

# FAQ チャットボットの誤りタイプの類型化と自動分類の検討

山岸 駿秀 貞光 九月 北岸 郁雄  
株式会社マネーフォワード

{yamagishi.hayahide, sadamitsu.kugatsu, kitagishi.ikuo}@moneyforward.co.jp

## 概要

FAQ チャットボットを導入した組織では、回答に用いる FAQ 集を継続的に改善する必要がある。FAQ 集の改善では、FAQ 集の内容等を網羅的に把握し、正しく回答できなかった事例に対して適切な改善策を選ぶ。本研究では、このような事例の誤りタイプを類型化し、誤りタイプ別に改善策をまとめる。また、誤りタイプの自動分類に必要な特徴量をまとめ、分類の可能性を議論する。

## 1 はじめに

企業等において、提供サービスに関するユーザの質問に自動で回答する FAQ チャットボットが普及している。従来からある Web 上の静的なサポートページはサービスの詳細な理解に役立つのに対し、チャットボットは高頻度な質問に簡潔かつ動的に回答できる。これらを図 1 のように併用することで、ユーザ自己解決率の向上などに繋げられる。

本研究で扱うチャットボットは、想定質問文と回答文からなる QA ペアを事前に登録した FAQ 集に対し、ユーザ質問文（クエリ）の関連 QA ペアを検索して回答する。本研究では、各クエリにつき一問一答で回答し、文脈等は考慮しないものとする。回答文ではクエリに直接関連する説明のみを行い、詳細は文中にリンクを張ったサポートサイトの記事内で説明することもある。回答後にはクエリ 1 つと回答文 1 つのペア（応対）が応対履歴として蓄積される。

チャットボットの運用担当者は、日常的には FAQ 集の整備による応対内容の改善を行う。例えば、新機能の情報が FAQ 集にないことが原因で誤答になる事例は、新機能に関する QA ペアを追加すれば改善できる。FAQ 集の改善は、FAQ 検索エンジンやチャットボットの UI 等の改良と比べて、運用担当者が速やかに実施しやすいことが利点となる。

チャットボットの運用担当者が FAQ 集を改善するためには、既存の FAQ 集の内容やサービスの機

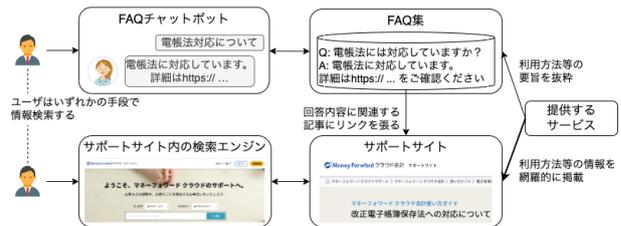


図 1 チャットボットとサポートサイトのイメージ図

能等を網羅的に把握している必要がある。しかし、サービスの機能が增加すると FAQ 集の内容も増加するため、全体の把握も困難になると考えられる。

本研究では、FAQ チャットボットの改善活動において、各応対事例の誤りタイプを類型化し、各誤りタイプに必要な改善策をまとめる。また、誤りタイプの自動分類に必要な特徴量を整理して分類器を構築し、自動分類の実現可能性を議論する。

## 2 誤りタイプの整理

FAQ 集を拡充する技術は盛んに研究されている。例えば、過去のクエリを「印刷関連」のようにジャンル別に整理し、拡充すべきジャンルを特定する方法 [1, 2] や、外部情報源から新規 QA ペアを抽出する方法 [3, 4] などがある。本研究では、FAQ 集の拡充で改善する応対事例は全体の一部であると考え、拡充が必要または不要な事例の特徴をまとめる。

チャットボットの改善方法を整理するため、応対履歴を分析して誤りタイプを調べる。チャットボットは、実験用として、BM25 による FAQ 検索をもとに構築した。FAQ 検索では、クエリと、FAQ 集の QA ペアに含まれる想定質問のみとの類似度を測り、類似質問に対応する回答文を得る。FAQ 集は、当社のカスタマーサポートがマネーフォワードクラウド会計・確定申告<sup>1)</sup>の質問に回答できるように整備したものを使う。FAQ 集の回答文には、同サービスのサポートサイト<sup>2)</sup>の関連記事（引用記事）のリンクを付与した。このチャットボットに、あらかじめ

1) <https://biz.moneyforward.com/>  
2) <https://biz.moneyforward.com/support/>

表1 FAQ チャットボットの回答誤りタイプ

カテゴリ	誤りタイプ	概要	クエリの例	回答文の例（抜粋）	改善策
改善可能	QA ペア検索失敗	FAQ 集にある正答の QA ペアを選択できていない	領収書のスキャンデータを保存したい	無料会員のご利用期間に制限はございません。	想定質問を拡充する
	引用記事のみ正解	回答文に解決方法の記述がないが、文内で引用された記事には記述がある	メンバーの追加ができません	3名まで基本料金に含まれます。詳細は <a href="https://...">https://...</a> をご覧ください	新規 QA ペアを追加
	あいまいな質問	質問文が曖昧で、適切な回答内容を判断できない	取得状態が設定エラーとなります	電子証明書の画面にエラーが表示される場合…	回答候補選択肢を出す
	回答手段なし	課題解決につながる QA ペアが FAQ 集になく、関連する記事もない	1月の役員借入金の合計を確認したい	各勘定科目の説明は以下の通りです。	新規 QA ペアまたは新規記事を追加
改善困難	サービス機能なし	ユーザ課題解決につながる機能がサービスにない	英語対応していますか？	現在日本語表示にのみ対応しています。	新規機能を開発する
	回答対象外	FAQ 集の対象外のサービスに関する質問	クラウド請求書の料金プラン	こちらはマネーフォワードクラウド会計の窓口です。	適切な窓口へ案内する
	質問の誤字脱字	質問文に誤字脱字があり、機械の可読性が低い	事業者登録しておらず、大会できません	事業者の情報を共有するためには…	再入力を促す

収集したユーザクエリ 1574 件を入力して対応履歴を得た。これらを目視で分析し、正解ではないと思われる対応 1030 件の誤りタイプを分析した。

分析を通して、各対応事例を、FAQ 集の改善により課題が解決できる**改善可能系**と、FAQ 集を改善しても課題が解決されない**改善困難系**の2つに大別した。さらに、**改善可能系**を4つ、**改善困難系**を3つの誤りタイプにそれぞれ細分類した。表1に、本研究で整理した計7つの誤りタイプを示す。

**改善可能系: QA ペア検索失敗** ユーザの課題解決に直結する QA ペアが FAQ 集内にあるが、FAQ 検索の精度不足で誤答となった場合が該当する。FAQ 検索の対象となる想定質問の拡充により、類似質問に対する関連 QA ペアのヒット率を高められる。

**改善可能系: 引用記事のみ正解** 回答文には「詳細は以下の記事をご覧ください」といった形でサポートサイトの記事（引用記事）へのリンクを含む場合がある。こうした回答文の一部では、課題解決につながる内容が回答文自体にはなく、引用記事のみ存在する状態が発生する。表1の例では、メンバー追加方法に関する質問に対してメンバー追加にかかる料金を説明している。この例の引用記事のタイトルは「「メンバー追加・管理」の使い方」であり、課題解決につながる内容を含む。しかし、ユーザは回答文を誤りと判断するため、引用記事に解決手段の記述があっても記事が読まれにくく、課題解決に繋がらない。回答文を読むだけで課題解決につながる QA ペアを追加することが改善策となる。

**改善可能系: あいまいな質問** クエリに、関連 QA ペアを絞るために必要な情報が不足している場合が

該当する。こうしたクエリでは、チャットボットの運用担当者でも、ユーザが望んでいる回答内容を判断できない。表1の例では、どの取得状態に関する質問かわからないため、電子証明書に関する回答が正しいかどうかもわからない。改善策には、限られた情報でも答えられる QA ペアの追加がある。前述の例では、代表的なエラー文を説明する QA ペアを追加すればよい。また UI を改良し、ユーザから追加情報を得る手段を実装することも改善策となる。

**改善可能系: 回答手段なし** 課題解決につながる QA ペアが FAQ 集になく、サポートサイトにも関連記事がないことから、チャットボットやサポートサイトに回答手段がない場合が該当する。FAQ 集は高頻度な質問への回答をまとめたものであるため、低頻度な質問や、ユーザの利用状況特有の質問などには対応できない。サポートサイトは FAQ 集よりも詳細な情報を持つ傾向にあるが、全てを網羅することは難しいため、対応できない内容もある。改善には、**回答手段なし**の中で高頻度な質問に関する QA ペアやサポートサイト記事を追加する必要がある。

**改善困難系: サービス機能なし** 提供するサービスに、ユーザの課題解決につながる機能がない場合が該当する。表1の例のように、回答としては現在非対応である旨を説明することが正しい。しかし、この回答ではユーザの課題は解決しない。根本的な改善には、サービスの改善が必要である。

**改善困難系: 回答対象外** FAQ 集が対象外としている質問が該当する。表1の例はマネーフォワードクラウド請求書に関する質問であるが、今回用いた FAQ 集はクラウド会計・年末調整が回答対象である

表2 各誤りタイプへの分類に必要な特徴量（検索ヒット数には各誤りタイプにおける平均を記載）

誤りタイプ	クエリ	整合性（○：整合性が高い、×：整合性が低い）			検索ヒット数
		クエリと回答文	クエリと引用記事	クエリと検索記事	
QA ペア検索失敗	✓	×	×	○	3.7 件
引用記事のみ正解	✓	×	○	○	2.9 件
あいまいな質問	✓	×	×	×	5.9 件
回答手段なし	✓	×	×	×	1.1 件

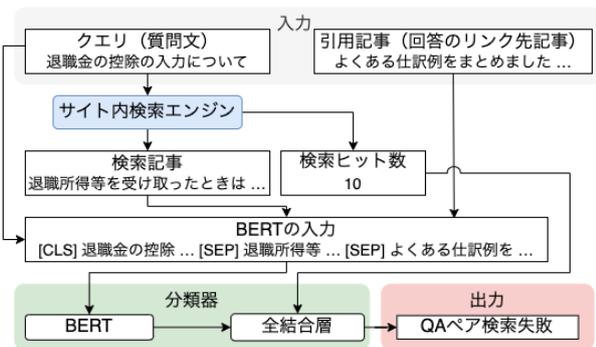


図2 誤りタイプ分類器の概要図

ため、回答できない。対象外の質問であることを検知し、適切な窓口へ誘導することが改善策となる。

**改善困難系: 質問の誤字脱字** 質問文に誤字脱字があることにより、適切なQAペアを発見できない場合が該当する。改善策としては、応対時に誤字脱字を検出して再入力を促すことが考えられる。

### 3 誤りタイプの自動分類の検討

本節では誤りタイプ分類の自動化を検討する。本研究ではFAQ集の改善で解決する誤りタイプを優先し、**改善可能系（QAペア検索失敗、引用記事のみ正解、あいまいな質問、回答手段なし）**の検出に必要な特徴量と分類器を設計する。**改善困難系**も実運用上は検出すべきであるが、今回は対象外とする。

#### 3.1 分類に必要な特徴量

分類に必要な特徴量の整理結果を表2に示す。

**検索記事の有無による分類** 2節の類型化から、誤りタイプの特定にはサポートサイトの記事が有用であると考え、4つの誤りタイプをユーザ課題の解決につながる記事（正解記事）の有無により2群に分ける。**QAペア検索失敗**と**引用記事のみ正解**は、どちらも正解記事を持つため、**記事特定可能群**とする。また、**回答手段なし**は正解記事がなく、**あいまいな質問**は正解記事を絞れないことから、どちらも正解記事がないとみなし、**記事不定群**とする。サポートサイトから正解記事を探索するため、質問文

を用いたサイト内検索を導入する。サイト内検索で得られた、ヒットした記事の件数（検索ヒット数）と、内容（検索記事）を特徴量に用いる。2群の区別には、クエリと検索記事の整合性を調べ、検索記事が正解記事かどうか確認することが重要となる。

**記事特定可能群内の分類** この群はクエリと引用記事の整合性を用いて分類を試みる。**QAペア検索失敗**の回答文は誤りであるので、引用記事はクエリと関連しない。**引用記事のみ正解**は回答文自体は誤りに見えるが、引用記事はクエリと関連する。

**記事不定群内の分類** この群は、クエリと検索ヒット数を用いて分類を試みる。**あいまいな質問**はクエリに含まれる情報が少なく、正解記事を絞れないため、検索ヒット数が多くなる。**回答手段なし**はクエリに含まれる情報が多く、全てに該当する記事が見つかりにくいいため、検索ヒット数が少なくなる。また、クエリに含まれる情報を直接抽出し、検索ヒット数の妥当性を確認することも重要である。

以上の議論から、クエリ、引用記事、検索記事、検索ヒット数の4つの特徴量を使うことで分類できると考えられる。なお、4つすべての誤りタイプでクエリと回答文との整合性は低いことが前提となるため、これらの分類に回答文は不要である。

#### 3.2 分類器の概要

図2に誤りタイプ分類器の概要を示す。分類器は、BERT [5] と1層の全結合層で構築する [6]。

BERTはクエリとサポートサイトの記事との整合性、または記事間の整合性を判定するために用いる。クエリと記事は特殊記号 [SEP] で繋いで入力する。複数の記事を用いる場合も、記事間を同じ特殊記号で繋ぐ。BERTの出力として、最終層の文頭記号 [CLS] に対応するベクトルを得る。

全結合層は、BERTから得られた整合性情報や検索ヒット数から最適な誤りタイプを推論する。入力には、検索ヒット数をBERTの出力ベクトルと結合させたものとする。学習時は、BERTと合わせて全体のfine-tuningを実施する。

表3 誤りタイプ分類の実験結果

実験パタン	A. クエリ	B. 引用記事	C. 検索記事	D. 検索ヒット数	Accuracy	F 値
1	✓	-	-	-	25.0	0.1859
2	✓	✓	-	-	45.0	0.4254
3	✓	-	✓	-	52.5	0.4966
4	✓	✓	✓	✓	60.0	0.5717
5	✓	✓	✓ (正解フラグを付与)	✓	77.5	0.7501

## 4 実験

### 4.1 実験設定

本節では、3節で述べた誤りタイプ分類の検証実験を行う。2節で用いた対応履歴から、各誤りタイプのデータ量を揃えるために各100件ずつサンプリングし、学習用、検証用、評価用データとして、320件、40件、40件に分けた。BERTの事前学習済みモデルにはbert-base-japanese-v2<sup>3)</sup>を使用した。このBERTは入力文長に512単語の制限がある。

記事特定可能群と記事不定群の分類には、クエリと検索記事の整合性を評価し、検索記事が正解記事かを判断することが重要となる。しかし、正解記事が得られるかはサイト内検索の検索エンジンの性能に大きく依存する。本実験では、検索で必ず正解記事が得られる状況を作り、検索エンジンの性能が極めて高い場合に整合性評価が機能することを確認する。記事特定可能群は、検索記事に正解記事を用いる。記事不定群は正解記事を持たないため、検索結果のうちヒット順1位の記事を検索記事とする。記事の検索にはWordpressの検索機能を用いた<sup>4)</sup>。

多くの記事は長文であるため、文長制限に収めるために検索結果のスニペットを検索記事とする。本実験では、各記事のクエリ関連箇所を目視で抜き出してスニペットとした。引用記事の取得には検索が不要のため、引用記事は記事の先頭5文を与える。

### 4.2 結果

表3に分類実験の結果を示す。3.1節で述べた特徴量を加えることで、分類精度が向上した。実験パタン1では、多くの分類結果がQAペア検索失敗に集まっていた。一方、引用記事を用いた実験パタン2では記事特定可能群内の分類精度が向上し、検索

記事を用いた実験パタン3では記事特定可能群と記事不定群を区別する精度が向上した。また、全ての特徴量を用いた実験パタン4では、各群内の分類精度の向上が見られた。例えば、表1の引用記事のみ正解のクエリ例「メンバーの追加ができません」は、パタン1では誤っているが、検索記事やヒット数を用いた実験パタン3,4,5では正しく分類できた。各実験結果の混同行列は付録5を参照されたい。

また、実験パタン5では、整合性評価が正しく行われた場合の精度を確認するため、記事特定可能群の検索記事の文頭に正解フラグ[TRUE]を付与し、整合性評価の正解を直接与えた。この実験で、整合性評価の精度を高めることで分類精度を高められることを確認した。実験パタン5でも誤分類となった事例は、あいまいな質問で多く見られた。例えば、表1のあいまいな質問のクエリ例「取得状態が設定エラーとなります」は、回答手段なしに分類された。あいまいな質問は検索ヒット数が他の誤りタイプより多い傾向にあるが、この例のヒット数は7と平均的であった。ヒット数が数十件の事例は正しく分類できていたため、平均的なヒット数では分類根拠としては不十分であったと思われる。したがって、記事不定群内の分類精度を高めるためには、クエリ文面から内容を抽出する精度向上と、記事検索の精度向上の双方が必要であると考えられる。

## 5 おわりに

本研究では、FAQチャットボットの対応の誤りタイプを類型化し、各誤りタイプに必要な改善策をまとめた。また、FAQ改善の効率化に向けて誤りタイプ自動分類を検討し、分類器の特徴量としてサポートサイトの検索結果が有用であることを示した。

今回の実験では、サポートサイトの記事検索精度が極めて高い状況を仮定したが、実運用に向けては、記事検索や、クエリと記事の整合性判定の精度向上を進める必要がある。また、今回議論の対象外とした誤りタイプも考慮した分類を進める。

3) <https://huggingface.co/cl-tohoku/bert-base-japanese-v2>

4) クエリには、質問文をMeCab(辞書:mecab-ipadic-neologd[7] 2020-09-14 Ver.)で形態素解析し、名詞・形容詞・動詞を抽出した結果を用いた。

## 謝辞

誤りタイプの体系の整理方法について繰り返し議論いただいた、株式会社マネーフォワード ビジネスカンパニー CS 本部の竹下晴基氏、古山三紗子氏に謝意を表します。

## 参考文献

- [1] 土居誉生, 森克利, 嘉門勇輝, 稲田徹. 対応ログからの FAQ 自動抽出システム NANQ. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 93, pp. 167–168, 2021.
- [2] 戸田隆道, 友松祐太, 杉山雅和. 提示候補とクエリの差分を用いたチャットボットの新規問い合わせ抽出手法. 言語処理学会第 27 回年次大会, pp. 1550–1553, 2021.
- [3] 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**, SIGIR'19, p. 1113–1116, New York, NY, USA, 2019. Association for Computing Machinery.
- [4] 坂田亘, 田中リベカ, 黒橋禎夫. 公式ウェブサイトをベースにした QA チャットボットの自動構築. 言語処理学会第 26 回年次大会, pp. 327–330, 2020.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [6] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.
- [7] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会 (NLP2017), pp. NLP2017–B6–1. 言語処理学会, 2017.

表 4 各実験の混同行列の抜粋

	1. クエリのみ				2. クエリ、引用記事			
True\Predict	検索失敗	記事正解	あいまい	手段なし	検索失敗	記事正解	あいまい	手段なし
検索失敗	7	0	2	1	3	3	2	2
記事正解	7	0	3	0	0	7	1	2
あいまい	8	0	2	0	1	3	2	4
手段なし	9	0	0	1	2	2	0	6
	3. クエリ、検索記事、ヒット数				4. 全て使用			
True\Predict	検索失敗	記事正解	あいまい	手段なし	検索失敗	記事正解	あいまい	手段なし
検索失敗	8	2	0	0	6	0	1	3
記事正解	7	3	0	0	1	6	2	1
あいまい	0	5	5	0	3	1	2	4
手段なし	0	1	0	9	0	0	0	10
	6. 全て使用（検索記事は正解フラグ付き）							
True\Predict	検索失敗	記事正解	あいまい	手段なし	検索失敗	記事正解	あいまい	手段なし
検索失敗	8	2	0	0				
記事正解	0	10	0	0				
あいまい	0	0	3	7				
手段なし	0	0	0	10				

## A 付録

### A.1 実験の混同行列

表 4 に、4.2 節の表 3 で示した各分類実験の混同行列をまとめた。

引用記事は、引用記事のみ正解とそれ以外を区別するために導入した。実験パタン 1 と 2 を比較すると、実験パタン 2 において引用記事のみ正解の分類精度が向上していることがわかる。また、回答手段なしは関連する記事がサポートサイトに存在しないため、引用記事が不正解となることから、こちらの分類精度も向上した。したがって引用記事は、引用記事のみ正解や回答手段なしを特定するために有用である。

検索記事は、正解記事が存在する記事特定可能群と、正解記事が絞り込めないまたは存在しない記事不定群を区別するために導入した。実験パタン 1 と 3 を比較すると、実験パタン 3 では両群を区別する精度が向上している。また、検索ヒット数は、記事不定群内を分類する目的で導入した。実験パタン 3 の各群を見ると、記事特定可能群内と比較して、記事不定群内の分類精度が高い。したがって、検索記事と検索ヒット数は、両群の区別や、記事不定群内の分類をする上で有用である。

実験パタン 5 は、クエリと検索記事の整合性評価が完全に機能した場合の分類精度を確認するため、整合性評価の正解を外部から与えた。整合性評価の正解を与えなかった実験パタン 4 と比較すると、他群へ誤って分類する例がなくなったことがわかる。したがって、検索記事の整合性評価の精度を高めることで、両群の区別の精度を高められる。一方、記事不定群では、あいまいな質問の多くが回答手段なしに分類された。これは、4.2 節で述べたとおり、あいまいな質問であっても検索ヒット数が少ない事例があったことが原因である。今後は、記事検索の精度を高めることで検索ヒット数の妥当性を高めるとともに、クエリの内容理解の精度を高めることで、検索ヒット数に過度に依存しないような分類器の構築が必要である。