Commonsense Reasoning with Implicit Knowledge in Natural Language

*Pratyay Banerjee, *Swaroop Mishra, *Kuntal Pal,
*Arindam Mitra, Chitta Baral
Can we answer commonsense questions regarding Abductive, Social Interactions and Physical Interactions?
Can we improve our answers with implicit knowledge in natural language form?
How can we find relevant knowledge?
How to reason with retrieved implicit relevant knowledge?
What kind of training strategies can we use?
How do they compare with respect to large / complex models?
Knowledge Retrieval Pipeline

- **Knowledge Source**
  - aNLI $\longrightarrow$ *ROC Stories + Story Cloze* (Directly Derived)
  - SIQA $\longrightarrow$ *Atomic* (Partially Derived)
  - PIQA $\longrightarrow$ *Wikihow* (Relevant)
Knowledge Infusion Modes

- The top 10 knowledge statements ($K_{ij}$) retrieved for each options ($A_i$) are infused with the transformer encoders in four different ways:

  - **CONCAT**: $[CLS] K_i [SEP] Q A_i [SEP]$  
  - **PARALLEL-MAX**: $[CLS] K_{ij} [SEP] Q A_i [SEP] → \max (CLS \text{ for each } A_i)$  
  - **SIMPLE-SUM**: $[CLS] K_{ij} [SEP] Q A_i [SEP] → \sum (CLS \text{ for each } A_i)$  
  - **WEIGHTED-SUM**: $[CLS] K_{ij} [SEP] Q A_i [SEP] → \text{wtd-sum (CLS for each } A_i)$
Training Strategies

- **REVISION**
  - Pre-Trained (MLM+NSP [BERT] and MLM [RoBERTa]) on $K_D$
  - Fine-Tuned on D

- **OPENBOOK**
  - Fine-Tuned on D+S, S is a subset of $K_D$

- **REVISION+OPENBOOK**
  - Pre-Trained (MLM+NSP [BERT] and MLM [RoBERTa]) on $K_D$
  - Fine-Tuned on D+S, S is a subset of $K_D$

$K_D$ - Respective knowledge sources for given 3 datasets D
Flow-Diagram of our Weighted-Sum Model

A sample from PIQA dataset
Results: Training Strategies with Knowledge infusion modes

- Revision with Openbook strategy works best across 3 datasets for each of 4 knowledge infusion modes
- Weighted sum model shows the best performance across all datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Strategy</th>
<th>BERT</th>
<th>RoBERTa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Concat</td>
<td>Max</td>
</tr>
<tr>
<td>aNLI</td>
<td>OPENBOOK</td>
<td>73.9±0.8</td>
<td>74.7±0.1</td>
</tr>
<tr>
<td></td>
<td>REVISION</td>
<td>72.7±0.3</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>REVISION &amp; OPENBOOK</td>
<td>74.4±0.2</td>
<td>74.3±0.1</td>
</tr>
<tr>
<td>PIQA</td>
<td>OPENBOOK</td>
<td>67.8±0.4</td>
<td>72.4±0.6</td>
</tr>
<tr>
<td></td>
<td>REVISION</td>
<td>74.5±0.3</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>REVISION &amp; OPENBOOK</td>
<td>67.7±0.1</td>
<td>73.8±0.8</td>
</tr>
<tr>
<td>SIQA</td>
<td>OPENBOOK</td>
<td>70.1±0.8</td>
<td>67.8±0.1</td>
</tr>
<tr>
<td></td>
<td>REVISION</td>
<td>69.5±0.9</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>REVISION &amp; OPENBOOK</td>
<td>68.8±0.4</td>
<td>66.6±0.4</td>
</tr>
</tbody>
</table>
Results: Weighted Sum model Performance Comparison

- WS with BERT and RoBERTa is better than baseline BERT and RoBERTa
- Better than huge parameterized model (T5 in aNLI)
- WS shows better performance than complex Graph based models like KagNet or GBR
Analysis

- For PIQA we find most of the errors are due to distracting knowledge.
- Over 50% of the correct predictions for each of the 3 datasets have implicit knowledge.

<table>
<thead>
<tr>
<th>Types of Error</th>
<th>aNLI</th>
<th>SIQA</th>
<th>PIQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotation</td>
<td>41%</td>
<td>38%</td>
<td>10%</td>
</tr>
<tr>
<td>Model Prediction</td>
<td>48%</td>
<td>27%</td>
<td>29%</td>
</tr>
<tr>
<td>Distracting Knowledge</td>
<td>11%</td>
<td>35%</td>
<td>61%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Knowledge</th>
<th>aNLI</th>
<th>SIQA</th>
<th>PIQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicitly Present</td>
<td>14%</td>
<td>11%</td>
<td>10%</td>
</tr>
<tr>
<td>Implicitly Present</td>
<td>55%</td>
<td>59%</td>
<td>51%</td>
</tr>
<tr>
<td>Fully Irrelevant</td>
<td>31%</td>
<td>30%</td>
<td>39%</td>
</tr>
</tbody>
</table>

Error Percentage

Knowledge Classification - Correct Predictions
Analysis of the Weighted Sum model

- We show the impact of increasing knowledge sentences on the accuracy with Revision training strategy for both BERT and RoBERTa
  - It increases for SIQA and PIQA
- We also show the weights learned by the weight layer of the WS model
  - We find that the model learns lower weights where there is very less lexical overlap between knowledge retrieved and Question and Answer options
Conclusion

In this paper through 3 commonsense reasoning dataset,

- We perform a thorough analysis of transformers’ (BERT and RoBERTa) ability to reason.
- We present four modes of knowledge infusion in transformer encoders which improves 2-9% of accuracy across the datasets.
- We carry out an extensive investigation to study the impact of different knowledge sources and pre-training on such knowledge sources.
Thank You !!!

Contacts：Pratyay Banerjee，Swaroop Mishra，Kuntal Pal，Arindam Mitra，Chitta Baral
Paper：https://openreview.net/pdf?id=a4-fFL7aCi0
Email：{pbanerj6, srmishr1, kkpal, chitta}@asu.edu, arindam.mitra@microsoft.com