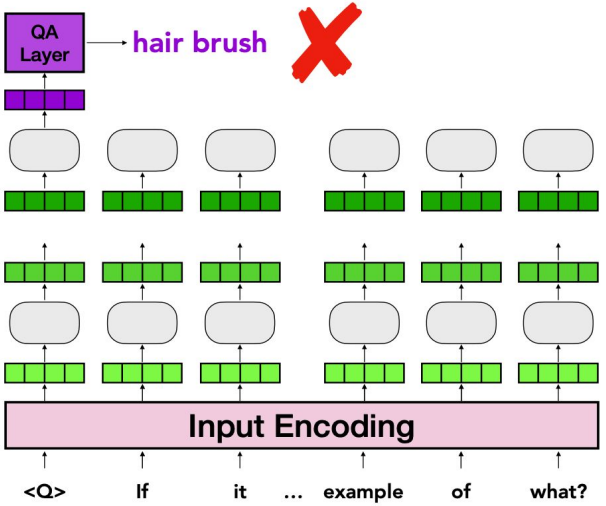


Benefit 2: Structured Reasoning

Motivation: Existing LMs struggle with negation [Kassner+2020]

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

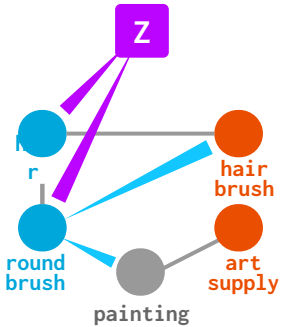
A. **hair brush** B. bathroom C. **art supplies*** D. shower



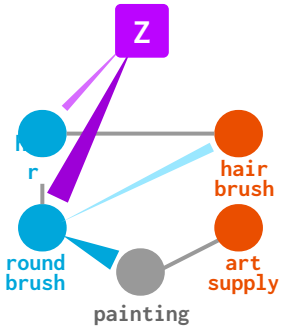
Benefit 2: Structured Reasoning

Original Question

If it is **not** used for **hair**, a **round brush** is an example of what?
A. hair brush B. **art supplies***



GNN **1st** Layer



GNN **Final** Layer

A. hair brush (#1)
B. art supplies (#2)
RoBERTa Prediction

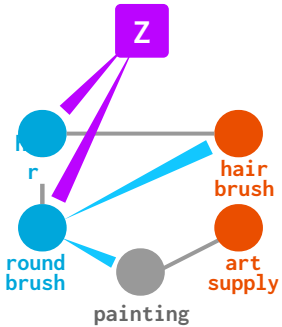
A. hair brush (#2)
B. art supplies (#1)

QA-GNN Prediction

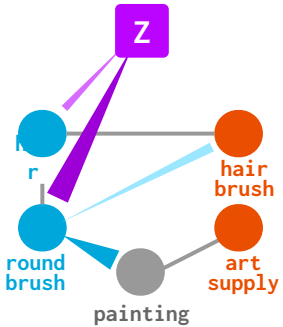
Benefit 2: Structured Reasoning

Original Question

If it is **not** used for **hair**, a **round brush** is an example of what?
A. hair brush B. **art supplies***



GNN **1st** Layer



GNN **Final** Layer

A. hair brush (#1)
B. art supplies (#2)
RoBERTa Prediction ✗

A. hair brush (#2)
B. art supplies (#1) ✓

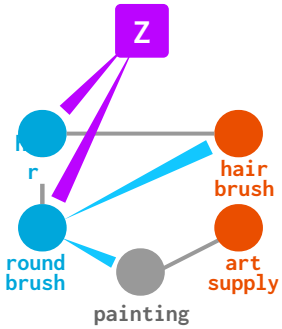
QA-GNN Prediction

After several layers of GNN,
attention weight from **text** over **hair** decreases, but
attention weight over **round brush** and **painting**
increases, adjusting for the negation in text.

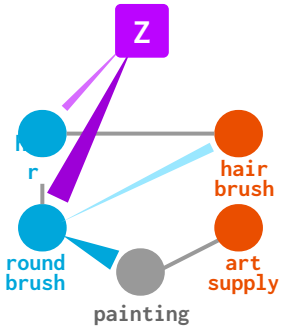
Benefit 2: Structured Reasoning

Original Question

If it is **not** used for **hair**, a **round brush** is an example of what?
A. hair brush B. **art supplies***



GNN **1st** Layer



GNN **Final** Layer

A. hair brush (#1)
B. art supplies (#2)
RoBERTa Prediction

✗

A. hair brush (#2)

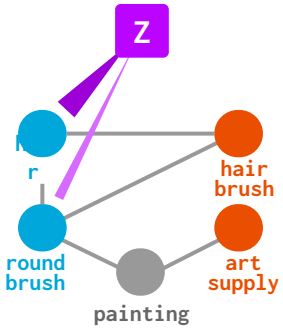
B. art supplies (#1) ✓

QA-GNN Prediction

After several layers of GNN,
attention weight from **text** over **hair** decreases, but attention weight over **round brush** and **painting** increases, adjusting for the negation in text.

Negation Flipped

If it is used for **hair**, a **round brush** is an example of what? A. hair brush B. art supplies



GNN **Final** Layer

A. hair brush (#1)
B. art supplies (#2)
RoBERTa Prediction

A. hair brush (#1) ✓

B. art supplies (#2)

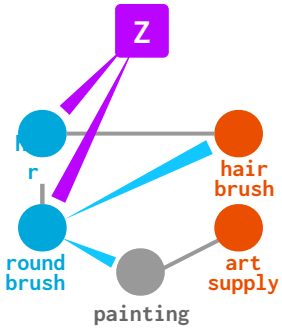
QA-GNN Prediction

Attention weight from **text** over **hair** now increases in the final layer of GNN.

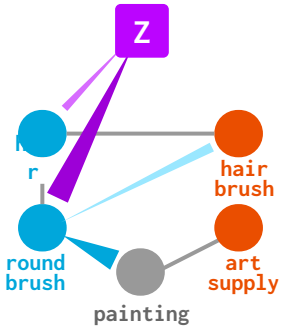
Benefit 2: Structured Reasoning

Original Question

If it is **not** used for **hair**, a **round brush** is an example of what?
A. hair brush B. **art supplies***



GNN **1st** Layer



GNN **Final** Layer

A. hair brush (#1)
B. art supplies (#2)
RoBERTa Prediction

✗

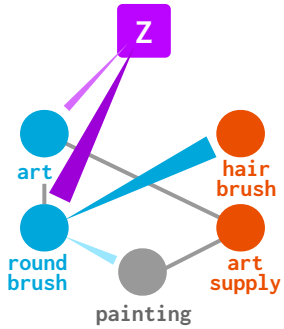
A. hair brush (#2)

B. art supplies (#1) ✓

QA-GNN Prediction

Entity Changed (hair → art)

If it is **not** used for **art**, a **round brush** is an example of what? A. hair brush B. art supplies



GNN **Final** Layer

A. hair brush (#1)
B. art supplies (#2)
RoBERTa Prediction

A. hair brush (#1) ✓

B. art supplies (#2)

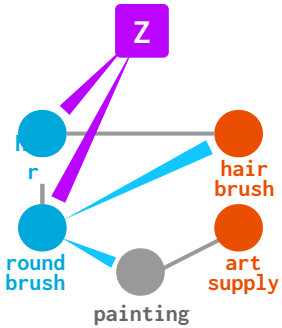
QA-GNN Prediction

Attention weight from **text** over **round brush** and from **round brush** to **hair brush** is high.

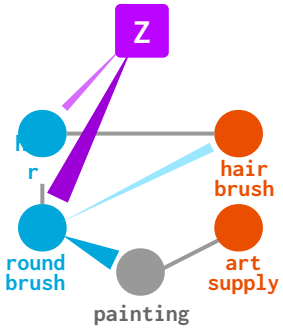
Benefit 2: Structured Reasoning

Original Question

If it is **not** used for **hair**, a **round brush** is an example of what?
A. hair brush B. **art supplies***



GNN **1st** Layer



GNN **Final** Layer

A. hair brush (#1)
B. art supplies (#2)
RoBERTa Prediction

✗

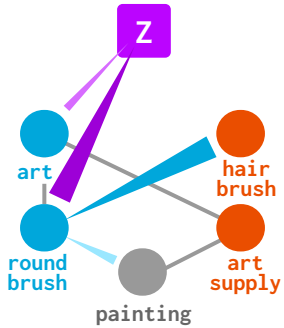
A. hair brush (#2)

B. art supplies (#1) ✓

QA-GNN Prediction

Entity Changed (hair → art)

If it is **not** used for **art**, a **round brush** is an example of what? A. hair brush B. art supplies



GNN **Final** Layer

A. hair brush (#1)
B. art supplies (#2)
RoBERTa Prediction

A. hair brush (#1) ✓

B. art supplies (#2)

QA-GNN Prediction

KG provides a scaffold for structured reasoning!

Benefit 2: Structured 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 (✗)

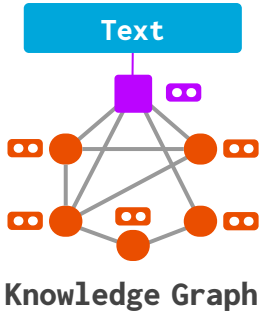
Takeaways

QA-GNN: combine KG and LM for general question answering

- Use LM to score and identify the relevant part of KG
- Jointly reason over LM and KG by using a GNN on text+KG joint graph

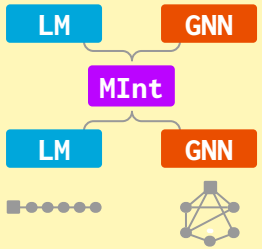
Ability to perform interpretable and structured reasoning

Outline



How to combine LM and KG for reasoning?

- **QAGNN**: Reasoning with Language Models and Knowledge Graphs for Question Answering [NAACL'21]

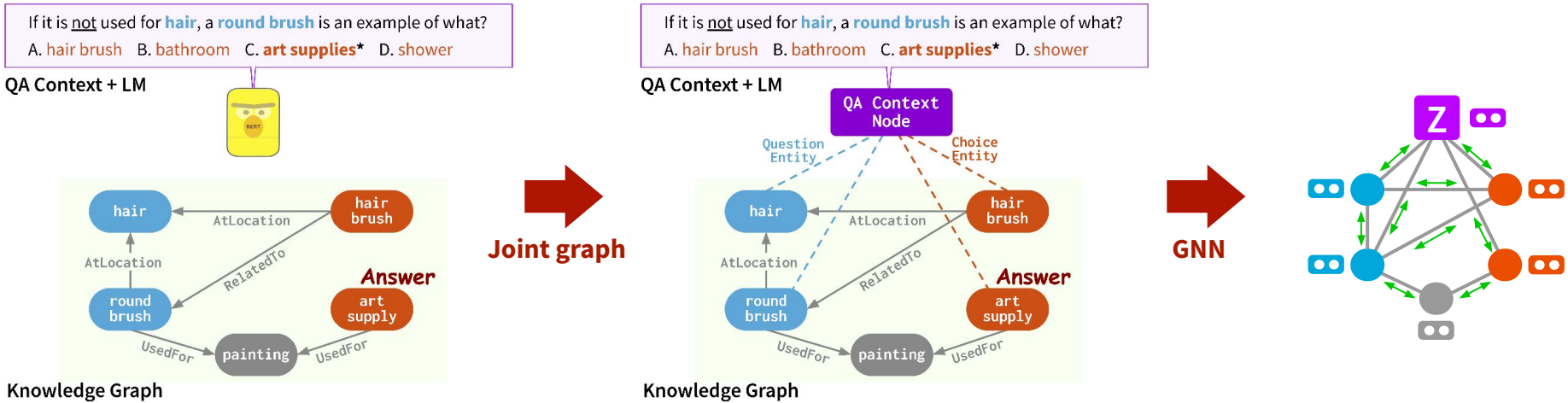


How to perform more expressive reasoning?

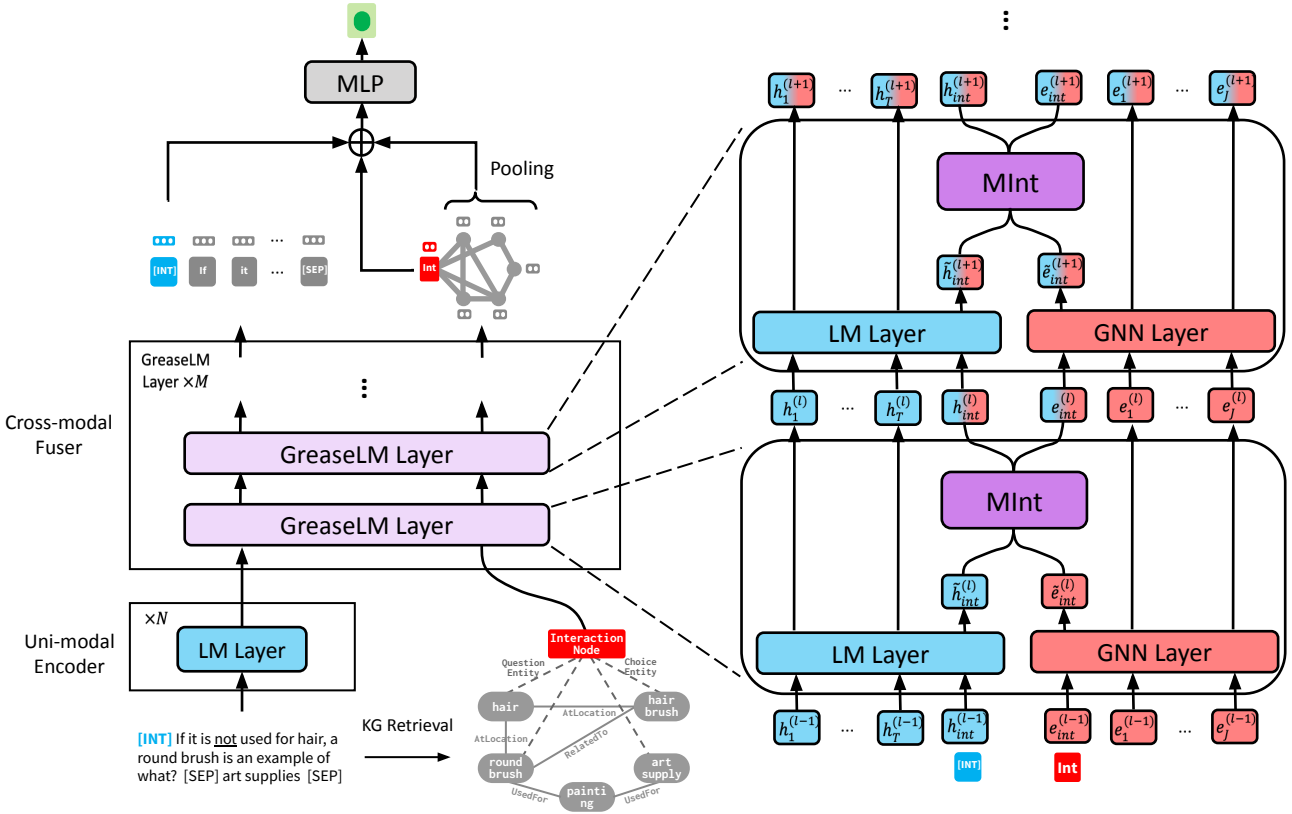
- **GreaseLM**: Graph Reasoning Enhanced Language Model for Question Answering [ICLR'22]

Limitation of QA-GNN

- QA-GNN uses a single pooled representation for the text
  - Cannot use KG information to update individual word representations in the text
- How to perform more expressive reasoning?



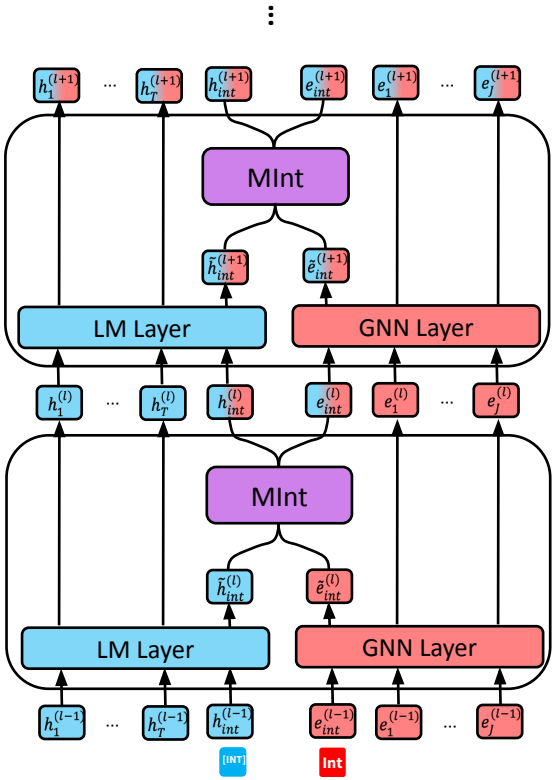
Our new model: GreaseLM



Our new model: GreaseLM

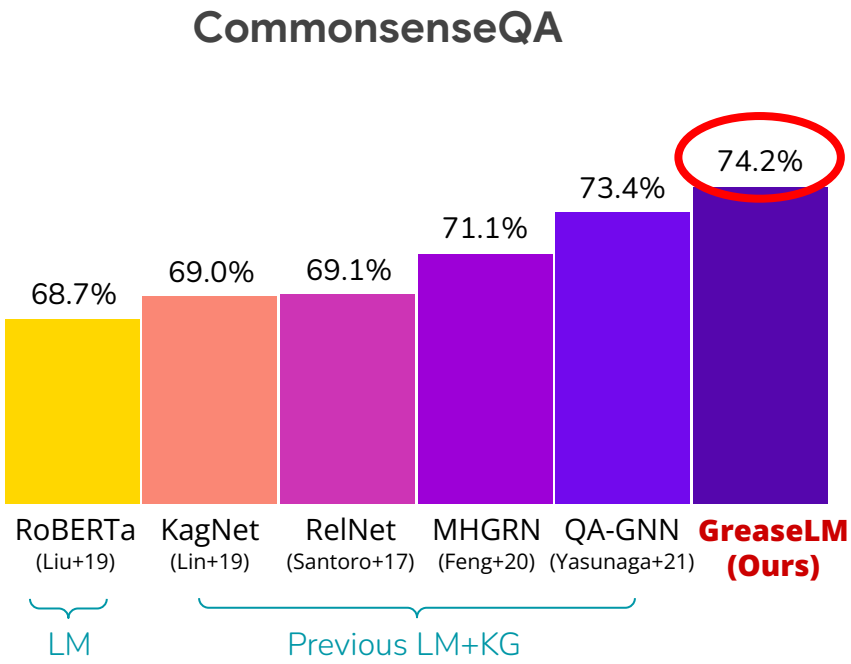
Key innovations

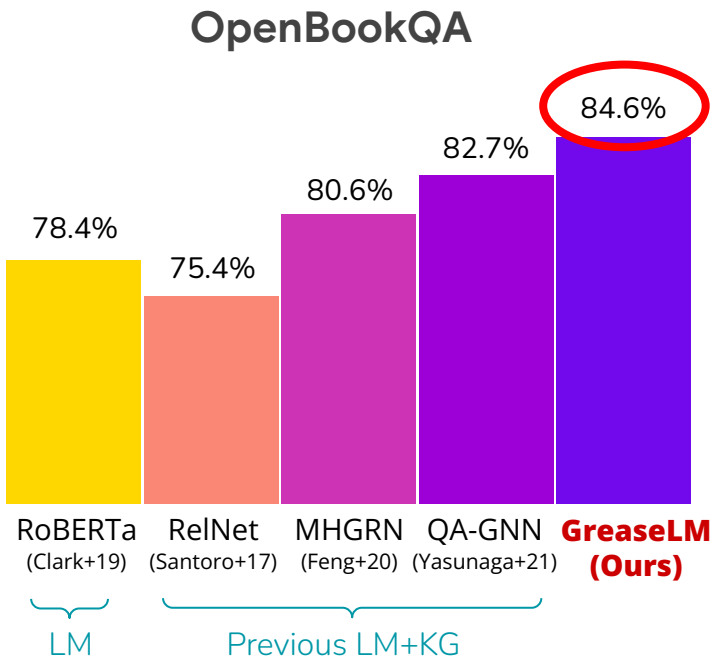
- Treat LM layer (over text) and GNN layer (over KG) at the equal level
- Modality interaction (MInt): Fuse and exchange information from LM and GNN for multiple layers
- Representations of all tokens in text and all nodes in KG are mutually updated



Performance

Improved performance on two QA tasks





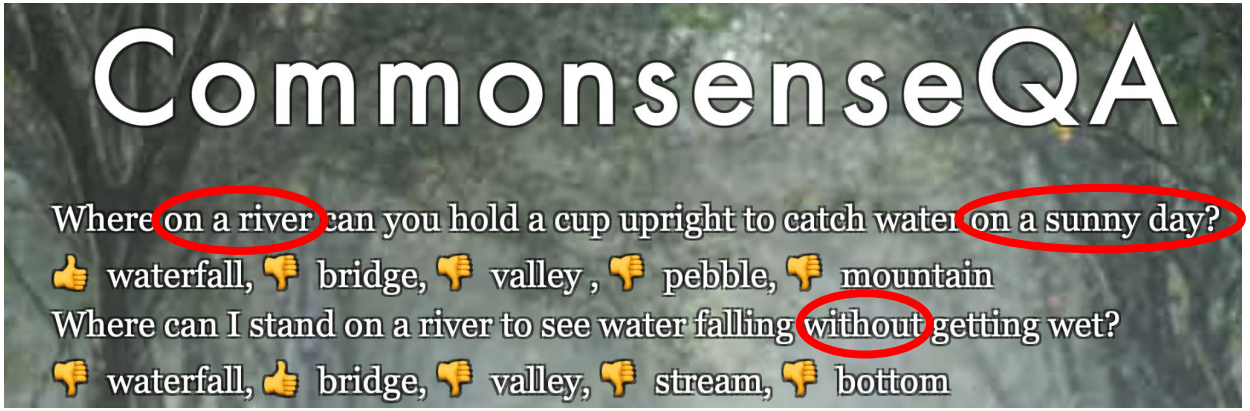
Ablation study

Ablation Type	Ablation	Dev Acc.
GREASELM	-	78.5
Modality Interaction	No interaction	76.5 ↓
	Interaction in every other layer	76.3 ↓
Interaction Layer Parameter Sharing	No parameter sharing	77.1
Graph Connectivity	Context node connected to all nodes in $\mathcal{V}_{\text{sub}}$, not only $\mathcal{V}_{\text{linked}}$	77.6
Node Initialization	Random	60.8
	TransE (Bordes et al., 2013)	77.7

Strength: Complex Reasoning

Questions requiring complex reasoning:

Prepositional phrases



Where would I not want a fox?
hen house, england, mountains,
english hunt, california

Negation terms

What is a place that usually does not have an elevator and that sometimes has a telephone book?
hotel, kitchen, library, telephone booth, house

Hedging terms

Strength: Complex Reasoning

GreaseLM solves various complex reasoning

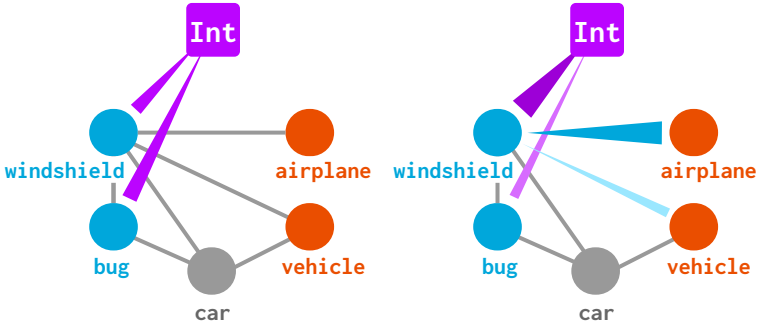
Model	# Prepositional Phrases					Negation Term	Hedge Term
	0	1	2	3	4		
RoBERTa-Large	66.7	72.3	76.3	74.3	69.5	64.6	69.5
QA-GNN	**76.7**	76.2	79.1	74.9	81.4	67.1	74.7
GREASELM (**Ours**)	75.7	**79.3**	**80.4**	**77.2**	**84.7**	**69.5**	**76.2**

Strength: Complex Reasoning

GreaseLM

What is unlikely to get **bugs** on its **windshield** due to **bugs'** inability to reach it when moving?

A. airplane ✅ B. motor vehicle

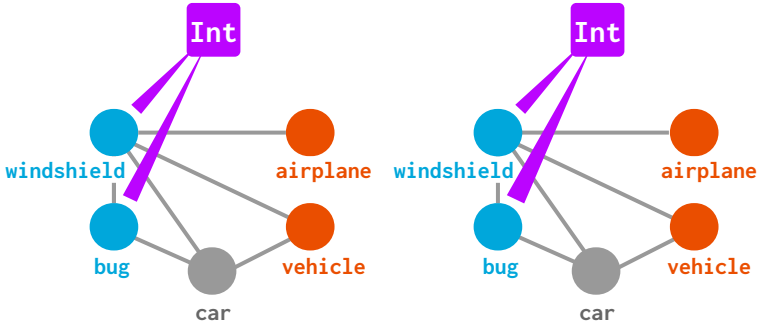


GNN **1st** Layer GNN **Final** Layer

QAGNN

What is unlikely to get **bugs** on its **windshield** due to **bugs'** inability to reach it when moving?

A. airplane **B. motor vehicle** ❌



GNN **1st** Layer GNN **Final** Layer

Attention weight from **text** over **bug** decreases, but attention weight over **windshield** and **airplane** increases, adjusting for hedging ("unlikely") in text.

Extension to Biomedical Reasoning

QA dataset: US Medical License Exam (USMLE)

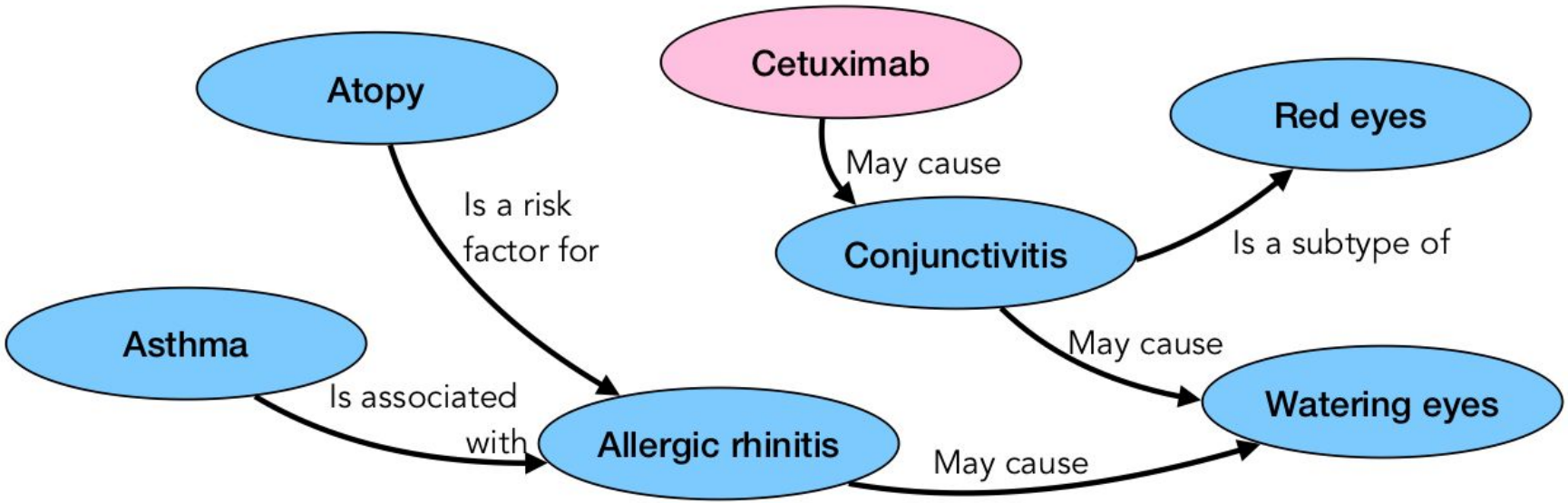
A 45-year-old woman presents to the emergency department with acute onset of severe right upper quadrant abdominal pain that radiates to the infrascapular region. Her medical history is significant for obesity, hypertension, obstructive sleep apnea, and gastric bypass surgery 2 years ago after which she lost 79 kg (150 lb). The patient complains of nausea and vomiting that accompanies the pain. Her temperature is 38.9°C (101.2°F), blood pressure is 144/88 mm Hg, heart rate is 76/min, and respiratory rate is 14/min (fever). Abdominal examination is significant for right upper quadrant tenderness along with guarding and cessation of inspired breath on deep palpation of the right upper quadrant.
Which test should be ordered first for this patient?

A) Abdominal ultrasound
B) CT scan of the abdomen
C) Hepato-iminodiacetic acid scan
D) MRI of the abdomen
E) X-ray film of the abdomen

Jin+2020

Extension to Biomedical Reasoning

Biomedical KG: DiseaseDatabase, DrugBank, UMLS



Bodenreider+2004, Wishart+2018

Extension to Biomedical Reasoning

Improved performance over LM and previous LM+KG

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**

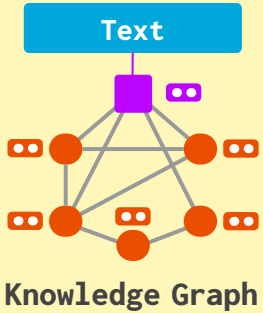
Takeaways

GreaseLM: improved architecture to perform expressive reasoning

- Treat LM (over text) and GNN (over KG) at the equal level
- Modality interaction (MInt): Fuse information from LM and GNN for multiple layers
- Representations of all tokens in text and all nodes in KG are mutually updated

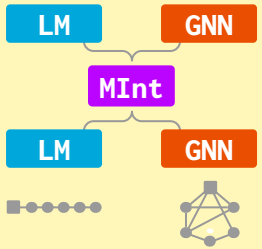
Ability to perform complex reasoning that requires both language understanding and knowledge (e.g. prepositional phrases, negation, hedging)

Outline



How to combine LM and KG for reasoning?

- **QAGNN**: Reasoning with Language Models and Knowledge Graphs for Question Answering [NAACL'21]



How to perform more expressive reasoning?

- **GreaseLM**: Graph Reasoning Enhanced Language Model for Question Answering [ICLR'22]

Thanks!


Michihiro Yasunaga


Xikun Zhang


Antoine Bosselut


Hongyu Ren


Percy Liang


Chris Manning


Jure Leskovec

Thank you to the members of the Stanford SNAP / NLP / P-Lambda groups, and the project MOWGLI team, as well as our anonymous reviewers. Funded in part by DARPA MCS.

Papers

- QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. NAACL 2021.
- GreaseLM: Graph Reasoning Enhanced Language Model. ICLR 2022.

Code: https://github.com/michiyasunaga/QAGNN; https://github.com/snap-stanford/GreaseLM

Website: https://snap.stanford.edu/QAGNN