Retrieval-augmented generation に基づく カスタマーサポートにおける返信メール自動生成の検討

小島淳嗣 [1]
[1] 株式会社マネーフォワード
kojima.atsushi@moneyforward.co.jp

概要

企業のカスタマーサポートにおけるカスタマーへの返信メール作成業務を支援するため、返信メールを自動で生成するシステムを検討する。システムは retriever model と プロダクトについて学習させた large language model (LLM) によって構成され、retrieval augmented generation (RAG) に基づいて返信メールを生成する。メールの生成において、retriever model は問い合わせメールの回答に該当する情報を含む topN のチャンクをサポートサイトのページと返信メールテンプレートから抽出する。次に LLM は retriever model によって選択された チャンクとカスタマーのメール文を入力として返信メールを生成する。LLM は事前に汎用ドメインで事前学習を行った後、サポートサイトのテキストを用いて継続事前学習を行う。さらに、カスタマーの問い合わせメールと実際の返信メールをそれぞれ prompt と completion とみなして supervised fine-tuning を行った。実験では、 継続事前学習や RAG の有効性、及びエラー分析の結果について報告する。

1 はじめに

large language model (LLM) の高いテキストの生成能力が着目されており、医療、金融、法律など様々な領域において活用が広がっている [1, 2, 3]。このような専門知識を要するタスクでは、LLM に正確な出力をさせるため、retrieval-augmented generation (RAG) [4] に基づく推論や LLM の fine-tuning が実施される。

本稿では、カスタマーサポートにおけるカスタマーへの返信メール作成業務を支援するため、返信メールを自動で生成するシステムを検討する。システムは retriever model とプロダクトについて学習させた LLM によって構成され、RAG に基づきカスタマーへの返信メールを生成する。なお、問い合わせメールには、しばしば confidential な情報が含まれるため、OSS の事前学習モデルをベースに独自で fine-tuning を実施した。

LLM のドメイン適応は、継続事前学習 [5] と supervised fine-tuning (sft) の 2 段階で行う。継続事前学習では、汎用ドメインのテキストで事前学習されたモデルに対して、サポートサイトや問い合わせ履歴を用いて学習を行う。sft では、継続事前学習で学習されたモデルに対して、メールの問い合わせと回答の履歴をそれぞれ prompt と completion とみなすことで学習を行う。

実験では、マネーフォワードのカスタマーサポートのデータを用いてシステムを実装し、継続事前学習や RAG の有効性等について評価する。また、実際のベテランのカスタマーサポート 1 名によるエラー分析の結果についても報告する。

2 返信メール生成システム

図 1 に返信メール生成システムの概要を示す。システムは、retriever model と プロダクトについて学習させた LLM によって構成され、RAG に基づきカスタマーへの返信メールを生成する。retriever model は、問い合わせメールの回答に該当する topN のチャンクをサポートサイトと返信メールテンプレートから選択する。次に、LLM は retriever model によって選択された チャンクとカスタマーのメール文を入力として返信メールを生成する。

3 LLM の学習

LLM のドメイン適応は、継続事前学習と sft の 2 段階で行う。継続事前学習では、あらかじめ 汎用ドメインのテキストで事前学習されたモデルに対して、サポートサイトや問い合わせ履歴を用いて学習を行う。sft では、メールの問い合わせと回答の履歴

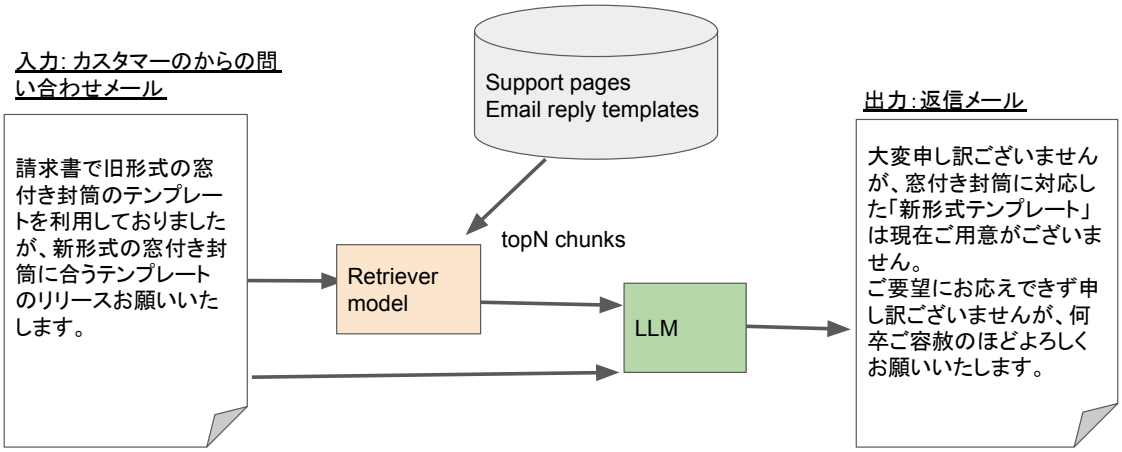

図 1　返信メール生成システムの概要

をそれぞれ prompt と completion とみなすことで学習を行う。

3.1　継続事前学習

継続事前学習ではチャット、メール、電話などのプラットフォームからのカスタマーの問い合わせ、カスタマーサポートによる実際の回答、サポートサイトのテキスト [1)]、返信テンプレートテキストを用いて学習を行った。表 1 にデータセットの内訳を示す。

表 1　継続事前学習のデータセット

Data	Disk size
サポートサイト	6.7 M
メール・チャット対応履歴	264.5 M
返信用テンプレートメール	59.4 M

3.2　sft

sft はメールの問い合わせと回答をそれぞれ prompt と completion とみなして実施する。図 2 に sft におけるモデルの入力フォーマットを示す。prompt と completion の境界を区別するために特殊な文字列として「### 返信メール:」を使用した。prompt はカスタマーからの問い合わせメールの内容と、問い合わせメールの回答を含む チャンクで構成される。チャンクは、学習時には ground truth を利用し、推論時には retriever model によってサポートサイトからチャンクとして選択される。学習では、カスタマーサポートによって、返信メールにおいて参考リンクとして記載されたページのテキストから効率的に ground truth のチャンクを作成した。sft では prompt を構成する トークンの loss を 0 にして、completion を構成するトークンの loss のみ最小化することでモデルを学習した。

```
### 指示: あなたは株式会社マネーフォワードのカスタマーサポートです.
ソースを考慮して, カスタマのメールへの丁寧な返信文を生成してください.

### カスタマからの問い合わせメール : <メール文>

### ソース1: <チャンク1>

### ソース2: <チャンク2>
        :
        :

### ソースN: <チャンクN>

### 返信メール:
```

図 2　sft における入力フォーマット

sft のデータの品質を担保するため、学習データのフィルタリングと整形を行った。具体的にはスパムメールや空メールは学習から覗き、署名などの情報は削除した。これらの処理を行い、カスタマー対応履歴から 8055 件の sft データを作成した。

4　実験

4.1　実験条件

全ての実験は、マネーフォワードのカスタマーサポートにおける実際のカスタマーからの問い合わせとその返信メールを用いて行った。評価データには、継続事前学習と sft の学習データ に含まれていない問い合わせデータ 100 件 を用いる。評価尺度には ROUGE-2 を採用し、実際にカスタマーサポートが返信したメールのテキストと システムによって生成された返信メールのテキストを用いて計算した。

1)　https://biz.moneyforward.com/support

LLM は、 C-4 や wikipedia、C-00 などの汎用ドメインのテキストで事前学習されたモデル [2] を用いて継続事前学習と sft を実施した。retriever model には Bidirectional Encoder Representations from Transformers (BERT) [6] に基づく埋め込みモデル [3] と BM25 を採用し、性能を比較した。チャンクはサポートサイトのテキストからチャンクサイズ 128、オーバラップ比率 25% で切り出すことで作成し、271122 のチャンクを得た。推論時には、チャンクの中から 最もスコアが高いものを prompt に入力するチャンクとした。ただし、スコアには閾値を設け、この値を上回るチャンクのみ選択し、該当するチャンクが存在しない場合には、prompt に チャンクを入力しないこととした。また、LLM のコンテキスト長を超えないように最大で選択するチャンク数は 4 つとした。

表 2 に LLM の fine-tuning の条件を示す．Adam

表 2 LLM の学習条件

Parameter	Value
Learning rate (継続事前学習)	0.00001
Learning rate (sft)	0.00002
Optimizer	Adam
Weight decay	0.01
Number of epochs (継続事前学習)	5
Number of epochs (sft)	2
Batch size	32

optimizer [7] のハイパーパラメータは $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10$ とした．

4.2 結果

表 3 に結果を示す。表において no retriever は、prompt に チャンクを入力しない時の結果を示す。ground truth は ground truth のチャンクを prompt に入力した時の結果を示す。retriever model を 用いない推論結果 (C0) に比べると、BM25 (D0) と BERT (E2) に基づく retriever model を用いて推論を行うことで、ROUGE-2 のスコアが高くなった。また、BM25 (D0) と BERT (E2) を比較すると、BERT が高い性能を示した。 また、ground truth のチャンク を prompt に入力した結果 (A0) に比べ、BERT に基づく retriever model によって チャンクを選択した結果 (E2) は、retriever model のエラーによってわずかに精度が低下した。

2） rinna/japanese-gpt-1b
3） intfloat/multilingual-e5-base

表 3 effect of retriever model

Exp ID	method	ROUGE-2
C0	no retriever	0.19
A0	ground truth	0.246
D0	BM25	0.22
E2	BERT	0.23

表 4 に継続事前学習の効果の比較を示す。継続事前学習を行わずに学習された sft モデル (F0) の結果に比べ、継続事前学習を行った sft モデル (E2) は、ROUGE-2 のスコアが高くなった。また、 事前学習モデルの結果 (B0) を比べると、継続事前学習の有無にかかわらず、sft を行うことで、ROUGE-2 のスコアが向上することがわかった (F0, E2)。

表 4 Effect of training process

Exp ID	継続事前学習	sft	ROUGE-2
B0	–	–	0.11
F0	–	✓	0.22
E2	✓	✓	0.23

表 5 に BERT に基づく retriever model における チャンクサイズの影響を示す。チャンクサイズを変えることで ROUGE-2 のスコアが最大 8 ポイント変化しており、チャンクサイズのチューニングが生成において重要であることがわかる。

表 5 BERT に基づく retriever model におけるコンテキスト長の影響

チャンクサイズ	ROUGE-2/recall/precision
64	0.2 /0.19/0.29
128	0.27/0.26/0.35
256	0.23/0.27/0.27
512	0.19/0.19/0.22

図 3 に実際にシステムの出力例を示す。システムの出力には、ユーザが LLM によって生成された内容が事実かどうかを効率的に確認できるように、retriever model によって選択されたチャンクが作成された元のページの URL とともに表示することとした。この例では、retriever model によって問い合わせに対する回答が記載されているサポートサイトから 適切なチャンクが選択され、返信メールが LLM によって生成できている。

カスタマーからの問い合わせ
インボイス制度に対応した新形式テンプレートでは「窓付き封筒対応」のご用意がありません。対応予定を教えてください。

生成結果
お問い合わせいただいた件について、ご案内いたします。
誠に恐れ入りますが、インボイス制度に対応した新形式テンプレートでは、「窓付き封筒対応」のご用意がございません。
ご要望にお応えできず申し訳ございませんが、今後の開発をお待ちいただけますと幸いです。
今後ともマネーフォワードをご愛顧くださいますようお願いいたします。

Reference
マネーフォワード クラウド請求書 API スタートアップガイド（v3ご利用者さま向け）| マネーフォワード クラウド請求書サポート
https://biz.moneyforward.com/support/invoice/guide/api-guide/a04.html

請求書の「郵送依頼」機能をリリースしました | マネーフォワード クラウド請求書 Plusサポート
https://biz.moneyforward.com/support/receivable/news/new-feature/n014.html

「窓付き封筒」にあったテンプレートは、ありますか？ | マネーフォワード クラウド請求書サポート
https://biz.moneyforward.com/support/invoice/faq/document/d11.html

図 3 システムの出力例

4.3 エラー分析

システムのさらなる改善のためカスタマーサポートに所属するベテランの社員一名によるシステムの出力分析を行った。その結果、2 つの問題によって不正確な返信文や hallucination が起きることがわかった。

1 つめの問題としては、retriever model がアクセスできないリソースの情報を必要とする問い合わせに対して、hallucination が起きうることがわかった。具体的には、カスタマーの契約状況やリアルタイムの情報 (障害状況) に関する問い合わせは、サポートサイトに情報が記載されていないため、生成されたテキストに hallucination が含まれることがある。例えば、ログインができないといった問い合わせに対してシステムは、理由がシステム障害であっても、パスワードとの入力ミスを確認するように返信することがある。この問題を解決するため、今後は、sft のデータセットを作成する際には、カスタマーの契約状況やリアルタイムの情報を必要とする関する問い合わせ対しては、completion を「回答できない」にして学習することで、この問題の解決を検討する。

2 つめの問題点として、異なるプロダクトの類似した機能に関する問い合わせに対して、retriever model が問い合わせ対象とは異なるプロダクトのサポートサイトから チャンク を抽出してきてしまうことで、LLM が 誤った回答を生成してしまうことがあった。今後は、チャンクにメタ情報として、プロダクトの種類を付与し、チャンクを選択する際に、問い合わせ対象のプロダクトに該当するサポートサイトから切り出されたチャンクに制限することで、この問題の解決を検討する。

5 おわりに

企業のカスタマーサポートにおけるカスタマーへの返信メール作成業務を支援するため、問い合わせメールへの返信を生成するシステムを検討した。システムは retriever model と プロダクトについて学習させた LLM によって構成され、RAG に基づいて返信メールを生成する。LLM は事前に汎用ドメインで事前学習を行った後、サポートサイトのテキストを用いて継続事前学習を行う。さらに、カスタマーの問い合わせメールと実際の返信メールをそれぞれ prompt と completion とみなして sft を行った。今後は、エラーの分析の結果に基づき、システムの改善に取り組む。

参考文献

[1] Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. CoRR, *arXiv preprint arXiv:2305.15062*.

[2] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, You Zhang. 2023. ChatDoctor: A medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. CoRR, *arXiv preprint arXiv:2303.14070*.

[3] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, Gideon Mann. 2023. BloombergGPT: A large language model for finance. CoRR, *arXiv preprint arXiv:2303.17564*.

[4] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Proceedings of the 34st International Conference on Neural Information Processing Systems, page 9459–9474.

[5] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342-8360.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

[7] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. CoRR, *arXiv preprint arXiv:1412.6980*.