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
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

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

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

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

Knowledge can be stored in:

- **Text & Pretrained Language Model (LM)**

- **Knowledge Graph (KG)**

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

[Bojacker+2008; Speer+2016]
(1) **Knowledge Graphs**

![Knowledge Graph Diagram]

**ConceptNet**
An open, multilingual knowledge graph

**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’, ‘population’, ‘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
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)

**Structured & Interpretable**
Knowledge Graph (KG)

[Wikipedia]

[ConceptNet]

[Freebase]

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

[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)
  - Complete Wikipedia and 11,038 books

- **Structured & Interpretable**
  - Knowledge Graph (KG)

- **WIKIPEDIA**

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

[ConceptNet]

[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

- 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

Retrieve k-hop neighbors/paths in KG

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

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

Retrieved KG

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
(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)
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

**Answer**
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**

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

**Knowledge Graph**
QA-GNN Message Passing

Message Passing

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

Attention \((s \rightarrow t)\)

Message \((s \rightarrow t)\)

Node type & relation-aware message

\[ m_{st} = f_m (h_{s}^{(\ell)}, u_s, r_{st}) \]

Node type, relation, & score-aware attention

\[ q_s = f_q (h_{s}^{(\ell)}, u_s, \rho_s) \]

\[ k_t = f_k (h_{t}^{(\ell)}, u_t, \rho_t, r_{st}) \]

\[ \alpha_{st} = \frac{\exp(\gamma_{st})}{\sum_{t' \in \mathcal{N}_s \cup \{s\}} \exp(\gamma_{st'})}, \quad \gamma_{st} = \frac{q_s^T k_t}{\sqrt{D}} \]

Node types
- QA Context
- Question entity
- Answer entity
- Other entity

25
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
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

**CommonsenseQA**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa (Liu+19)</td>
<td>68.7%</td>
</tr>
<tr>
<td>KagNet (Lin+19)</td>
<td>69.0%</td>
</tr>
<tr>
<td>RelNet (Santoro+17)</td>
<td>69.1%</td>
</tr>
<tr>
<td>MHGRN (Feng+20)</td>
<td>71.1%</td>
</tr>
<tr>
<td><strong>QA-GNN (Ours)</strong></td>
<td><strong>73.4%</strong></td>
</tr>
</tbody>
</table>

**OpenBookQA**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa (Clark+19)</td>
<td>78.4%</td>
</tr>
<tr>
<td>RelNet (Santoro+17)</td>
<td>75.4%</td>
</tr>
<tr>
<td>MHGRN (Feng+20)</td>
<td>80.6%</td>
</tr>
<tr>
<td><strong>QA-GNN (Ours)</strong></td>
<td><strong>82.7%</strong></td>
</tr>
</tbody>
</table>
# Ablation study

Performance drops to the score of previous LM+KG models

<table>
<thead>
<tr>
<th>Graph Connection</th>
<th>Dev Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>No edge between Z and KG nodes</td>
<td>74.81</td>
</tr>
<tr>
<td>Connect Z to all KG nodes</td>
<td>76.38</td>
</tr>
<tr>
<td>Connect Z to QA entity nodes (final)</td>
<td><strong>76.54</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relevance scoring</th>
<th>Dev Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nothing</td>
<td>75.56</td>
</tr>
<tr>
<td>w/ contextual embedding</td>
<td>76.31</td>
</tr>
<tr>
<td>w/ relevance score (final)</td>
<td><strong>76.54</strong></td>
</tr>
<tr>
<td>w/ both</td>
<td>76.52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GNN Attention &amp; Message (§3.3)</th>
<th>Dev Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node type, relation, score-aware (final)</td>
<td><strong>76.54</strong></td>
</tr>
<tr>
<td>- type-aware</td>
<td>75.41</td>
</tr>
<tr>
<td>- relation-aware</td>
<td>75.61</td>
</tr>
<tr>
<td>- score-aware</td>
<td>75.56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GNN Layers (§3.3)</th>
<th>Dev Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L = 3$</td>
<td>75.53</td>
</tr>
<tr>
<td>$L = 4$</td>
<td>76.34</td>
</tr>
<tr>
<td>$L = 5$ (final)</td>
<td><strong>76.54</strong></td>
</tr>
<tr>
<td>$L = 6$</td>
<td>76.21</td>
</tr>
<tr>
<td>$L = 7$</td>
<td>75.96</td>
</tr>
</tbody>
</table>
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.

- Question with ≤ 10 entities:
  - KG as sentences: 72.5%
  - QA-GNN: 73.4%

- Question with > 10 entities:
  - KG as sentences: 70.8%
  - QA-GNN: 73.5%

- All Question:
  - KG as sentences: 71.7%
  - QA-GNN: 73.4%

- Question with negation:
  - KG as sentences: 55.2%
  - QA-GNN: 58.8%
Benefit 1: Interpretability

Attention visualization direction: BFS from Q

Where would you find a **basement** that can be accessed with an **elevator**?  
A. closet  B. church  C. **office building***
Benefit 1: Interpretability

Attention visualization direction: $Q \rightarrow O$ and $A \rightarrow O$

**Crabs** live in what sort of environment?
A. saltwater*  B. galapagos  C. fish market
Benefit 2: Structured Reasoning

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

![Diagram showing GNN 1st Layer and GNN Final Layer predictions compared to QA-GNN Prediction](image)

- GNN 1st Layer: hair brush
- GNN Final Layer: hair brush
- QA-GNN Prediction: art supplies (1)
- RoBERTa Prediction: hair brush (1), art supplies (2)
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**  **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**

**QA-GNN Prediction**

---

**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**

**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.*

**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*

**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 1st Layer  GNN Final Layer  QA-GNN Prediction

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

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

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

---

* GNN stands for Graph Neural Network, which is a type of neural network that processes graph-structured data. In this context, GNNs are used to analyze relationships between different types of entities, such as hair and art supplies. The **1st Layer** and **Final Layer** refer to different stages of processing within the GNN framework. The **QA-GNN** prediction is a model that combines GNNs with question-answering techniques to answer questions about graph data.
**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*

<table>
<thead>
<tr>
<th>GNN 1st Layer</th>
<th>GNN Final Layer</th>
<th>QA-GNN Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>round brush</td>
<td>round brush</td>
<td>A. hair brush (1)</td>
</tr>
<tr>
<td>painting</td>
<td>painting</td>
<td>B. art supplies (2)</td>
</tr>
<tr>
<td>hair brush</td>
<td>hair brush</td>
<td>RoBERTa Prediction</td>
</tr>
<tr>
<td>art supply</td>
<td>art supply</td>
<td></td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>GNN Final Layer</th>
<th>QA-GNN Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>round brush</td>
<td>A. hair brush (1)</td>
</tr>
<tr>
<td>painting</td>
<td>B. art supplies (2)</td>
</tr>
<tr>
<td>hair brush</td>
<td></td>
</tr>
<tr>
<td>art supply</td>
<td></td>
</tr>
</tbody>
</table>

---

**KG provides a scaffold for structured reasoning!**
## Benefit 2: Structured Reasoning

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

<table>
<thead>
<tr>
<th></th>
<th>CommonsenseQA</th>
<th>OpenBookQA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LM</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoBERTa (Liu+19)</td>
<td>68.7%</td>
<td>78.4%</td>
</tr>
<tr>
<td>KagNet (Lin+19)</td>
<td>69.0%</td>
<td>75.4%</td>
</tr>
<tr>
<td>RelNet (Santoro+17)</td>
<td>69.1%</td>
<td>80.6%</td>
</tr>
<tr>
<td>MHGRN (Feng+20)</td>
<td>71.1%</td>
<td>82.7%</td>
</tr>
<tr>
<td>QA-GNN (Yasunaga+21)</td>
<td>73.4%</td>
<td></td>
</tr>
<tr>
<td>GreaseLM (Ours)</td>
<td><strong>74.2%</strong></td>
<td><strong>84.6%</strong></td>
</tr>
</tbody>
</table>

**Previous LM+KG**

<table>
<thead>
<tr>
<th></th>
<th>CommonsenseQA</th>
<th>OpenBookQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa (Clark+19)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RelNet (Santoro+17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHGRN (Feng+20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QA-GNN (Yasunaga+21)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GreaseLM (Ours)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**LM**
# Ablation study

<table>
<thead>
<tr>
<th>Ablation Type</th>
<th>Ablation</th>
<th>Dev Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GREASELM</strong></td>
<td>-</td>
<td>78.5</td>
</tr>
<tr>
<td><strong>Modality Interaction</strong></td>
<td>No interaction</td>
<td>76.5 ↓</td>
</tr>
<tr>
<td></td>
<td>Interaction in every other layer</td>
<td>76.3 ↓</td>
</tr>
<tr>
<td><strong>Interaction Layer Parameter Sharing</strong></td>
<td>No parameter sharing</td>
<td>77.1</td>
</tr>
<tr>
<td><strong>Graph Connectivity</strong></td>
<td>Context node connected to all nodes in $\mathcal{V}<em>{\text{sub}}$, not only $\mathcal{V}</em>{\text{linked}}$</td>
<td>77.6</td>
</tr>
<tr>
<td><strong>Node Initialization</strong></td>
<td>Random</td>
<td>60.8</td>
</tr>
<tr>
<td></td>
<td>TransE (Bordes et al., 2013)</td>
<td>77.7</td>
</tr>
</tbody>
</table>
Strength: Complex Reasoning

Questions requiring complex reasoning:

Prepositional phrases

Where on a river can you hold a cup upright to catch water on a sunny day?
👍 waterfall, 👎 bridge, 👎 valley, 👎 pebble, 👎 mountain

Where can I stand on a river to see water falling without getting wet?
👎 waterfall, 👍 bridge, 👎 valley, 👍 stream, 👎 bottom

Negation terms

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

Hedging 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
Strength: Complex Reasoning

GreaseLM solves various complex reasoning

<table>
<thead>
<tr>
<th>Model</th>
<th># Prepositional Phrases</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Negation Term</th>
<th>Hedge Term</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>Term</td>
<td>Term</td>
</tr>
<tr>
<td>RoBERTa-Large</td>
<td>66.7</td>
<td>72.3</td>
<td>76.3</td>
<td>74.3</td>
<td>69.5</td>
<td>64.6</td>
<td>69.5</td>
</tr>
<tr>
<td>QA-GNN</td>
<td>76.7</td>
<td>76.2</td>
<td>79.1</td>
<td>74.9</td>
<td>81.4</td>
<td>67.1</td>
<td>74.7</td>
</tr>
<tr>
<td>GREASELM (Ours)</td>
<td>75.7</td>
<td>79.3</td>
<td>80.4</td>
<td>77.2</td>
<td>84.7</td>
<td>69.5</td>
<td>76.2</td>
</tr>
</tbody>
</table>
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

---

**QAGNN**

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

A. airplane ✓ B. motor vehicle ❌

---

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
Extension to Biomedical Reasoning

Biomedical KG: DiseaseDatabase, DrugBank, UMLS
# Extension to Biomedical Reasoning

Improved performance over LM and previous LM+KG

<table>
<thead>
<tr>
<th>Methods</th>
<th>Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baselines</strong> (Jin et al., 2021)</td>
<td></td>
</tr>
<tr>
<td>Chance</td>
<td>25.0</td>
</tr>
<tr>
<td>PMI</td>
<td>31.1</td>
</tr>
<tr>
<td>IR-ES</td>
<td>35.5</td>
</tr>
<tr>
<td>IR-Custom</td>
<td>36.1</td>
</tr>
<tr>
<td>CLINICALBERT-BASE</td>
<td>32.4</td>
</tr>
<tr>
<td>BIOROBERTA-BASE</td>
<td>36.1</td>
</tr>
<tr>
<td>BIOBERT-BASE</td>
<td>34.1</td>
</tr>
<tr>
<td>BIOBERT-LARGE</td>
<td>36.7</td>
</tr>
<tr>
<td><strong>Baselines</strong> (Our implementation)</td>
<td></td>
</tr>
<tr>
<td>SapBERT-Base (w/o KG)</td>
<td>37.2</td>
</tr>
<tr>
<td>QA-GNN</td>
<td>38.0</td>
</tr>
<tr>
<td><strong>GREASELM (Ours)</strong></td>
<td><strong>38.5</strong></td>
</tr>
</tbody>
</table>
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

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