E コマースの自然文検索における LLM 生成クエリを用いたファインチューニングの有効性検証

福光嘉伸　浅野孝平　作本猛　稲田和明
株式会社 MonotaRO
{yoshinobu.fukumitsu, kohei.asano, takeshi.sakumoto, kazuaki.inada}@monotaro.com

概要

近年のセマンティック検索の普及により，商品の用途を含む自然文のクエリが増加するなど，E コマースにおけるユーザの検索行動に変化が生じている．また，セマンティック検索の性能向上を目的として，利用目的に応じたデータを用いて汎用的なテキスト埋め込みモデルに対するファインチューニングが実施される．しかし，自然文のクエリは検索ログ全体に占める割合が少ないため，ファインチューニングに十分な規模の学習データを得られない．そこで本研究では，検索クエリと商品のクリックログから，大規模言語モデル（LLM）を用いて自然文のクエリを生成することで学習データを拡張する．評価実験では，実際の E コマースのログを用いた生成クエリによるファインチューニングを行い，自然文のクエリに対する性能向上を確認した．

1　はじめに

近年の E コマースでは，検索クエリの意味や意図を理解するセマンティック検索が重要である [1, 2]．密ベクトル表現モデルによるセマンティック検索は，テキストを高次元のベクトル空間に埋め込むことで意味的な類似度計算を可能にし [3]，テキストマッチング [4] では捉えられない意味的関連性を考慮した商品検索を実現できる [5]．しかし，多くの事前学習済み埋め込みモデルは一般的なテキストで学習されているため，専門性の高い商品を扱う E コマースでは，専門用語や業界特有の表現に対応させるために，ファインチューニングなどの性能向上の対応が必要となる [6]．

近年のセマンティック検索の普及により，ユーザの検索クエリも多様化している．本研究では検索クエリを 2 つの形式に分類し取り扱う．「昙間テープ シャッター」「水栓 レンチ」 のようにスペース区切りの単語からなる検索クエリを**単語クエリ**，「シャッター下の昙間をふさぐ」「水栓を開ける道具」といった用途や目的を含むフレーズからなる検索クエリを**自然文クエリ** と定義する．モノタロウ[1)] の検索ログを分析した結果，2023 年から継続的に自然文クエリが増加していることが確認された．自然文クエリでの検索が増加するなかで，より良い検索体験をユーザに提供するためには自然文クエリに対する検索性能の向上が重要になる．しかし，自然文クエリが検索ログ全体に占める割合は現状では少なく，ファインチューニングに必要な量の学習データの確保が課題となる．

本研究では，自然文クエリの学習データ不足の課題に対し，大規模言語モデル（LLM）を用いたクエリ生成の有効性を検証する．具体的には，検索ログから抽出した検索クエリとクリックされた商品のペアを入力として LLM で自然文クエリを生成し [7]，テキスト埋め込みモデルのファインチューニングに活用する．評価実験では，生成した自然文クエリが検索性能に与える影響と，学習データにおける単語クエリと自然文クエリの構成比率が検索性能に与える影響を分析する．

2　関連研究

密ベクトル検索において，Reimers と Gurevych[3] の Sentence-BERT は Siamese network を用いた効率的なテキスト埋め込み手法を提案した．日本語環境では，塚越と笹野 [8] による Ruri が大規模な日本語コーパスで学習された汎用的なテキスト埋め込みモデルとして注目されている．Ruri は一般的なウェブ文書や Wikipedia で事前学習されており，汎用テキストに対して高い性能を示す．しかし，E コマース商品検索において汎用モデルは専門用語の理解不足により検索性能が低下するため，ドメイン適応が必

1） https://www.monotaro.com/

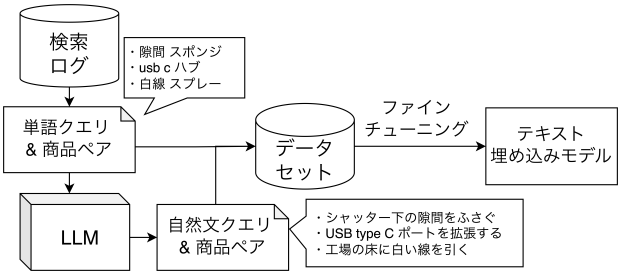

図 1: 手法の概要

要であることが示されている [6]．本研究では，埋め込みモデルのファインチューニング有無や用いるデータの形式による検索性能を比較し，専門性の高い E コマースにおけるドメイン適応の必要性を検証する．

西川ら [9] はユーザ行動ログから検索意図が整合するクエリペアを正例とした対照学習によるファインチューニングを提案した．また，Jin ら [10] の LADER は，検索ログから類似した過去のクエリを特定し，クリックされた文書情報を集約して学習データを構築する．これらの手法は大量の検索ログを前提としているため，特定のクエリ形式のログが少ない環境では適用が困難である．本研究では，このようなデータ不足の課題に対し，LLM によるデータ拡張で対応する．

LLM による学習データ拡張として，Bonifacio ら [11] は GPT の few-shot learning によって既存文書から合成クエリを生成する手法を提案した．また，検索ログに依存せずテキスト生成モデルで質問をマイニングする手法も提案されている [12, 13]．E コマース分野では，Jagatap ら [14] が商品・クエリインタラクションに基づいて LLM をファインチューニングし，ドメイン特化の合成クエリを生成している．しかし，これらの研究では日本語の専門性の高い E コマースの用語への対応は検証されていない．本研究では，当該ドメインにおいて LLM を用いたデータ拡張の有効性を検証する．

3 クエリ生成によるデータセット構築とファインチューニング

本研究で用いる手法の概要を図 1 に示す．はじめに，検索ログから単語クエリを抽出し，単語クエリとクリックされた商品のペアから LLM を用いて自然文クエリを生成する．次に，生成した自然文クエリと元の単語クエリを組み合わせてテキスト埋め込みモデルをファインチューニングする．

検索ログには，ユーザが入力したクエリとクリックした商品のペアが大量に蓄積されている．これらのペアはユーザの検索意図と商品の関連性を暗黙的に示している [15, 16]．本手法では，ログから抽出した単語クエリと商品名のペアを入力として，LLM で自然文クエリを生成する．表 1 に自然文クエリの生成例を示す．生成された自然文クエリは，元の単語クエリの検索意図を保持しつつ，「〜に使う」「〜する」といった用途を含む表現に変換される．単語クエリと生成した自然文クエリからデータセットを構成することで，同じ商品に対して多様な検索クエリを学習データとして取得できる．

テキスト埋め込みモデルのベースモデルには，日本語に特化した事前学習済みモデルを使用する．ファインチューニングでは，データセット内のクエリと商品のペアを正例，同一バッチ内の他の商品を負例として対照学習を行う [17]．

4 評価実験

テキスト埋め込みモデルのファインチューニングを通じて，ドメイン適応の効果や生成した自然文クエリの検索性能への影響を検証する．4.1 節では単語クエリと自然文クエリで学習したモデルの性能を比較し，4.2 節では学習データセットにおける単語クエリと自然文クエリの構成比率が検索性能に与える影響を調査する．

データセットを構成する単語クエリとクリックされた商品のペアは，モノタロウの検索ログから抽出する．自然文クエリの生成には，Gemini 2.5 Flash-Lite[2] を使用し，温度パラメータは 0.1 とする．

ベースとなるテキスト埋め込みモデルは Ruri-large[3] を用いる．商品検索対象として，モノタロウの約 60,000 件の商品を使用する．テストデータは，単語クエリ 10,000 件と自然文クエリ 200 件を検索ログから学習データと重複しないように抽出する．単語クエリは，スペース区切りの単語数が 1〜2 個（スペース 0〜1 個）で各単語の文字数が 6 文字以下であるクエリを抽出した．自然文クエリは，スペースを含まず文字数が 10 文字以上であるクエリを抽出し，助詞や動詞を含む自然文表現であることを目視で確認して採用した．評価クエリに対する正例は，それぞれのクエリの検索でクリック実績のある商品とした．評価指標として **Precision@k** と **Recall@k** を用いる．各指標はそれぞれ検索結果の上位 k 件の

2） https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash-lite
3） https://huggingface.co/cl-nagoya/ruri-large

表 1: LLM を用いた自然文クエリ生成の例

単語クエリ	商品名	生成された自然文クエリ
クリップ リムーバ	メガドラ クリップリムーバー	クリップを外す
l 型 ホース	L 型ホース継手	ホースの接続に使う
usb c ハブ	USB-C PD イーサネット ハブ	USB-C ポートを拡張する

適合率と再現率である．

4.1 クエリ形式毎の性能検証

クエリ形式（単語クエリ・自然文クエリ）の違いがテキスト埋め込みモデルのファインチューニングの性能に与える影響を評価する．実験では以下の 3 つの設定を比較する．

- **Base**：ファインチューニングなし（Ruri-large をそのまま使用）
- **Term**：検索ログから収集した単語クエリ・商品ペア 50,000 件でファインチューニング
- **Sentence**：LLM で生成した自然文クエリ・商品ペア 50,000 件でファインチューニング

各テキスト埋め込みモデルの学習データは同一の商品セットを用い，クエリ形式のみを変更する．Precision@k と Recall@k を $k = 1$ から 30 まで評価し，各モデルを比較する．また，実際のクエリに対する検索結果を各モデルで比較し，定性的に評価する．

図 2 に示す単語クエリに対する評価結果では，Term が最も高い検索性能を示し，Sentence はそれに次ぐ性能を示した．一方，Base はファインチューニングしたモデルに比べ大幅に低い性能に留まる結果となった．図 3 に示す自然文クエリに対する評価結果では，Sentence が最も高い検索性能を示し，Term はそれに次いだ．Base は単語クエリに対する評価と同様に大幅に劣る性能となった．

ベースラインモデルがファインチューニングをしたモデルに対して大幅に劣ったことは，ドメイン適応の重要性を示している．学習クエリと評価クエリの形式が一致する場合に性能が最大化される理由として，モデルがそのクエリ形式に適した表現を学習できることが考えられる．また，異なるクエリ形式で学習したモデルも比較的高い性能を示した．これは，単語と自然文という異なるクエリ形式でも意味的な共通性があり，モデルがその共通性をある程度捉えられることを示唆する．

表 2 に，「シャッター下の昙間をふさぐ」という自然文クエリに対する各モデルの検索結果上位 3 件

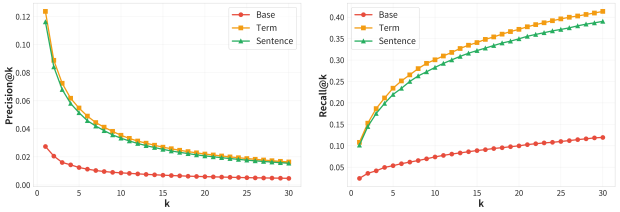

図 2: 単語クエリに対するモデル比較（左: Precision@k，右: Recall@k ）

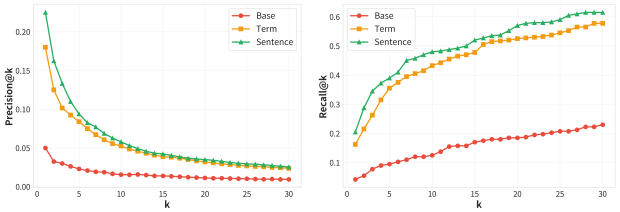

図 3: 自然文クエリに対するモデル比較（左: Precision@k，右: Recall@k ）

を示す．表中の記号は，クエリと商品との関連度を示しており，〇は関連性が高い，△は関連性がある，×は関連性が低いことを表す．Base では，3 商品すべてがクエリとの関連性が低いことが分かる．一方，Term と Sentence はいずれも関連性の高い商品を上位に検索できているが，Sentence の方が「昙間をふさぐ」という用途により適した商品を検索できている．

以上より，LLM で生成したクエリによるファインチューニングが実際の自然文クエリに対する検索性能の向上に有効であることと，各クエリ形式に対しては同じ形式での学習が効果的であることが示された．

4.2 データ構成比率の影響

学習データにおける単語クエリと自然文クエリの構成比率が検索性能に与える影響を調査する．実際には 2 つの形式が混在するため，両形式に対してバランスの取れた性能が求められる．

実験では，単語クエリと自然文クエリの両データを組み合わせた学習データを用いてファインチューニングを行い，検索性能を評価する．単語クエリの割合を 0% から 100% まで 10% 刻みで変化させ，それぞれの割合に応じて両クエリ形式を合計 50,000

表 2: 各モデルによる「シャッター下の昙間をふさぐ」の検索結果と関連度

順位	Base	Term	Sentence
1	×; A1 判 ポスターフレーム	△; ダクト挿入型逆風防止シャッター	○; シャッタースポンジ
2	×; 飛散防止フィルム	○; シャッタースポンジ	○; ワイドすき間モヘアシール
3	×; 拡散板	△; シャッターガード	○; 高機能スキマテープ (シャッター用)

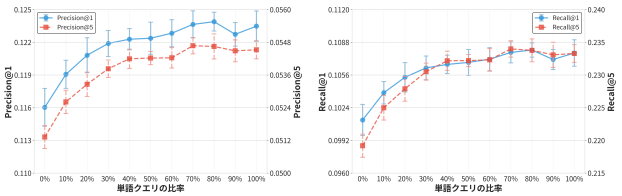

図 4: 単語クエリに対する学習データセットの構成比率毎の性能（左: Precision@k，右: Recall@k ）

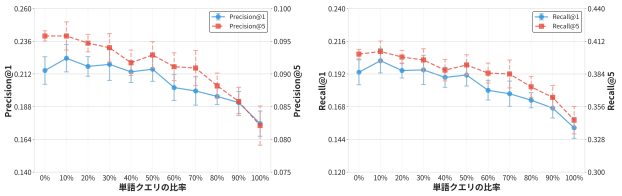

図 5: 自然文クエリに対する学習データセットの構成比率毎の性能（左: Precision@k，右: Recall@k ）

件となるように組み合わせて学習データとする．各構成比率において，データをランダムにサンプリングして学習する試行を 10 回繰り返し，Precision@k と Recall@k（$k = 1, 5$）の平均値と標準偏差を算出する．

図 4 に単語クエリに対する 10 回の試行における評価値を示す．図 4 より，単語クエリの学習データ比率が高いほど検索性能が向上する傾向が確認された．単語クエリの比率が 100% 付近で最も高い性能が得られ，自然文クエリの比率が増加するにつれて性能は低下した．

一方，図 5 の自然文クエリに対する評価結果では，自然文クエリの学習データ比率が高いほど性能が向上し，単語クエリの比率が 0% から 10% の範囲で最も高い性能を示した．単語クエリの比率が増加すると性能は低下し，特に 90% 以降では性能低下の度合いが大きくなった．

図 4 と図 5 で共通する点として，性能が大幅に変化する比率が存在する．単語クエリの評価では単語クエリの比率が 20% 付近，自然文クエリの評価では 90% 付近を境に性能低下の度合いが大きくなる．これは，各クエリ形式が性能を維持するために一定量以上の学習データが必要であることを示す．また，学習データにおいて多数派のクエリ形式は性能が向上する一方で，少数派のクエリ形式では性能が低下することを示している．本実験により，想定されるクエリ形式の出現頻度に応じて比率設定を決定することでバランスの取れた性能が期待できる．

5 おわりに

本研究では，検索ログに不足している自然文クエリを LLM で生成し，テキスト埋め込みモデルのファインチューニングに活用する手法の有効性を検証した．専門性の高い E コマースを対象に，LLM で生成した自然文クエリによる検索性能向上と，学習データにおける単語クエリと自然文クエリの構成比率が検索性能に与える影響を分析した．

検索ログを用いた評価実験により，LLM で生成した自然文クエリを学習データとして使用することで，実際の自然文クエリに対する検索性能を向上できることを示した．また，専門性の高い E コマースの商品検索においても，ドメイン適応の重要性が示唆された．さらに，学習データにおける単語クエリと自然文クエリの構成比率が，各クエリ形式に対する検索性能に大きく影響することを明らかにした．特に，性能を維持するために一定量以上の学習データが必要であることから，想定されるクエリ形式の出現頻度に基づいた学習データ設計の重要性が示された．これらの結果は，ファインチューニングに十分な規模の自然文クエリが検索ログに不足している環境においても，LLM を活用することで効果的な意味的検索を実現できることを示唆している．

今後の課題として，商品画像などのマルチモーダル情報を活用し，色や形などの商品画像からのみ得られる情報を含んだクエリを生成することでさらなる検索性能の向上が期待できる．また，単語クエリか自然文クエリかをリアルタイムに識別し，専用のモデルを用いることで性能向上が期待できる．

謝辞

本研究は，愛媛大学大学院 眞鍋光汰氏とのインターンシップを起点として開始されたものであり，同氏には深く感謝申し上げる．

参考文献

[1] Priyanka Nigam, Yiwei Song, Vijai Mohan, Vihan Lakshman, Weitian (Allen) Ding, Ankit Shingavi, Choon Hui Teo, Hao Gu, and Bing Yin. Semantic product search. In **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, KDD '19, pp. 2876–2885, New York, NY, USA, 2019. Association for Computing Machinery.

[2] Han Zhang, Songlin Wang, Kang Zhang, Zhiling Tang, Yunjiang Jiang, Yun Xiao, Weipeng Yan, and Wen-Yun Yang. Towards personalized and semantic retrieval: An end-to-end solution for e-commerce search via embedding learning. In **Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '20, pp. 2407–2416, New York, NY, USA, 2020. Association for Computing Machinery.

[3] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[4] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. **Found. Trends Inf. Retr.**, Vol. 3, No. 4, p. 333–389, April 2009.

[5] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics.

[6] Thanh Nguyen, Nikhil Rao, and Karthik Subbian. Learning robust models for e-commerce product search. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6861–6869, Online, July 2020. Association for Computational Linguistics.

[7] Huanhuan Cao, Derek Hao Hu, Dou Shen, Daxin Jiang, Jian-Tao Sun, Enhong Chen, and Qiang Yang. Context-aware query classification. In **Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '09, pp. 3–10, New York, NY, USA, 2009. Association for Computing Machinery.

[8] 塚越駿, 笹野遼平. Ruri: 日本語に特化した汎用テキスト埋め込みモデル. 言語処理学会 第 31 回年次大会 発表論文集, pp. 1622–1627, 2025.

[9] 西川荘介, 平子潤, 鍛治伸裕, 渡邉幸暉, 浅野広樹, 山城颯太, 佐野峻平. ユーザ行動ログに基づくクエリ理解のための検索クエリ埋め込み. 言語処理学会 第 31 回年次大会 発表論文集, pp. 3293–3298, 2025.

[10] Qiao Jin, Ashley Shin, and Zhiyong Lu. LADER: Log-augmented DEnse retrieval for biomedical literature search. In **Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '23, pp. 2092–2097, New York, NY, USA, 2023. Association for Computing Machinery.

[11] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. InPars: Data augmentation for information retrieval using large language models, 2022.

[12] 大村和正, 石原祥太郎. 検索クエリログを用いない自然な質問のマイニングの検討. 言語処理学会 第 31 回年次大会 発表論文集, pp. 2433–2438, 2025.

[13] Said Faraby, Kang Adiwijaya, and Ade Romadhony. Review on neural question generation for education purposes. **International Journal of Artificial Intelligence in Education**, Vol. 34, , 2023.

[14] Akshay Jagatap, Srujana Merugu, and Prakash Mandayam Comar. Improving search for new product categories via synthetic query generation strategies. In **Companion Proceedings of the ACM Web Conference 2024**, WWW '24, pp. 29–37, New York, NY, USA, 2024. Association for Computing Machinery.

[15] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. Context-aware query suggestion by mining click-through and session data. In **Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining**, pp. 875–883, 2008.

[16] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In **Proceedings of the 2008 Eighth IEEE International Conference on Data Mining**, p. 263–272, USA, 2008. IEEE Computer Society.

[17] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In **International Conference on Learning Representations**, 2019.