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

NAACL 2021

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, Jure Leskovec
Stanford University
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**
Question Answering with Knowledge

Question

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

Knowledge sources

Pre-trained Language Model (LM)

[Devlin+19; Liu+19; Brown+20; ...]
Question Answering with Knowledge

**Question**

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

**Knowledge sources**

Pre-trained Language Model (LM)

Knowledge Graph (KG)

[Devlin+19; Liu+19; Brown+20; ...]

[Bollacker+08; Speer+16]
Goal: Leverage Both Knowledge Sources

Knowledge sources

Pre-trained Language Model (LM)

Knowledge Graph (KG)

Broad Coverage

Structured & Interpretable

Complete Wikipedia and 11,038 books

[Devlin+19; Liu+19; Brown+20; ...]

[Bollacker+08; Speer+16]
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
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:** C. art supplies

**QA Context + LM**

**Knowledge Graph**

- **hair** → AtLocation → **hair brush**
- **round brush** → AtLocation → **hair brush**
- **round brush** → RelatedTo → **art supply**
- **painting** → UsedFor → **art supply**
- **painting** → UsedFor → **art supply**

*Note: The asterisk (*) indicates the correct answer.*
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**

- **Answer**

- How to identify relevant KG subset?
- How to **jointly reason** over the text and KG?
Overview of Our Approach: QA-GNN

Idea

QA context $[q; a]$
Overview of Our Approach: QA-GNN

Idea

QA context: \([q; a]\)

- LM Encoding
- KG Retrieval

QA-GNN
Overview of Our Approach: QA-GNN

Idea
1) Language-conditioned KG node relevance scoring
Overview of Our Approach: QA-GNN

Idea
1) Language-conditioned KG node relevance scoring
2) Joint Reasoning:
   Connect text and KG to form a joint graph (*working graph*)
Overview of Our Approach: QA-GNN

Idea

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)
Overview of Our Approach: QA-GNN

Idea
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)
Overview of Our Approach: QA-GNN

Idea
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)
Overview of Our Approach: QA-GNN

Idea
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)
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
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

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

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

Retrieve k-hop neighbors/paths in KG
Existing KG Retrieval

**QA Context**

A **reversing 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 neighbors/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
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

Some entities are irrelevant to the given 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

Our Solution: Language-conditioned KG Node Scoring

Some entities are irrelevant to the given 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

Our Solution: Language-conditioned KG Node Scoring

Some entities are irrelevant to the given 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

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
   → Can improve model efficiency (time/space)

Option 2. Incorporate as auxiliary feature of KG node
   → General way to weight information on KG. We’ll do this in our GNN
Overview of Our Approach: QA-GNN

Idea
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)
Overview of Our Approach: QA-GNN

Idea
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: C. art supplies
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
QA-GNN Message Passing

Message passing
Message passing

\[ h_{t}^{(\ell+1)} = f_n \left( \sum_{s \in N_t \cup \{t\}} \alpha_{st} m_{st} \right) + h_{t}^{(\ell)} \]
QA-GNN Message Passing

Message passing

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

Attention (s → t)

Message (s → t)

Node type & relation-aware message

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

Node types

- QA Context
- Question entity
- Answer entity
- Other entity
QA-GNN Message Passing

Message passing

\[ h_t^{(\ell+1)} = f_n \left( \sum_{s \in N_t \cup \{t\}} \alpha_{st} m_{st} \right) + h_t^{(\ell)} \]

Attention (s → t)
Message (s → t)

Node type & relation-aware message

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

Node type, relation, & score-aware attention

Node types
- QA Context
- Question entity
- Answer entity
- Other entity
QA-GNN Message Passing

Message passing

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

Attention (s → t)
Message (s → t)

Node type & relation-aware message

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

Node type, relation, & score-aware attention

\[ q_s = f_q(h_{s}^{(\ell)}, u_s, \rho_{s}) \]
\[ k_t = f_k(h_{t}^{(\ell)}, u_t, \rho_t, r_{st}) \]

Node types
- QA Context
- Question entity
- Answer entity
- Other entity
QA-GNN Message Passing

Message passing

\[ h_t^{(\ell+1)} = f_n \left( \sum_{s \in N_t \cup \{t\}} \alpha_{st} m_{st} \right) + h_t^{(\ell)} \]

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

\[ q_s = f_q(h_s^{(\ell)}, u_s, \rho_s) \]
\[ k_t = f_k(h_t^{(\ell)}, u_t, \rho_t, r_{st}) \]

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

Node types
- QA Context
- Question entity
- Answer entity
- Other entity
Overview of Our Approach: QA-GNN

Idea
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)
Experimental Setup

QA Datasets

- **CommonsenseQA**: reasoning with commonsense knowledge
  
  Train / Dev / Test  8500 / 1221 / 1241

- **OpenBookQA**: reasoning with elementary science knowledge
  
  Train / Dev / Test  4957 / 500 / 500

---

What do people typically do while playing guitar?

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

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

[Talmor+18]  
[Mihaylov+18]
Experimental Setup

Knowledge graph:

ConceptNet (English) \([\text{Liu+04; Speer+16}]\)

~800,000 nodes
17 relation types

KG processing

- Link entities in the question to KG
- Extract a subgraph of 2-hop paths
Experimental Setup

Compared systems

- Fine-tuned LM
  - RoBERTa [Liu+19]

- Previous LM+KG models
  - RelNet [Santoro+17]
  - KagNet [Lin+19]
  - MHGRN [Feng+20]

**Difference: QA-GNN**

1. Uses the KG node score (relevance given the question)
2. Mutually updates the LM and KG representations
**Results**

Improved performance on two QA tasks

- **CommonsenseQA**: reasoning with commonsense knowledge

  - RoBERTa (Liu+19)
  - KagNet (Lin+19)
  - RelNet (Santoro+17)
  - MHGRN (Feng+20)
  - QA-GNN (Ours)

  - 68.7%
  - 69.0%
  - 69.1%
  - 71.1%
  - 73.4%

- **OpenBookQA**: reasoning with elementary science knowledge

  - RoBERTa (Liu+19)
  - GconAttn (Wang+19)
  - RelNet (Santoro+17)
  - MHGRN (Feng+20)
  - QA-GNN (Ours)

  - 64.8%
  - 64.8%
  - 65.2%
  - 66.80%
  - 70.5%
## Ablation Study

<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>

*Close to the previous LM+KG models*
When is KG Node Scoring or Joint Graph Helpful?

- Node scoring tends to help when retrieved KG is big

<table>
<thead>
<tr>
<th>Question with ≤ 10 entities</th>
<th>No KG score</th>
<th>QA-GNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>72.8%</td>
<td>73.4%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question with &gt; 10 entities</th>
<th>No KG score</th>
<th>QA-GNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>71.5%</td>
<td>73.5%</td>
<td></td>
</tr>
</tbody>
</table>
When is KG Node Scoring or Joint Graph Helpful?

- Node scoring tends to help when retrieved KG is big
  - Question with ≤ 10 entities: 72.8% vs. 73.4%
  - Question with > 10 entities: 71.5% vs. 73.5%

- Joint graph tends to help when question has negation
  - All Question: 71.5% vs. 73.4%
  - Question with negation: 55.5% vs. 58.8%
Case Study: Interpreting Reasoning Patterns

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***

*Correct answer.
Case Study: Interpreting Reasoning Patterns

Attention visualization direction: \( Q \rightarrow O \) and \( A \rightarrow O \)

**Crabs** live in what sort of environment?
A. **saltwater***  B. **galapagos**  C. **fish market**
Case Study: Robust Reasoning (Negation)

Original Question

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

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

QA-GNN Prediction
A. hair brush (#2)
B. art supplies (#1)
Case Study: Robust Reasoning (Negation)

**Original Question**

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

- A. hair brush
- B. art supplies*

**Negation Flipped**

If it is 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

RoBERTa Prediction

A. hair brush (#1)
B. art supplies (#2)
If it is not used for hair, a round brush is an example of what?

A. hair brush  B. art supplies

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

A. hair brush  B. art supplies
Case Study: Robust Reasoning (More Examples)

<table>
<thead>
<tr>
<th>Example (Original taken from CommonsenseQA Dev)</th>
<th>RoBERTa Prediction</th>
<th>Our Prediction</th>
</tr>
</thead>
</table>
| [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?)  
B. art supply (✓) | 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?)  
B. art supply (✓) | 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 (✓) | A. interested (✓) |
| [Negation ver 2] If you have to read a book that is **not** dry you may become what? | B. bored (✗)  
A. interested (✓) | 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 (✗) | A. interested (✗) |
Takeaways

Problem: Question answering using language models and knowledge graphs

Innovations of QA-GNN

1. **Use language models to score KG node** relevance conditioned on a question. A general method to weight information on KG (“soft” retrieval or pruning)
2. **Joint Reasoning with LM and KG**: connect the LM (text) and KG to form a joint graph, and mutually update their representations with GNN

Insights

- Node scoring: tends to help when retrieved KG is big and noisy
- Joint LM-KG message passing: tends to help when the question requires rich reasoning (e.g. negation)
Thanks!

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

Funded in part by DARPA and Funai Foundation Fellowship.

Code: https://github.com/michiyasunaga/QAGNN
Website: https://snap.stanford.edu/QAGNN