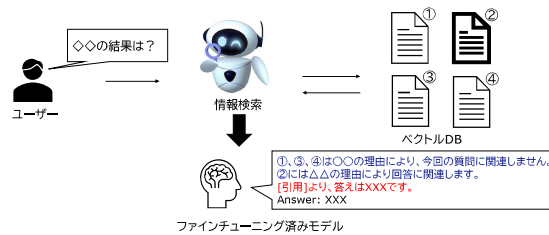


RAGの検索ノイズ抑制のためのファインチューニング

岡本怜奈, 大塚真子, 富田勇希, 濱松大介, 中本裕大 (SCSK株式会社)

1. 概要

- ◆ RAGの高精度化手法として、埋め込みモデルや検索アルゴリズムに依存しない、関連文書検索後におけるチャンクのフィルタリングに着目する
- ◆ 文書から回答に必要な情報を引用し、無関係な情報を無視する推論過程 (CoT) を学習させることで、ノイズを抑制し質問に正確に回答できるモデルの構築を目指す
- ◆ ファインチューニング (FT) の結果、未学習モデルと比較し、チャンクから回答に必要な情報を特定し、関連しないノイズ情報を回答に含めないようにすることにより、回答性能が向上した



2. 背景

- ◆ RAGの精度向上に向け、検索アルゴリズムの改良や検索後のノイズ除去など様々な工夫がされているが、質問に関連しないチャンクの参照による誤答やハルシネーションが依然として課題である
- ◆ 高度化手法のうち特定の埋め込みモデルや検索アルゴリズムに依存しない、汎用性が高いアプローチとして、検索後のチャンク選別に着目した
- ◆ 既存FT手法の限界
 - RAFT¹⁾: 回答に必要な関連情報のみを引用できるが、不要情報への明示的な除外理由の言及がない
 - CoN²⁾: 質問とチャンクの関連性判定はできるが、関連チャンクから回答に必要な箇所のみを絞り込みが限定的
- ◆ 今回試したアプローチ

RAFT ✖ CoN → 両者の弱点を補完した高精度な回答生成

3. 実験

3-1. 学習データ準備

wikipedia-utils^{※1}により日本語Wikipediaのテキストを取得・前処理を行ったコーパス「passages-c400-jawiki-20240401」^{※2}に対して下記の処理を適用

- ◆ 300文字以上のチャンクが6つ以上存在する記事のみを抽出 (300文字未満のチャンクを含む記事は除外)
- ◆ 1チャンクをランダムに選択し、チャンク記述内容を問う質問をLLMで作成
- ◆ 抽出後の409記事に対し、質問の回答生成に必要/不必要なチャンクを判断する思考過程を2パターン(RAFT, RAFT+CoN)用意
- ◆ 質問・思考過程・回答の作成には、Qwen-3-Next-80B-A3B-Instructを使用

<RAFTの設計方針>

- 質問回答に寄与する記述箇所を抽出すること
- 回答に必要なと判断した理由を言語化すること

<RAFT+CoNの設計方針>

- 左に加え、以下の指示を追加
- 不要なチャンクについても理由を言語化すること
 - 不要情報を排除し、回答に必要な部分のみ引用すること

RAFT+CoNの学習データ例

```

"question": "アーサー・ゴアはウィンブルドンで男子シングルスは何回勝ったか?",
"chunks": [
  "ゴアは1884年からウィンブルドン選手権に出場し始め...",
  "この大会では「チャレンジ・ラウンド」準決勝まで勝ち進み、...",
  "アーサー・ゴア(Arthur Gore, 1868年1月2日 - 1928年12月1日)は、...",
  "デビスカップの第1回大会は、..."
],
"model_response": "<think>
アーサー・ゴアがウィンブルドン選手権で男子シングルスは何回勝ったかを特定するために、chunkを段階的に分析します。まず、...
chunk3に、ゴアのウィンブルドン選手権での成績に関する明確な記述があります。
必要/不必要情報の判断理由
##begin_quote##アーサー・ゴア(Arthur Gore, 1868年1月2日 - 1928年12月1日)は、イングランド・リン
ドハースト出身の男子テニス選手。ウィンブルドン選手権で男子シングルス3勝、男子ダブルス1勝を
挙げ、同選手権の歴史を通じて「最長期間出場選手」の記録を持つ人である。##end_quote##
回答に必要な情報
この文は、ゴアがウィンブルドン選手権で男子シングルスで**3回**優勝したと明言しています。
他のchunk (chunk1、2、4) は、...直接関係しません。
したがって、chunk3の記述が質問に対する唯一の直接的な根拠となります。...</think>
思考過程の要約
<answer>3回</answer>"
  
```

※1: <https://github.com/singletongue/wikipedia-utils>

※2: <https://huggingface.co/datasets/singletongue/wikipedia-utils>

Wikipedia content is licensed under the CC-BY-SA 3.0 and GFDL licenses

3-2. 実験設定

- ◆ CoT学習データ880件、評価データ211件を使用
 - 正解チャンク+ノイズチャンク3件、ノイズチャンク4件を8:2の割合で含む
- ◆ ファインチューニングはLoRAで実施($lora_rank=16$, $lora_alpha=32$, $epoch=3$)
- ◆ V4-8b-decay2m-ipt_v3.1-instruct4に対し、質問とチャンクを入力として受け取った後の、思考過程+回答生成を2パターン学習(RAFT, RAFT+CoN)
- ◆ RagasのAnswer Accuracyによる定量評価を実施
 - 模範解答とモデル出力の最終回答の言及内容の一致度を比較

4. 結果・考察

4-1. 結果

自動評価のLLMにはQwen3.5-122B-A10Bを使用

モデル	Answer Accuracy		
	完全回答(件)	部分回答(件)	不正解(件)
未学習モデル	146	36	29
未学習モデル (RAFTプロンプト)	136 (-10)	43 (+7)	32 (+3)
未学習モデル (RAFT+CoNプロンプト)	137 (-9)	40 (+4)	34 (+5)
RAFT	148 (+2)	39 (+3)	24 (-5)
RAFT+CoN	154 (+8)	39 (+3)	18 (-11)

4-2. 考察

【Answer Accuracyの向上理由】

- ◆ 未学習モデルは回答に不要な情報を含んでいたが、ファインチューニングモデルは必要な情報のみで回答できるものが増えた
- ◆ 検索結果に回答に必要な情報が一切含まれない場合に、未学習モデルはハルシネーションで架空の答えを生成することが多かったが、ファインチューニングモデルは正しく「回答できない」と判断できるようになっていった

【RAFT+CoNの新規正解例】

質問: 誰が「にぎりっ屁」の元祖とされていますか?
回答: chunkには「にぎりっ屁」の元祖に関する情報が含まれていないため、答えを導けません。

RAFTの思考過程と回答

chunk1の解釈 (桜金造の「怪談の語り手」としての活動について述べられている)

chunk3~4の解釈 (桜金造の俳優業や私生活、病歴について述べている)

chunk2の解釈 (この文は、桜金造の芸名の由来について述べている)

よってchunk全体に情報はありません。

しかし、chunk全体を通じて、桜金造がコメディアンとして活動し、俳優としても活動し、怪談の語り手としても活動しているという情報があります。¥n¥n「にぎりっ屁」は、彼の代表的な芸の一つであり、chunkに明示されていないとしても、chunk2の芸名の由来と芸歴の記述から、彼がその芸の元祖であると推測するのが自然です。...

<answer>桜金造</answer>

モデルが関係ない文脈から勝手に推測をしてしまっている

RAFT+CoNの思考過程と回答

chunk1の解釈 (桜金造の芸名の由来について→関係なし)

chunk2の解釈 (桜金造のプロフィールと芸歴→関係なし)

chunk3の解釈 (2013年の病状と復帰について→関係なし)

chunk4の解釈 (桜金造の「怪談の語り手」としての活動について→関係なし)

桜金造はコメディアンとして知られ、特に「アゴ&キンゾー」時代の芸風は荒唐無稽なネタで有名でしたが、その芸風の中に「にぎりっ屁」という特定の芸が含まれているという記述は文脈にありません。また、他の芸人や人物との関連性も示されていません。¥n¥nしたがって、提供されたchunkからは「にぎりっ屁」の元祖が誰かを特定する根拠が一切得られません。

<answer>unknown</answer>

結果

RAFTは関係ない文脈の参照により誤った思考を行ってしまっている

RAFT+CoNは各チャンクで述べている内容と質問との関連性を明示的に評価することで、思考過程の要約を生成する際に、 unnecessary 情報の参照を行わなくなった

参考文献

1. Zhang et al., "RAFT: Adapting Language Model to Domain Specific RAG", arXiv preprint, 2024.
2. Yu et al., "Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models", EMNLP, 2024.