6R-09

RAG を用いた有用性の高い技術記事を推薦するモデルとデータセットの開発

竹内悠人[†1]　菅原朔[†2]
灘高等学校[†1]　国立情報学研究所[†2]

1. はじめに

Large Language Models（LLM）の台頭，特に GPT-4[1]のようなモデルがテキスト生成の分野で大きな進歩を遂げ，Retrieval-Augmented Generation （RAG）のように，外部の知識ベースから事実を検索し，LLM に回答を生成させる技術が発展している[2]．しかし,これらのモデルはドメイン特有のデータに対しては課題を抱えており,たびたび Hallucination と呼ばれる虚偽の内容や[3], 知識が古いこと[4], ドメイン固有の専門知識不足によって[5]，役に立たない内容を生成する傾向がある．特に正確かつ最新の情報が常に求められる技術分野において，これは重要な課題であり，適切な文書選定の必要性を意味している．

一方で，RAG を医療や政治などの特定ドメインに適用する研究は多数存在するが[6]，プログラミングや開発といった技術分野に特化した研究は未だ見られない. 加えて, RAG で望ましい情報の候補を取得してきたとき,その評価においてはユーザーが有用と思うかが重要になる。しかし有用性を評価することは本質的に主観的なタスクであり，モデルにとって困難なタスクである．また，ウェブコンテンツ自体の有用性を評価する研究は多数存在するが，各々が様々な複雑な指標を有用性の評価に用いるため，評価プロセスは煩雑になっているのが現状である [11][11].

本研究ではまず，技術記事の有用性を定義したデータセットはこれまで存在しなかったため，技術記事の有用性を特徴づける評価基準を選定し，人手によるラベリングを行って，日本語技術記事データセットを作成した．そしてこのデータセット中の，最終的な記事との質に強い相関のある評価基準を利用して, GPT-3.5 の fine-tuning [7]を行うことで，有用な技術記事の推薦を可能とし，技術分野に関する高品質な日本語の RAG モデル作成へのベースラインを提供する．

2. 関連研究

2.1 Retrieval-Augmented Generation

LLM に含まれた知識はいずれ古くなり，Hallucination を引き起こす可能性がある[3][4]．また RAG は外部の知識を参考にすることで，より正確な応答を生成することができる．さらに、LLM は回答を作成するときに提供された検索結果に依存する傾向があり，これらの結果の質が信頼性に大きな影響を与えることがわかっている[8].一方で RAG そのものの性能を評価する取り組みも行われており，RAG を LLM に適用することの性能の限界も見え始めている[9]．また医療や政治といった特定ドメインに RAG をなどの適用する研究は存在するものの[6]，技術分野に特化する研究はまだなされていない．

2.2 ウェブサイトの品質

過去 30 年で，ウェブサイトは社会に情報を広め，サービスを提供するインターネット上で最大のプラットフォームになった．それにもかかわらず，ウェブサイトの品質に関する基準はまだ完全に統合されていない[10][11]．またこれらの評価に関する研究は，コンテンツの品質、サービスの品質、技術的な品質などの多面的な要素を組み合わせて評価に用いることができる[12]．こうした中で Morales-Vargas らは,ウェブサイトの品質に関する 305 の出版物の文献レビューを行い, 120 個以上の品質を特徴づける評価基準及び包括的な評価フレームワークを定義し，この分野における品質評価の統合を推進した. [13].そこで我々はこのフレームワークに基づいて技術記事の有用性評価に用いることができる評価基準を選定し，新たな指標を定義して追加した．また今回は「人が記事を有用だと判断する」という前提のもとで話を進めることとする．

3. 提案手法

3.1 データセットの作成

以下の手順でデータセットを作成した．

（1）日本語 stack overflow[14]の人気タグ上位 40 件を取得し，それぞれのタグ上位 5 件の質問に対して，質問のタイトルを検索クエリとして適切に変形した．変換した検索クエリを Google 検索にかけた後，検索結果の上位 3 件及び 18,19,20 番目の 6 つの記事を取得した. これは記事の質が高いと予測されるものと，質が低いと予測されるものの両方を取得するためである．

（2）Morales-Vargas らが提案するフレームワーク[13]に基づき, 120 個以上のウェブサイトの品質を特徴づける評価基準の中から，技術記事の有用性評価に用いることができる要素を選定した．加えてここに，技術記事に特化させる目的でいくつかの評価基準を追加した．こうして得られた評価基準を図 1 に示す．

How to Prepare Your National Convention of IPSJ Reports in MS-Word
†1 YUTO TAKEUCHI, Nada High School.
†2 SAKU SUGAWARA, National Institute of Informatics

選定したパラメーター	追加したパラメーター
参考・引用文献が書かれているか？	検索キーワードに対して最新の情報が反映されているか？
図表が用いられているか？	コードスニペットが書かれているか？
明確に記事の内容を理解できたか？	より深い学びに繋がるか?
記事の中身は網羅性があったか？	検索キーワードに対する関連性はあったか？
記事は簡潔であったか？	

図 1 評価基準一覧

（3）こうして得られた 1,200 記事を，図 1 に示した 9 つの評価基準に加え，「最終的にその記事を読む価値があったか」という設問を用意し，ソフトウェアエンジニア 21 人の実験参加者によって 5 段階での評価を行った．追加した設問に対する答えが得られなかったものに関しては，既存の 9 つの評価基準への回答を元に gpt-3.5-turbo-1106 の fine-tuning を行って補完した．補完したものについて，人手のラベルで品質をチェックし，問題ないものがほとんどであることを確認した．

3.2 データセットの評価

以下の手順でデータセットを評価した．

（1）図 1 に示した 9 つの評価基準のうち，どれが最終的な記事の質と強い相関関係があるか，それぞれの相関関係を計算して検証した．

（2）（1）で得られた，最終的な記事の品質と強い相関関係を持つ評価基準を対象に，記事の本文から各基準での評価の生成を，個別で gpt-3.5-turbo-1106 の fine-tuning を行って学習した．これに加え，各基準のスコアのみから最終的な記事の品質を判断する fine-tuning を行った．これらを併用することで，記事の本文から記事の品質を予測することが可能になる．

4. 結果

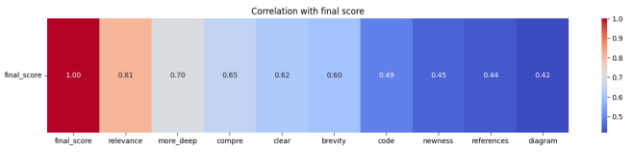

図 2 各評価基準の，最終的な記事の質との相関

技術記事の有用性を特徴づける各評価基準の，最終的な記事の品質との相関は図 2 のようになった．
また fine-tuning での予測結果を，5 段階から 3 段階の評価に変形したものの F1 score は 0.44 となった

5. 考察

今回 fine-tuning の F1score が伸び悩んだ原因は，図 1 の評価基準一覧の一部のみを用いたことが原因だと考えられる．また Google 検索での検索順位も評価基準として用いることで，性能の向上が見込まれる．

6. 結論

本研究では日本語技術記事の有用性を特徴づけるデータセットの作成とモデルの提案を行った．今後は図 1 の評価基準の全て及び Google 検索での検索順位も評価基準として用い，性能向上を目指していきたい．

謝辞

本研究は，国立研究開発法人科学技術振興機構グローバルサイエンスキャンパス（GSC）「情報科学の達人」育成官民協働プログラム（国立情報学研究所，情報処理学会，情報オリンピック日本委員会）及び株式会社ラック　サイバー・グリッド・ジャパン　IT スーパーエンジニア・サポートプログラム「すごうで」並びに国立研究開発法人情報通信研究機構（NICT）「SecHack365」の支援を受けた．

参考文献

[1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.
[2] OpenAI. 2023. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.
[3] Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. ACM Comput. Surv., 55(12).
[4] He, H.; Zhang, H.; and Roth, D. 2022. Rethinking with Retrieval: Faithful Large Language Model Inference. arXiv:2301.00303.
[5] Shen, X.; Chen, Z.; Backes, M.; and Zhang, Y. 2023. In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT. arXiv:2304.08979.
[6] Siriwardhana, S.; Weerasekera, R.; Wen, E.; Kaluarachchi, T.; Rana, R.; Nanayakkara, S. 2022. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. arXiv:2210.02627
[7] OpenAI Models〈https://platform.openai.com/docs/models〉（参照 2024-01-08）
[8] Ren, R.; Wang, Y.; Qu, Y.; Zhao, W. X.; Liu, J.; Tian, H.; Wu, H.; Wen, J.-R.; and Wang, H. 2023. Investigating the Factual Knowledge Boundary of Large Language Models with Retrieval Augmentation. arXiv:2307.11019.
[9] Jiawei Chen, Hongyu Lin, Xianpei Han, Le Sun.2023. Benchmarking Large Language Models in Retrieval-Augmented Generation. arXiv:2309.01431
[10] Law, R., Qi, S. and Buhalis, D. (2010), “Progress in tourism management: a review of website evaluation in tourism research”, Tourism Management, Vol. 31 No. 3, pp. 297-313.
[11] Semerádová, T. and Weinlich, P. (2020), “Looking for the definition of website quality”, in Semerádová, T. and Weinlich, P. (Eds), Website Quality and Shopping Behavior: Quantitative and Qualitative Evidence, SpringerBriefs in Business, Cham, pp. 5-27.
[12] Rocha, Á. (2012), “Framework for a global quality evaluation of a website”, Online Information Review, Vol. 36 No. 3, pp. 374-382.
[13]Morales-Vargas, Pedraza-Jimenez, and Codina. (2023) “Website quality evaluation: a model for developing comprehensive assessment instruments based on key quality factors”.
[14]stack overflow〈https://ja.stackoverflow.com/〉（参照 2024-01-08）.