Overview

Can we let models make less mistakes?

Q: How many legs can an easel have?

QA Model

We propose a new method that substantially enhances model’s commonsense reasoning performance.

Simple Effective Generalizable

knowledge base special supervision

Main idea: Prompt an existing QA model with model-generated background knowledge.

Step 1: Knowledge Generation

• We develop a task-specific prompt.
  • For a new question, we plug it into the prompt and sample multiple knowledge statements from GPT-3’s text continuation to the prompt:
    \[ K(q) = \{ k_m : k_m \sim p_G(k | \text{prompt}, q) \} \]
  • where \( q \) is the question, \( K(q) \) is the set of knowledge statements, and \( p_G(\cdot) \) is the generative distribution given by GPT-3 and decoding with nucleus sampling.

Step 2: Inference with Knowledge Prompting

• Each knowledge statement individually prompts the QA model.
  • We collect the confidence scores of choices in all model passes.
  • Aggregated score for choice \( a \):
    \[ P_a = \max_{1 \leq m \leq M} P_{a,m} \]
  • Final prediction:
    \[ \hat{a} = \max_{1 \leq a \leq A} P_a \]

Experimental Results

Tasks: NumerSense, CommonsenseQA, CommonsenseQA 2.0, QASC

Main Results

• New SOTA among non-retrieval methods.
• Comparable with or outperforms retrieval methods.

Ablations

Three key factors in the success of our method: knowledge quantity, knowledge quality, aggregation method.

Human Evaluation

Most knowledge are helpful to the QA model in a human-interpretable way.

Qualitative examples

Our generated knowledge support various reasoning procedures: induction, deduction, abduction, analogy, etc.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Question / Knowledge</th>
<th>Prediction</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>NumerSense</td>
<td>an easel can have [M] or four legs.</td>
<td></td>
<td>Commonsense Deduction</td>
</tr>
<tr>
<td>CommonsenseQA</td>
<td>The master of a master is a master.</td>
<td></td>
<td>Commonsense Deduction</td>
</tr>
<tr>
<td>CommonsenseQA 2.0</td>
<td>I did not need a servant, I was not a what?</td>
<td></td>
<td>Commonsense Abduction</td>
</tr>
<tr>
<td>QASC</td>
<td>[M] is used for transportation.</td>
<td></td>
<td>Commonsense Analogy</td>
</tr>
</tbody>
</table>

Code/Data

https://github.com/liujch1998/GKP

Large pre-trained language models as source of flexible, high-quality knowledge