P6-25

自己認知は LM as KB の信頼性を高めるか

RebelsNLU JAIST

○井之上直也 (JAIST/理研), 原口大地, 田中健史朗, 白井清昭, Natthawut Kertkeidkachorn (JAIST)

問い

同一の知識を表す複数の言語表現に対し
一貫性のない回答を返す問題にどう対処？

仮説

単語生成確率により回答一貫性を推定
単純な言語表現に分解して Retry or Abstain

結果

回答率と精度の両立に成功
GPT-3.5/4, StrategyQA にて実証

1. 目的: 信頼できる LM as KB

❖ LM as KB: 知識ベースとしての LM (Petroni+2019)

- 知識の保存: パラメタの学習 / 問い合わせ: トークン予測
- 利点: 柔軟に問い合わせ可能

❖ LM as KB の問題点

- 訓練事例にない回答を作話 (Ji+2022)
- LLM の回答は言語表現に鋭敏 (Hagström+2023)

❖ 理想の LM as KB

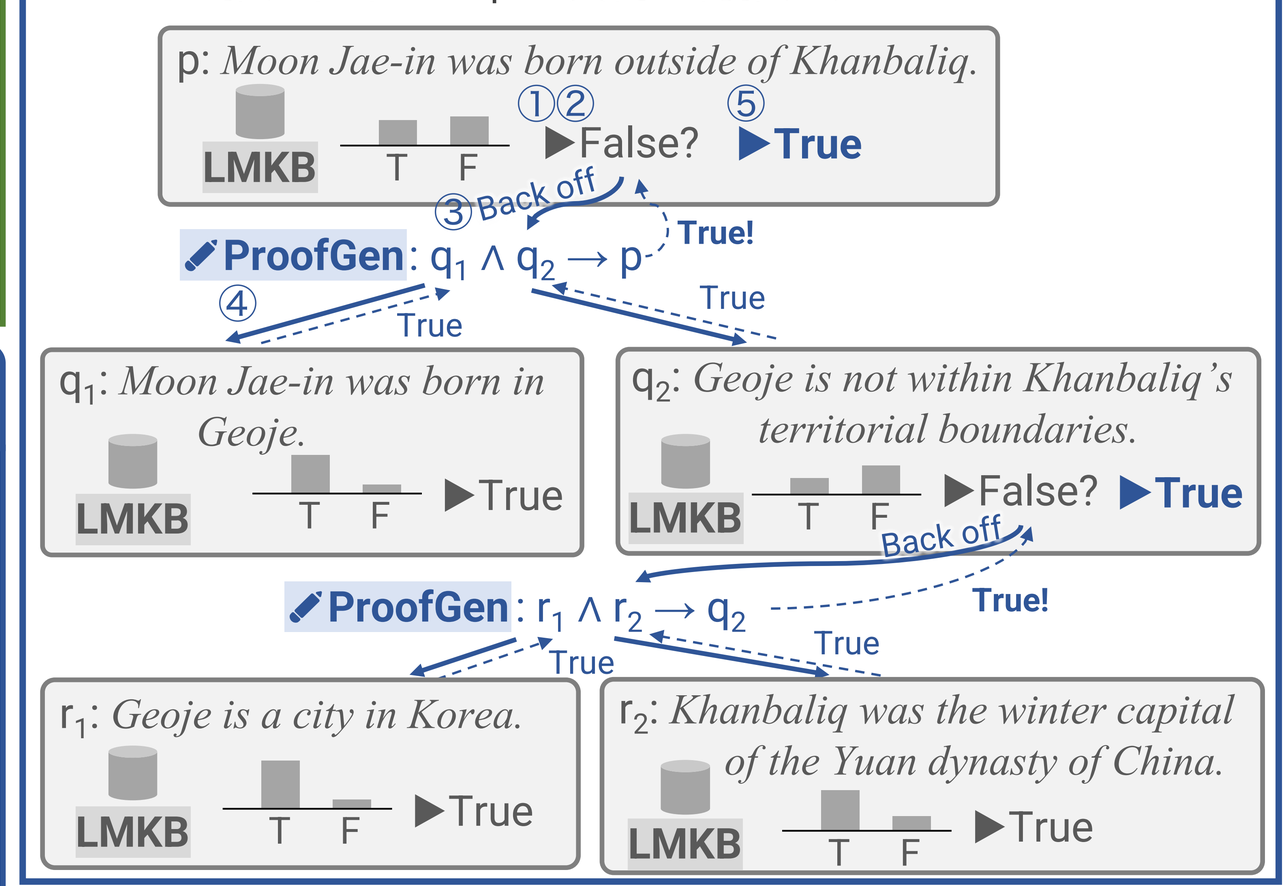
- “知識の知識”: 無知, 情報源, 自己矛盾, 意見多様性, ...
- 言語表現に依存しない形で知識の問い合わせができる
- 推論の過程を追跡できる

2. 提案法:定理証明器 + LM as KB

❖ 全体の流れ

① 命題 p の真偽を問い合わせ
② 結果に自信あり → そのまま解答
③ 結果に自信なし → 間接証明に Back-off: $q_1, q_2, ..., q_n$

- $q_1, q_2, ..., q_n$ の真偽を推定 (→ Step 1; 再帰呼出)
- 推定結果から p の真偽を再推定、上書き



3. 実装: LLM + Zero-shot Prompting

❖ LMKB (知識ベース)

- LLM + Zero-shot Prompting

> For the following question, provide your best guess.
> Give ONLY the guess. No other words or explanation.
> { p } . True or False?

- トークン “True”, “False” の生成確率分布 → $\pi(Y)$
- 自己認知機構: 正規化エントロピーに基づく確信度

$$\mathrm{LMKB}(p) = \begin{cases} \underset{y \in \{\mathrm{True},\mathrm{False}\}}{\mathrm{argmax}} \ \pi(Y = y) & \text{if } 1 - \mathrm{NE}(Y) \geq \tau \\ \mathrm{Unknown} & \text{otherwise.} \end{cases}$$

❖ ProofGen (間接証明器)

- LLM + Zero-shot Prompting

> To conclude a statement X, what premises are needed?
> Write two atomic premises P and Q.
> P and Q can be false facts. P and Q should contain
> all the world knowledge required to prove X. ...(省略)
> X: { p }

- 様々な分解方法、真偽推定方法がありうる
  - 十分条件: $\{q_1 \wedge q_2 \rightarrow p, q_1, q_2\} \vdash p$ (本研究)
  - 同値言い換え: $\{q \leftrightarrow p, q\} \vdash p$ (例: 誕生 ↔ 生まれる)
  - Modus Tollens: $\{p \rightarrow q, \neg q\} \vdash \neg p$ (例: 犬 → 4つ足)
  - 選言三段論法: $\{q \vee p, \neg q\} \vdash p$ (例: パン ∨ ごはん)

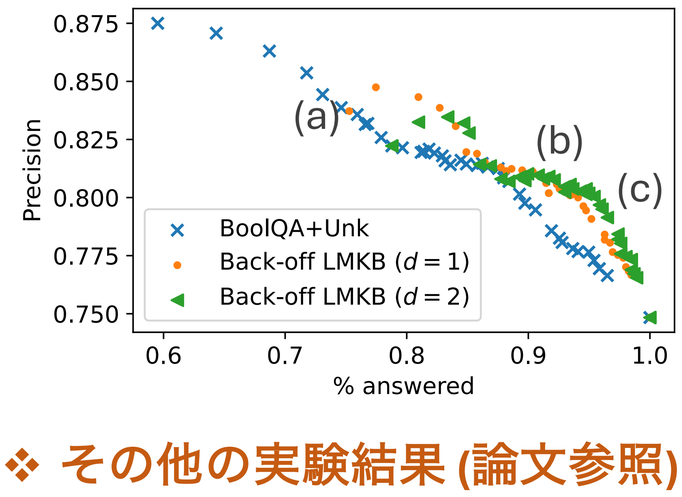
謝辞 本研究は JSPS 科研費 19K20332 の助成を受けたものです。

4. 実験結果: 回答率と精度を両立

❖ 実験設定

- 言語モデル: GPT-3.5, GPT-4 / データ: StrategyQA(Geva+2021)
- 評価指標: 回答率-精度曲線
- ベースライン: BoolQA+Unk (= LMKB)

❖ 主な実験結果



a) エントロピーが確信度として機能
b) 間接証明による正確な知識問い合わせ
c) 2次のバックオフも有効

❖ その他の実験結果 (論文参照)

- GPT-3.5 による実験, 間接証明の精度, エラー分析

5. 議論/今後: 自己認知とは...

- 確信度 == 無知?「分からない」とは？
  - トークン “True” “False” の生成確率は無知を表現？
- 今後の課題: 多様な証明戦略の採用, 証明の展開方策の工夫