Graph-Based Reasoning over Heterogeneous External Knowledge for Commonsense Question Answering

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Songlin Hu

Published: AAAI 2020

Presenter: Anushree Hede
April 20, 2020
Motivation

Combining evidence from ConceptNet & Wikipedia gives the option C.

Commonsense QA
- Collect background knowledge and reason over it

Structured KBs: relations beneficial for reasoning
- But low coverage is an issue

Unstructured text: abundant coverage
Contributions

• **Main**: Combine heterogeneous knowledge sources together into the same representation space
• **Graph modules to leverage structure for reasoning**
  – Context representation learning module
  – Inference module
• **New state-of-the-art performance**: 75.3%
Contents

• Overview of Approach
• Heterogeneous Knowledge Extraction
• Graph-Based Modules
• Experiments and Results
• Conclusions
• Related Work
• Issues
• Discussion
Problem: Overview

- **Dataset:**
  - CommonsenseQA [1]
  - Questions lack evidence, rely on background knowledge

- **Evaluation:**
  - Accuracy
  - Ablation Study
  - Error Analysis

Output option $o_i$

Graph-Based Reasoning

Knowledge Extraction

Question $Q = \{q_1, \ldots, q_m\}$
Answer options $A = \{a_i\}$

For all $i = 1$ to $5$
Knowledge Extraction ConceptNet → Concept-Graph

- Commonsense Knowledge Base
- Locate and search for path from question entities → answer choice entities (< 3 hops)
- Merge triples as nodes in graph
  - Edge from $s_i$ to $s_j$ if they contain same entity
- Convert triples to natural language sentences
Knowledge Extraction \( \text{Wikipedia} \rightarrow \text{Wiki-Graph} \)

- Top 10 Wiki sentences from Elastic Search for (question + choices)
- Semantic Role Labeling: Nodes are **subject**, **predicate**, **object**
- Edges:
  - (subject, predicate)
  - (predicate, object)
  - Node A is contained in node B and the \#words(A) > 3
  - Node A and node B only have one different word and \#words(A) and \#words(B) > 3
Graph-Based Reasoning

- Evidence
  - Concept-Graph
  - Wiki-Graph
- Context Representation Learning
- Inference
  - Graph Convolutional Network
  - Graph Attention
- Output
• If $p \in s_i$, $q \in s_j$ and $(p,q)$ is an edge in Wiki-Graph, then $(s_i, s_j)$ is an edge in sentence.

• **Topological sort** on Concept-Graph & sentence graph

• Goal: Shorten distance between semantically similar nodes
Contextual Representation Learning Module

- **XLNet:** captures long term dependencies

  ![Diagram](image)

  - Topologically sorted sentences from Concept-Graph and Sentence graph
  - Question
  - Answer choice $a_i$

- **Goal:**
  - Obtain better contextual word representations
  - Fuse two knowledge sources in same representation space
Algorithm 1 Topology Sort Algorithm.

Require: A sequence of nodes \( S = \{s_1, s_2, \ldots, s_n\} \); A set of relations \( R = \{r_1, r_2, \ldots, r_m\} \).

1: function DFS(node, visited, sorted_sequence)
2:     for each child \( s_c \) in node’s children do
3:         if \( s_c \) has no incident edges and visited[\( s_c \)] == 0 then
4:             visited[\( s_c \)] = 1
5:             sorted_sequence.append(0, \( s_c \))
6:             Remove the incident edges of \( s_c \)
7:             DFS(\( s_c \), visited, sorted_sequence)
8:         end if
9:     end for
10: end function
11: sorted_sequence = []
12: visited = [0 for i in range(n)]
13: S,R = to_acyclic_graph(S,R)
14: for each node \( s_i \) in \( S \) do
15:     if \( s_i \) has no incident edges and visited[i] == 0 then
16:         visited[i] = 1
17:         sorted_sequence.append(\( s_i \))
18:         DFS(\( s_i \), visited, sorted_sequence)
19: end if
20: end for
21: return sorted_sequence
Inference Module

- **Graph Convolutional Networks (GCNs)**
  - Use Concept-Graph and Wiki-Graph
  - Update graph node representations using features of neighboring nodes

- The $i^{th}$ node representation in layer 0

\[
    h_i^0 = \sigma(W \sum_{w_j \in s_i} \frac{1}{|s_i|} h_{w_j}) \tag{1}
\]

- **Subsequent layers**

\[
    z_i^l = \sum_{j \in N_i} \frac{1}{|N_i|} V^l h_j^l, \tag{2}
\]

\[
    h_i^{l+1} = \sigma(W^l h_i^l + z_i^l). \tag{3}
\]
Inference Module

Graph Attention (multiplicative)
- Attention function: alignment score between \(<\text{cls}>\) and final GCN representation of \(i^{th}\) node
- Aggregate over all nodes of graph
- Obtain normalized score, compare across options

\[
\alpha_i = \frac{h_i^c \sigma(W_1 h_i^L)}{\sum_{j \in N} h_j^c \sigma(W_1 h_j^L)},
\]

\[
h^g = \sum_{j \in N} \alpha_j^L h_j^L.
\]

Q: Animals who have hair and don’t lay eggs are what?
A: Mammals

Correct option = \(\arg\max_{a \in A} p(q, a)\)

Normalized scoring
\[
score(q, a) = \text{MLP}(h_g, h_c)
\]

\[
p(q, a) = \frac{e^{score(q, a)}}{\sum_{a' \in A} e^{score(q, a')}}.
\]

Importance of node \(i\)

\[
\text{ConceptNet}
\]

- mammals is a kind of animals
- mammals has hair
- animals has fur
- very few mammals

\[
\text{Wikipedia}
\]

- laying eggs
- on eastern towhees

\[
\text{node attention weight}
\]

\[
\begin{array}{c|c}
\text{mammals} & 0.17 \\
\text{mammals has hair} & 0.11 \\
\text{animals has fur} & 0.06 \\
\text{very few mammals} & 0.27 \\
\text{laying eggs} & 0.17 \\
\text{on eastern towhees} & 0.05 \\
\end{array}
\]
Experiments

<table>
<thead>
<tr>
<th>Group</th>
<th>Model</th>
<th>Dev Acc</th>
<th>Test Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>SGN-lite</td>
<td>-</td>
<td>57.1</td>
</tr>
<tr>
<td></td>
<td>BECON (single)</td>
<td>-</td>
<td>57.9</td>
</tr>
<tr>
<td></td>
<td>BECON (ensemble)</td>
<td>-</td>
<td>59.6</td>
</tr>
<tr>
<td></td>
<td>CSR-KG</td>
<td>-</td>
<td>61.8</td>
</tr>
<tr>
<td></td>
<td>CSR-KG (AI2 IR)</td>
<td>-</td>
<td>65.3</td>
</tr>
<tr>
<td>Group 2</td>
<td>BERT-large</td>
<td>-</td>
<td>56.7</td>
</tr>
<tr>
<td></td>
<td>XLNet-large</td>
<td>-</td>
<td>62.9</td>
</tr>
<tr>
<td></td>
<td>RoBERTa(single)</td>
<td>78.5</td>
<td>72.1</td>
</tr>
<tr>
<td></td>
<td>RoBERTa(ensemble)</td>
<td>-</td>
<td>72.5</td>
</tr>
<tr>
<td>Group 3</td>
<td>KagNet</td>
<td>-</td>
<td>58.9</td>
</tr>
<tr>
<td></td>
<td>BERT + AMS</td>
<td>-</td>
<td>62.2</td>
</tr>
<tr>
<td></td>
<td>RoBERTa + CSPT</td>
<td>76.2</td>
<td>69.6</td>
</tr>
<tr>
<td>Group 4</td>
<td>Cos-E</td>
<td>-</td>
<td>58.2</td>
</tr>
<tr>
<td></td>
<td>BERT + OMCS</td>
<td>68.8</td>
<td>62.5</td>
</tr>
<tr>
<td></td>
<td>HyKAS</td>
<td>-</td>
<td>62.5</td>
</tr>
<tr>
<td></td>
<td>AristoBERTv7</td>
<td>-</td>
<td>64.6</td>
</tr>
<tr>
<td></td>
<td>DREAM</td>
<td>73.0</td>
<td>66.9</td>
</tr>
<tr>
<td></td>
<td>RoBERT + KE</td>
<td>77.5</td>
<td>68.4</td>
</tr>
<tr>
<td></td>
<td>RoBERTa + CSPT</td>
<td>76.2</td>
<td>69.6</td>
</tr>
<tr>
<td></td>
<td>RoBERTa + IR</td>
<td>78.9</td>
<td>72.1</td>
</tr>
<tr>
<td></td>
<td>Our Model</td>
<td>79.3</td>
<td>75.3</td>
</tr>
</tbody>
</table>
Ablation Studies

Components of graph-based reasoning

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLNet + E</td>
<td>75.8</td>
</tr>
<tr>
<td>XLNet + E + Topology Sort</td>
<td>77.7</td>
</tr>
<tr>
<td>XLNet + E + Graph Inference</td>
<td>77.2</td>
</tr>
<tr>
<td>XLNet + E + Topology Sort + Graph Inference</td>
<td>79.3</td>
</tr>
</tbody>
</table>

Heterogenous knowledge sources

<table>
<thead>
<tr>
<th>Knowledge Sources</th>
<th>Dev Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>68.9</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>75.3</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>73.5</td>
</tr>
<tr>
<td>ConceptNet + Wikipedia</td>
<td>79.3</td>
</tr>
</tbody>
</table>

- Topology sort change the relative position between words for better contextual word representation
- GCN and graph attention can aggregate both word and node representations to infer answers
- Both together: complementary

- None: XLNet large model
- Both sources individually bring about improvement
- Combining both: much larger benefit
Conclusion

- Knowledge Extraction into graphs
  - ConceptNet (structured)
  - Wikipedia (unstructured)

- Graph-based reasoning
  - Contextual word representation learning module (Top. Sort + XLNet)
  - Inference module (GCN + Attention)

- State-of-the-art performance: 75.3%

- Graph structure of evidence sentences: basis for reasoning in commonsense question answering task
Issues

• Opening example in paper:
  – Claim: “Dataset built in a way that answer choices share the same relation with question concept”

• Semantic Role Labeling: typing errors
  – “Subjective” refers to → “subject”
  – “Objective” refers to → “object”

• Wiki-Graph example
  – “Node A is contained in node B and the #words (A) > 3”

• Uses only entities in question to extract knowledge
  – Replacing “typically” with “never” would not change Concept-Graph, rely only on Wiki-Graph

• Removal of stopwords during Wikipedia (Elastic Search)
  – Words like “not” would be skipped, this would give opposite results
  – BERT-Large baseline can’t deal with negation either [1]

• Robustness: case studies of failed examples absent

Does not reflect well
Discussion

• Error Analysis (in paper): extracted evidence lack answer; two options too similar
• Limitations (opinion) for other graph-based reasoning (not commonsense)
  ➢ Question Answering via Integer Programming over Semi-Structured Knowledge [3]
  ➢ Question Answering as Global Reasoning over Semantic Abstractions [4]
  • Support graph mathematically rigorous than Concept/Wiki graphs
• Both use structure of graph to formulate ILP problem
• XLNet representations vs. ILP
  • Pre-trained models perhaps perform better, but representations/constraints not explainable
• Using SRL for unstructured → structured knowledge: important advantage
• Does it address limitations of those papers?
  • Reasoning fails to exploit requisite knowledge from graph ❌
  • Natural language modules fail to represent the underlying phenomena of context ❌
References


