20240725Meeting

GENERATE RATHER THAN RETRIEVE: LARGE LANGUAGE MODELS ARE STRONG CONTEXT GENERATORS

Introduction

- Retrieve-then-read
  - 從大型資料庫中檢索相關文檔，然後閱讀這些文檔以回答問題
- Generate-then-read(GENREAD)
  - 利用大型語言模型（InstructGPT）來生成問題的上下文文檔

GENREAD

- two main advantages
  - generated contextual documents contain the correct answer more often than the top retrieved documents
  - outperforms directly generating answers from large language models
- recall performance does not scale as well with generated documents

  → clustering-based prompt method

GENREAD

1. 生成文檔（Generate）
   - 通過大型語言模型（如InstructGPT）生成基於給定問題的上下文文檔
2. 閱讀文檔（Read）
   - 使用生成的文檔和問題來預測最終答案
   - 零樣本設定（Zero-shot Setting）
     - 在沒有訓練數據的情況下，使用生成-閱讀管道可以提升大模型的回答精度
   - 監督設定（Supervised Setting）
     - 在有訓練數據的情況下，小型閱讀模型（如FiD）可以通過生成的文檔進行微調，以提高性能

Zero-shot setting

- 在沒有任何訓練數據的情況下，直接使用預訓練的大型語言模型生成回答問題所需的上下文文檔，再使用生成的文檔來推理答案
- Step1 : Generate
  - 使用大型語言模型（如InstructGPT）生成與問題相關的上下文文檔
- Step2 : Read
  - 使用生成的句子與輸入問題一起從語言模型中生成最終答案
  - Prompt : Refer to the passage below and answer the following question. Passage: {background placeholder} Question: {question placeholder}
  - 將提示文本輸入語言模型以生成答案

Supervised setting

- 利用具體任務的訓練數據來微調預訓練的大型語言模型使模型在生成上下文文檔和回答問題時更具針對性和準確性
- 由於直接在下游數據集上微調大型語言模型可能成本較高，所以利用小型閱讀模型（例如 FiD）在監督設置下來閱讀生成的文件
- 擴大檢索文件的數量可以帶來更好的性能，但要求大型語言模型生成多個高質量的上下文是一項困難的任務

  → 提出diverse human prompt和clustering-based prompt

Diverse human prompt

- 人工提供不同的prompt
- 缺點
  - 需人工撰寫不同的提示，在不同的知識密集型任務中不容易泛化
  - 不同的大型語言模型可能對不同的提示詞敏感

Clustering-based prompts

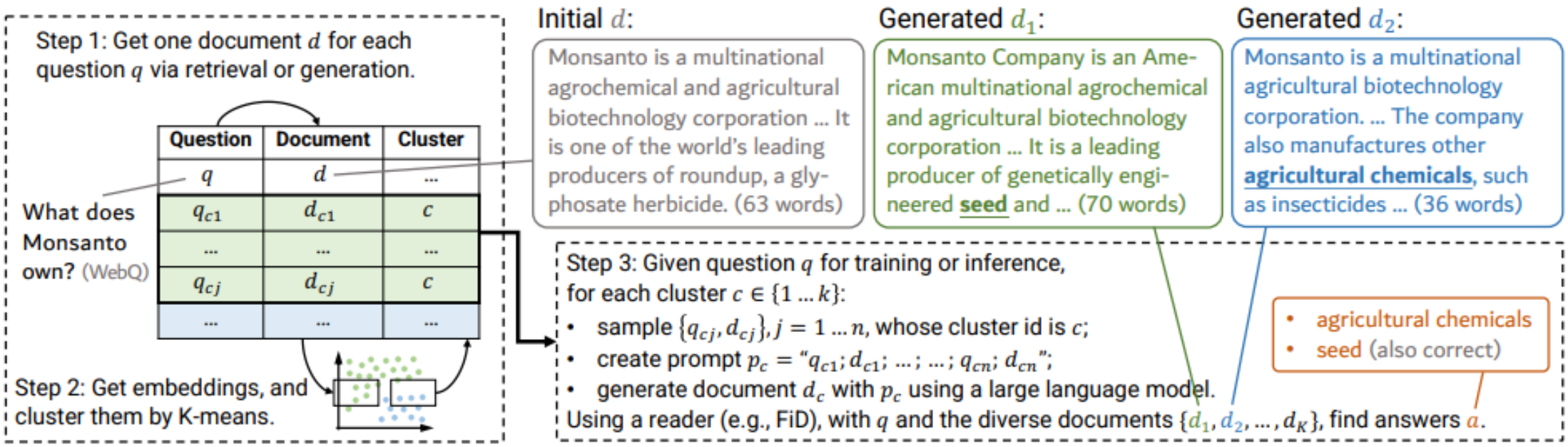

Figure 1: An overall framework of clustering-based prompting method. It leverages distinct question-document pairs sampled from each embedding cluster as in-context demonstrations to prompt a large language model to generate diverse documents, then read the documents to predict an answer.

Limitation

- 在新知識狀態和適應新領域的能力存在局限性
- 大型語言模型包含所有這些知識添加新知識可能需要經過再訓練
- 生成的文件可能會出現幻覺，導致錯誤的預測