過去クエリを介した関連文書検索システム

須賀幹太 [1,2] 宮本琳太郎 [1] 桂尚輝 [1] 梅澤慶介 [1]
[1] 株式会社メルカリ [2] 早稲田大学
thickstem@asagi.waseda.jp , {r-miyamoto,n-a-katsura,k-umezawa}@mercari.com

概要

多くのサービスではユーザーが困った際に質問フォームから問い合わせをすることができる。問い合わせに対して関連するヘルプガイドを自動で検索・推薦できれば、問題解決までの時間削減に寄与できる。また、ChatGPT を始めとする LLM を用いた文章生成において、関連ガイドをコンテキストとして渡すことで正確な返答文章の生成が可能になる。そこで本研究ではユーザーからの新規クエリに対して、過去のクエリとその返信文を経由することで関連ガイドを高精度に検索する手法を提案する。提案手法と新規クエリから直接ガイド文章との類似度を計算する既存手法を比較した結果、提案手法は高精度に文章を検索できることが分かった。

1 はじめに

サービス系アプリケーションでは、ユーザに生じた問題点をフリーテキストで運営側に問い合わせることができることが多い。現在それらの問い合わせに対する返答は、必要性に応じてカスタマーサポート等の担当部署が一件ずつ作成している。しかし近年の急激な IT 化に伴って、ユーザ数と共に問い合わせの件数も増加しており、人力で返答文を作成していると返信までに時間を要するという問題がある。ユーザの問い合わせに関連するヘルプガイドを自動的に検索・推薦することができれば、ユーザは問題が即座に解決を解決することができ、カスタマーサポート側にとっても人力で対応するべき問い合わせ数の減少効果が期待できる。また、近年台頭してきた ChatGPT を始めとする LLM では自然な文章を高速に生成することができるが、その生成には事前学習時に含まれている情報しか使用できないため、各サービス固有の問題に対する QA ボットとして直接使うことは難しい。その対応策として外部知識をコンテキストとして LLM に与えることで回答の生成に使用する情報を拡張させる Retrieval Augumented Language Model という手法が多く研究されている [1, 2, 3]。ユーザの問い合わせ文章から関連するガイドを検索し、その情報を LLM に与えることによってより正確な情報を含んだ回答が自動的に生成できると考えられる。現在、関連するガイドの検索手法としては問い合わせ文章と検索対象を単語の一致度やベクトル表現などで直接評価する手法 [4] が一般的であるが、文意を考慮できないため正しいガイドを推薦できないことがある。そこで本研究では問い合わせ文章を過去の問い合わせと返信文を経由することで文意を考慮した上で関連ガイドを高精度に推薦する手法を提案する。

2 関連研究

2.1 文章類似度検索

従来の文章検索分野においては、単語の出現頻度を元にした疎ベクトルで文章を表すことが一般的であった [5]。単語の重要度をどのように重み付けるかによって様々な手法が存在し、最も有名なものは TF-IDF[6] やそれを拡張した BM25[7] である。しかしこれらの疎ベクトルを用いた手法は文章が正確に一致していないと検索できないという制約がある。そこで、近年ではニューラルネットワークを用いた文章検索手法が多く提案されている [8, 9, 10]。これらは文章を潜在空間内の密ベクトルとして表現し、ベクトル間の距離を計算することで類似度を判定している。本研究では後者のニューラルネットワークをベースとした文章検索技術を改良した手法を提案する。

2.2 Retrieval Augmented Language Model

Retrieval Augmented Language Model(RA-LM) は言語モデル (LM) に外部から取得した関連文章を与えることで、生成文章の質を向上させることができる枠組みである [11]。近年、LM のパラメーター数の急激な増加によりモデルの再学習コストが高くなっ

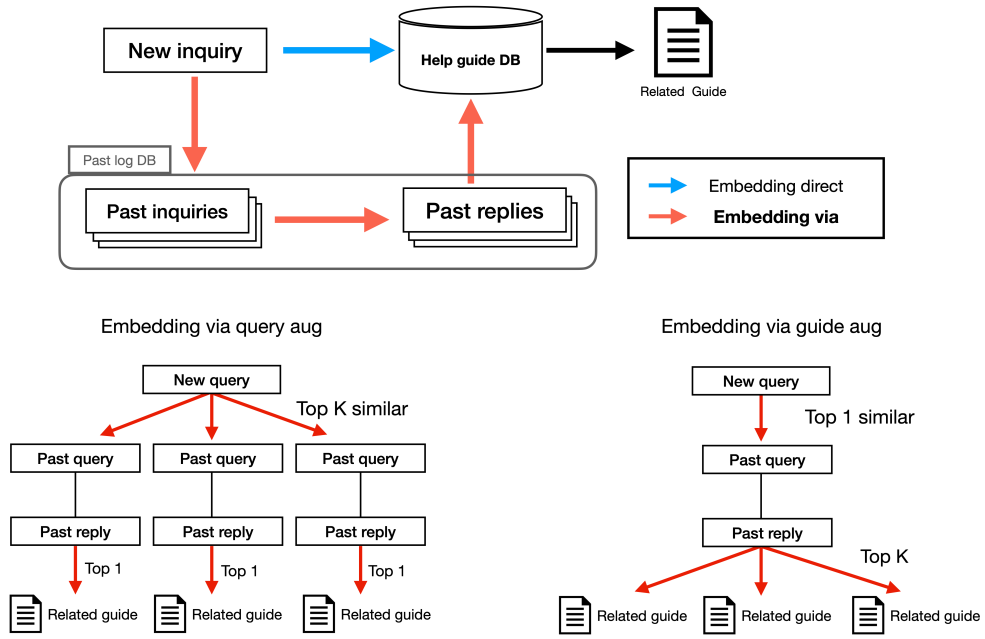

図 1　提案手法の概念図

ているため、簡易的に性能向上が見込める RA-LM が注目されている。RA-LM のうち Retriever 部分に関して盛んに研究がされており、代表的なものとしてはニューラルネットワークベースで教師あり学習が可能な DPR[12] や教師ラベルなしで学習可能な Retrieval Augmented Generation[13]、対照学習を行う Contriever[14] などがある。

3　提案手法

本研究ではユーザからの新規の問い合わせに関連するヘルプガイドを検索する際に、過去のユーザからの問い合わせとその返信文を経由する手法 **Embedding via** を提案する。提案手法の概要を図 1 に示す。従来の文章検索手法としては問い合わせ文章と検索対象群のベクトル表現に変換しその中で類似度を直接計算することが一般的である [5]。しかし質問と回答の文章では文体が異なることがあるため、類似度を直接計算すると正しく評価されない場合がある。そこで、問い合わせは問い合わせ同士、回答は回答同士で文章の類似度を比較することで類似ガイド推薦の精度を上げるため、本手法を提案する。なお、本論文において、以後ユーザからの問い合わせを**クエリ**と表現する。

3.1　Embedding via の検索過程

過去のクエリとそれに紐づいている返信文を経由することで類似度上位 K 個の関連文章を推薦する際に、検索の幅を増やす方向としては

1. 新規クエリと類似度が高い過去クエリ上位 k 件を取得
2. 返信文と類似度が高いガイド上位 k 件を取得

という 2 種類が存在する。本研究では各方向で検索幅を増やす手法をそれぞれ Embedding via query aug, Embedding via doc aug と名付け性能を比較する。

3.1.1　Embedding via query augment

新規のクエリを q_n、過去のクエリ群を Q_p として設定する。文書のベクトル表現を獲得する関数を $E(\cdot)$ とし、ベクトル同士の類似度を計算する関数を $S(\cdot)$ とする。式 3 の様に、ベクトル空間に埋め込んだ q_n と $q_p(\in Q_p)$ 間で類似度を計算し、類似度の高かった上位 K 件を $QP_k(k = 1, 2, , , K)$ とする。QP_k に一意に紐づいている返信文を $RP_k(k = 1, 2, , , K)$ とし、RP_k それぞれに対して検索対象のヘルプガイド $g_i(i = 1, , , N_g)$ との類似度を計算する。各 RP_k から得られた類似度スコアが最も高い 1 件のガイドを集積し、検索の結果 $\mathbb{G}$ として出力する。本研究では文書のベクトル表現獲得のための $E(\cdot)$ に OpenAI 社

が提供する text-embedding-ada-002[15] の API を採用し、類似度関数 $S(\cdot)$ にはコサイン類似度を用いる。

$$QP_k = \{q_i | q_i \in rank(S(E(q_n), E(q_i))) < k\} \quad (1)$$

$$\mathbb{G} = \{g_i | g_i \in \arg\max_{r_k \in RP_k} (E(r_k), E(g_i))\} \quad (2)$$

3.1.2 Embedding via doc augment

3.1.1 の Embedding via query augment と同様に新規のクエリを q_n、過去のクエリ群を Q_p とし、ベクトル空間に埋め込む。Embedding via doc augment では式 3 の様に、ベクトル空間に埋め込んだ q_n と q_p 間で類似度を計算し、類似度の高かった上位 **1** 件を $\hat{q_p}$ とする。$\hat{q_p}$ に一意に紐づいている返信文を r_p とし、r_p と検索対象のヘルプガイド $g_i (i = 1, , , N_g)$ の類似度を計算する。類似度スコア上位 k 件のガイドを検索の結果 $\mathbb{G}$ として出力する。ベクトル表現獲得のための $E(\cdot)$、類似度関数 $S(\cdot)$ も Embedding via query augment と同一である。なお、k=1 のとき、検索結果は Embedding via query augment と一致する。

$$\hat{q_p} = \arg\max_{q_i \in Q_p} S(E(q_n), E(q_i)) \quad (3)$$

$$\mathbb{G} = \{g_i | g_i \in rank(S\left(E(r_p), E(g_i)\right)) < k\} \quad (4)$$

4 実験

提案手法をメルカリ[1]に寄せられたお問い合わせ履歴に適用し評価を行った。また、提案手法の文章検索性能を比較するために、ベースラインとしてクエリと対象群の距離を直接計算する手法 Embedding direct と、疎ベクトルベースの検索手法である TF-IDF[6] と BM25[7] でも同様の実験を行う。

表 1 データセットの規模と文長

	文章数	文長平均値	文長中央値
検索対象ガイド	2,091	214	151
新規クエリデータ	10,000	159	109
過去返信文データ	100,000	569	507

4.1 データセット

社内で集積されたデータを用いて実験用データセットを構築した。使用する文章は全て事前に埋め込みモデルでベクトル化した。各データセットにおけるサンプル数および文章の長さの特徴を表 1 に示す。

検索対象ガイド 社内で使用しているヘルプガイドのうち、2023 年 10 月時点で外部公開状態になっている 676 件[2]を使用。1 つのヘルプガイドは複数のセクションから成っており、本研究ではセクション単位で切り出した 2,091 件の文章を検索対象とした。これらの文章は text-embedding-ada-002 によって予めベクトル化して保持した。

クエリ文章 実験に用いるクエリ文章としてはメルカリに対して実際にユーザから受けたもののうち、関連するヘルプガイドが紐付け可能なもののみを使用した。1 つのクエリ文章に対して正解ガイドが 1 件のみ対応している。新規クエリ q_n は 2023 年 10 月中に受けた問い合わせから 10,000 件をランダムにサンプリング、過去クエリ q_p は 2023 年 8-9 月中に受けた問い合わせから 100,000 万件をランダムにサンプリングして用意した。使用した問い合わせには全て社内のエージェントが作成した返答文が紐づいている。

4.2 比較手法

Embedding direct Embedding direct では、新規クエリ q_n から 3.1 と同様にベクトル表現 $E(q_n)$ を獲得し、これと検索対象群との類似度を計算する。その結果、類似度が上位 k 件であったガイドを検索の結果 $\mathbb{G}$ として出力する。

BM25 BM25 は文章 d を形態素解析によって単語レベルに分解した上で以下の式 5 によって計算される。ここで、$IDF(w_i)$ は検索対象の全文章中において単語 w_i が希少であるかを示す値であり、$f(w_i, d)$ は文章 d 中での単語の出現頻度、$|d|$ は文章 d の単語数である。BM25 を用いた比較手法では、式 5 でクエリ文章と検索対象ガイドの類似度スコアを計算し、あったガイドを検索の結果 $\mathbb{G}$ として出力する。

$$Score = \sum_i IDF(w_i) \frac{(k_1 + 1) f(w_i, d)}{f(w_i, d) + k_1 (1 - b + b \frac{|d|}{avg(dl)})} \quad (5)$$

4.3 評価方法

各手法で類似度が高いと算出された上位 k 件 $(k = 5, 10)$ の候補ガイドに対して Success Rate(SR)@k と Mean Reciprocal Rank(MRR)@k を計算して評価する。

SR とは Sakata らの研究 [16] で用いられている指標であり、検索された上位 k 件の候補文章の内に一

1） https://jp.mercari.com/

2） https://help.jp.mercari.com/guide/

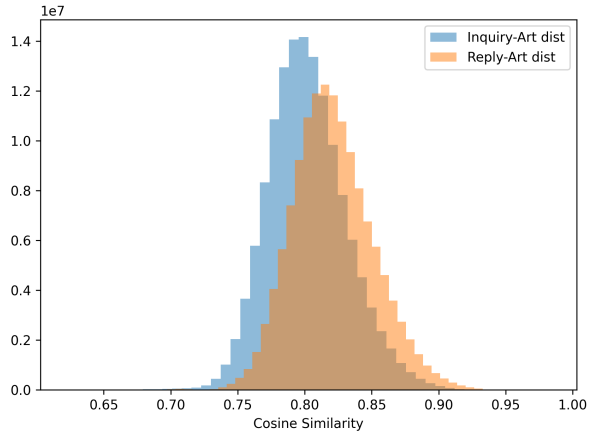

図 2　類似度の分布

つでも正解の関連文章 $\hat{g}$ が含まれていたクエリの割合である (式 6)。また、MRR@k は式 7 に示すように、検索された上位 k 件の文章を関連度に沿って降順に見て行った際に、最初に関連文書が含まれていた順位 $rank_i$ の逆数を加算した値である。

$$SR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} s_i, s_i = \left\{\begin{matrix} 1\ (\hat{g} \in \mathbb{G}) \\ 0\ (\hat{g} \notin \mathbb{G}) \end{matrix}\right\} \tag{6}$$

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{7}$$

4.4　結果と考察

提案手法と比較手法を用いて文章検索を行った時の評価結果を表 2 に示す。結果から Embedding via query augment/doc augment は共にベースラインの手法より性能が高いことがわかる。これはクエリとヘルプガイドの距離よりも返答文とヘルプガイドの距離が近いため、より正確に類似度を計算できているためだと考えられる。その検証としてクエリと返答文それぞれに対してヘルプガイド文章と類似度を計算した時の分布を図 2 に示す。図 2 から、返答文とヘルプガイドの類似度の方が全体として高い傾向が見られた。

また、提案手法の中では doc augment よりも query augment の方が性能が大幅に高い傾向が見られた。これは新規クエリに対して複数の類似過去クエリから関連ガイドを検索することによって推薦結果に多様性が生まれ、正解の文章が含まれている確率が高くなったからではないかと推測される。この考察を検証するため、Embedding via において 10 件の関連関連を検索する際に過去クエリと類似ガイド

表 2　評価結果

	@5		@10	
	SR	MRR	SR	MRR
BM25	0.147	0.064	0.302	0.84
Emb direct	0.324	0.188	0.496	0.211
Emb via doc aug	0.359	0.236	0.462	0.250
Emb via query aug	**0.428**	**0.259**	**0.549**	**0.275**

表 3　Embedding via n-query m-doc の検証結果

	query	doc	SR@10	MRR@10
Emb via	1	10	0.462	0.250
	2	5	0.495	0.249
	5	2	0.511	0.253
	10	1	0.549	0.275

の取得数の組み合わせをそれぞれ n 及び m 件に拡張し、表 2 と同様のデータセット、評価指標で比較実験を行った。その結果を表 3 に示す。この表における (query=1, doc=10) と (query=10, doc=1) はそれぞれ Embedding via doc augment (3.1.2) と Embedding via query augment (3.1.1) と同義である。結果として同じ件数のガイドを検索する際に、過去の類似クエリの数を多くするほど性能が高い傾向が見られた。

5　おわりに

本研究ではユーザからの新規のクエリに対して、類似する過去クエリと返信文を経由することで関連するヘルプガイド文章を検索する手法を 2 種類提案した。大規模な社内データを用いて文章検索実験を行った結果、提案手法は共にベースラインである直接的に文章の類似度を検索する手法よりも良い性能であり、特に過去類似クエリを多く取得する query augment の手法が高い性能を示した。類似する過去クエリ方向に検索幅を広げることの重要性を検証するため、使用する過去クエリと返信文からの類似ガイドの件数を複数変えて同様の実験を行った結果過去クエリの数が多くなると共に性能が上がる傾向が見られた。

本手法を用いて提案されたヘルプガイドは、ユーザの問題解決に有用となるだけでなく、LLM を Retrieval Augumented Generation システムと組み合わせることにより正確な回答文章を自動的に行うことができると考えられる。

今後の課題としては本手法が他環境のデータセットでも同様に有効であることを実証する必要がある。

謝辞

本研究を遂行するにあたり協力, 指導いただいた株式会社メルカリ社内の全ての方に御礼申し上げます。

参考文献

[1] Bohan Li, Yutai Hou, and Wanxiang Che. Data augmentation approaches in natural language processing: A survey. **Ai Open**, Vol. 3, pp. 71–90, 2022.

[2] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. **arXiv preprint arXiv:2007.00808**, 2020.

[3] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. **arXiv preprint arXiv:2010.08191**, 2020.

[4] S. Ibrihich, A. Oussous, O. Ibrihich, and M. Esghir. A review on recent research in information retrieval. **Procedia Computer Science**, Vol. 201, pp. 777–782, 2022.

[5] Amit Singhal, et al. Modern information retrieval: A brief overview. **IEEE Data Eng. Bull.**, Vol. 24, No. 4, pp. 35–43, 2001.

[6] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, Vol. 28, No. 1, pp. 11–21, 1972.

[7] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. **Nist Special Publication Sp**, Vol. 109, p. 109, 1995.

[8] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In **Proceedings of the 22nd ACM international conference on Information & Knowledge Management**, pp. 2333–2338, 2013.

[9] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. Learning semantic representations using convolutional neural networks for web search. In **Proceedings of the 23rd international conference on world wide web**, pp. 373–374, 2014.

[10] Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Vol. 24, No. 4, pp. 694–707, 2016.

[11] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. **arXiv preprint arXiv:2101.00774**, 2021.

[12] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. **arXiv preprint arXiv:2004.04906**, 2020.

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474, 2020.

[14] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. **arXiv preprint arXiv:2112.09118**, 2021.

[15] OpenAI. New and improved embedding model[3], 2022.

[16] Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. Faq retrieval using query-question similarity and bert-based query-answer relevance. In **Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval**, pp. 1113–1116, 2019.

3) https://openai.com/blog/new-and-improved-embedding-model