問い合わせログの集約における クラスタリングを用いた重要回答抽出

竹中 一秀　　林 岳晴　　大段 秀顕　　湯浅 晃
株式会社 NTT データ
{Kazuhide.Takenaka, Takeharu.Hayashi, Hideaki.Odan, Akira.Yuasa}@nttdata.com

概要

本研究では、問い合わせデータを集約し FAQ を構築するユースケースに焦点を当てた回答候補を提示する手法を提案する。FAQ はコンタクトセンタにて蓄積される質問と回答が対となる問い合わせログを用いて作成される。本提案手法では、問い合わせログの質問を対象にクラスタリングを行い、類似する質問ごとに集約されたクラスタ内の回答に対し、再度クラスタリングを実施し、複数種類の重要な回答を抽出することで、可読性を保ちつつ、重要な回答を複数抽出できることが確認された。併せて、回答をクラスタリングした際に得られる各クラスタのクラスタサイズが、回答を提示する上での重要性を示す指標として有効であることが確認された。

1 はじめに

近年、企業のコンタクトセンタでは、問い合わせ対応業務にチャットボット等の対話型 AI システムを活用されている。チャットボットには、ユーザが選択する質問に対して、あらかじめ設計されたシナリオに従って回答を提示するシナリオ型と、ユーザの入力する質問に対して、類似する FAQ を提示する FAQ 型がある。その中でも FAQ 型のチャットボットは、その導入時や運用時において FAQ の作成に、多大な労力を必要とするという課題がある。そこで、本稿では FAQ 型のチャットボットに焦点を当てることにした。

FAQ はコンタクトセンタの問い合わせログに代表される、質問文と回答文が対となるデータから作成される。一般的な手法としては、問い合わせログの質問文に対してクラスタリングを行い、クラスタ重心に最近傍の質問文を抽出し、その質問文に紐づいた回答文とセットで FAQ 候補とする。FAQ 作成担当者は、得られた FAQ 候補を確認し、内容や文の修正を行うことで最終的な FAQ を完成させる。しかし、この手法では 1 つの質問に対し、1 つの回答のみ提示されるため、質問と回答が本来 1 対 n となる場合に、適切な回答が得られない可能性がある。質問と回答が 1 対 n になる場合の FAQ の例として、「ソフトが強制終了されてしまいます。どうすればよいでしょうか。」という質問に対して回答を作成する場合には、ソフトバージョンや OS バージョン、データのバックアップは取得しているか等の幅広い状況に応じた回答を揃える必要がある。

そこで、本研究では、FAQ 作成の負荷を低減することを目的として、可読性を保ちつつ、冗長性を排除し、重要な回答を漏らさないように回答候補を提示する手法を提案する。

2 関連研究

本研究では、FAQ 作成における回答文作成支援のユースケースに焦点を当てている。ここでは、FAQ 抽出に関する関連研究の手法について言及する。

まず、FAQ 候補となる QA を抽出する研究として、マニュアルやガイド等の回答源となる文書から QA セットの作成を行い、問い合わせログから抽出した代表質問文と類似する QA セットを FAQ 候補として提示する土居ら[1]の手法や、特徴的な単語の係り受けパターンの抽出を行い、頻出するパターンを持つ QA を抽出する長谷川ら[2]の手法がある。友松ら[3]は問い合わせログの質問文を BERT によって分散表現を獲得し、分散表現のコサイン類似度をもとに階層クラスタリングを行うことで、類似する質問ごとに集約されたクラスタを作成し、そのクラスタから TF-IDF を用いて代表文の選定を行い FAQ の抽出を行っている。これらの手法は、回答文については、抽出した質問文に紐づいたもののみを使用しているため、複数の回答文候補を抽出するという点で我々の手法と異なっている。また、壹岐ら[4]は、問い合

わせログから質問文と回答文の抽出を行うことを要約問題として捉え、教師あり学習による抽出型要約を用いて、不要文の多い問い合わせログの質問文と回答文の中から本質的な部分文字列を抽出することで QA の作成を行っている。我々の手法とは、文字列単位ではなく、問い合わせログの回答文単位での回答抽出を行う点で異なるが、重要な回答文を抽出するという点では類似しているため、文書要約を用いた手法を提案手法の比較対象とすることにした。

3 提案手法

本研究では、FAQ 作成の負荷を低減することを目的として、重要な回答を複数抽出する手法を提案する。本章では、その具体的な実現方法について説明する。

3. 1 回答抽出の流れ

提案手法の回答文抽出の流れを図 1 に示す。まず、提案手法適用の前提として、問い合わせログの質問を対象にクラスタリングを行い、各クラスタ重心の最近傍の質問文が得られているものとする。このクラスタ重心に最近傍となる質問文を本稿では代表質問文と呼ぶ。得られた質問クラスタ内の回答文を対象にさらにクラスタリングを行う。その結果、類似する回答ごとに集約されたクラスタが得られる。得られた回答クラスタの重心に最近傍となる回答文を代表文としてそれぞれ抽出することで、抽出された代表質問文に対して、複数の回答候補を取得する。

3. 2 手法検討

本研究では、FAQ 作成における回答文作成支援のユースケースに焦点を当て、ビジネス適用における現実的な制約条件を考慮した上で手法検討を行った。分散表現の取得手法として、BERT 等の大規模言語モデルは、モデルの学習を行うために、ドメインに特化した大量の学習データが必要になる。そこで、本研究では、汎用コーパスで学習されたモデルで比較的上質な分散表現が得られることが知られている手法として[5]、Universal Sentence Encoder(USE)[6]を用いた。クラスタリングは、階層クラスタリングのうち、分類感度が高い ward 法を選定した。

また比較対象とする文書要約手法については、抽出型文書要約手法の中で一般的に高性能である LexRank を用いた。

同クラスタ内の回答文から要約文を抽出し、FAQ 作成において、提案手法とどちらが有力であるか比較を行う。

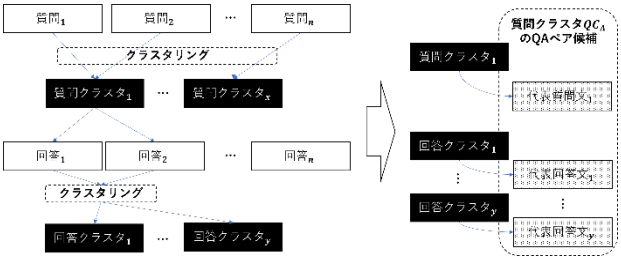

図 1　提案手法の回答文抽出イメージ

4 実験

提案手法の有効性検証のため、既存手法である、抽出した代表質問文に元々紐づいていた回答文のみ提示する場合と、比較対象とした文書要約手法それぞれを用いて、回答文の抽出を行い、FAQ 作成の支援を目的とした場合に、どの手法が有力であるか比較して検証を行った。抽出型文書要約手法 LexRank の実装に関しては python モジュールの sumy を用いた。

4. 1 データセット

手法の有効性検証のために、NII の Yahoo!知恵袋データ(第 2 版)[8]で提供されているデータのうち、質問カテゴリが「企業と経営」かつ、回答は質問者が選んだベストアンサーの条件で絞込んだものを用いた。また、実際の問い合わせログは、端的に短くまとめられていることが多いため、質問文の文書長が 10～50 トークンかつ、回答文の文書長が 10～120 トークンからなる QA をサンプリングした。その結果、データサイズは 4153 件となった。

次に、サンプリングしたデータの質問文を対象に USE＋ward 法を用いてクラスタリングを行い、類似する質問文ごとに集約された各クラスタのベクトル重心に最近傍の点を代表質問文として、抽出を行った。そのうち、クラスタサイズが大きいクラスタから 6 件を対象に実験に使用した。各クラスタのクラスタサイズと抽出した代表質問文を表 1 に示す。

表 1　実験に使用した質問集約のクラスタ

クラスタ ID	クラスタ サイズ	代表質問
1	18	会長と社長はどう違うのですか?・組織的に・対外的に・権力的に
2	17	前株と後株はどう違うのですか?何か意味があるのでしょうか?どちらの方が良いとかあるのでしょうか?
3	12	有限会社がなくなると前々から聞いているのですが本当ですか?これからは個人と株式だけになるのでしょうか?
4	10	「資本金の額」が大きい場合の企業メリット&デメリットを教えて下さいませんか?
5	10	子会社とグループ会社って何か違いはあるのですか?実質、同じのような気がするんですが・・・。
6	9	ISO9001.14001 という言葉をよく耳にしますが、何のことだかさっぱりわかりません。わかり易く教えてください。

4.2 評価方法

FAQ 作成支援において、有効性を評価するために、それぞれの手法で出力した回答文を可読性と、回答として重要な文がどのくらい含まれているかを示す重要情報包含率の観点で評価を行った。可読性については、表 2 に示す基準で定性的に評価を行った。また、抽出した回答文に、「文間の内容が矛盾しており、一貫性が無い」「文間に談話関係が存在せず、無関係である」といった問題がある場合を「不自然である」と評価し、上述の問題が無い場合には「自然である」と評価した。重要情報包含率の算出方法は、抽出した代表質問文から必要となる回答箇所を人手により正解として定義し、その中で各手法によって抽出できた割合を算出することで評価した。

表 2　可読性の評価基準

評価基準	判定
文章として、自然である	○
文章として、不自然である	×

5 実験結果と評価

提案手法の有効性検証の比較対象として、大量の文から重要な文を抽出する文書要約アルゴリズムの LexRank[7]を用いた。実験では、質問集約で得られたクラスタ内の回答文を全て結合した文を入力とした。また、ハイパーパラメータの出力文数は、出力された要約文が提案手法の出力した回答文と近しいトークン数になる文数を採用した。

提案手法のクラスタリングを行う際のクラスタ数については、Silhouette 係数が最大になる場合を採用した。既存手法については、表 1 で得られた代表文に紐づいた回答文のみを出力結果として扱った。

実験結果について、可読性の評価結果を表 3、情報含有率の評価結果を表 4 に示す。情報包含率については、各手法ごとに平均化したものを示す。また、実験によって得られた回答文の一部を表 6 に示す。太字部分は重要情報となる箇所、下線部分は可読性を低下させている箇所を示している。

表 3　可読性評価

クラスタ ID	質問集約 代表 QA ペア	LexRank	USE+word (Proposal)
1	○	×	○
2	○	×	○
3	○	×	○
4	○	×	○
5	○	○	○
6	○	○	○

表 4　重要情報含有率評価

質問集約 代表 QA ペア	LexRank	USE+word (Proposal)
0.72	0.50	**0.86**

表 3 によると、提案手法は可読性については問題ないことが示されている。LexRank は要約結果を文単位で得るため、文頭に「つまり」「これにより」があることや、関連性のない文間に「よって」「ゆえに」等の接続詞があり、文間の論理的な関係が破綻していることが確認された。

表 4 によると、提案手法が最も良い結果となっている。既存手法の抽出した代表質問文に紐づいた回答 1 件のみでは、情報として不足する場合も少なくなかった。表 6 を例に、既存手法と比較して提案手法では、「特例有限会社は、そのまま存続できます

し、一部例外を除き変更登記する必要もございません。」「特例有限会社から株式会社へ移行すること（商号変更）も出来ます。」との追加の重要文が抽出できていることが確認できる。補足として、既存手法で用いた回答は yahoo 知恵袋に投稿した質問者が選んだベストアンサーであるが、ベストアンサーは投稿された回答群の中から 1 つ選ばなければいけない仕組みであるため、情報として不足していてもベストアンサーに選ばれる場合は十分にある。

6 付帯効果と課題

FAQ を作成する上で、複数種類の回答がある場合に、重要度の高い回答を優先して提示することが重要であると考えられる。この場合、提案手法では、クラスタサイズが大きいクラスタほど、重要度が高い文が含まれている傾向を確認したため、重要度を表す指標として、回答クラスタサイズを用いることが、有効であると考えられる。また、提案手法の出力結果を文単位で確認したところ、内容が重複する文があることが判明した。その数を表 5 に示す。出力する回答数が増加するほど、重複文数も増加し、FAQ 作成担当者が確認する文書量も比例して増加してしまう。そのため、クラスタリング手法の改善をすべきであり、それでも対処しきれない場合は文間の類似度を算出してフィルタリングを行うなどの工夫を行う必要がある。

表 5　重複内容文数

クラスタ ID	出力 回答数	重複文数
1	3	2
2	2	1
3	2	1
4	1	0
5	2	0
6	1	0

7 おわりに

本研究では、問い合わせログデータを用いた FAQ 作成の支援を目的とした回答文抽出において、重要な回答文を複数抽出する手法を提案し、検証・評価を行った。既存手法や文書要約手法と比較して、提案手法は可読性を保ちつつ、重要な情報を抽出できることが確認された。また、精度以外のメリットとして、クラスタサイズ（回答数）を重要度のスコアとして提示できる点が挙げられる。

今後の課題として、提示する回答文が多くなる程、内容が重複した不要な文が増えてしまい、FAQ 作成担当者が確認する情報量が増えてしまうため、文間の類似度を算出してフィルタリングを行うなどの工夫が必要である。

表 6　各手法での回答文抽出結果（クラスタ ID = 3 の場合）

太字部分は重要情報となる箇所、下線部分は可読性を低下させている箇所を示している。

手法	抽出した回答文
質問集約 代表 QA ペア	**会社法が施行されてからは、新たに有限会社を設立することはできなくなりました。従前の有限会社は特例有限会社として、概ね従前の有限会社法が適用されまる会社ですのでご安心下さい。**また株式会社以外にも合同会社 合資会社 合名会社などがあります。
LexRank	これにより、有限会社は自動的に株式会社に組織変更されたものとして扱われています。**現在有限会社として残っているならば「特例有限会社」という形です。今ある、有限会社は、有限会社のままでもいれますし、株式会社になる事もできます。**それ以前に設立された有限会社は、会社名にきちんと「有限会社」と書かねばなりません。**特例有限会社は、そのまま存続できますし、一部例外を除き変更登記する必要もございません。「特例有限会社」という名前の株式会社です。会社法が施行されてからは、新たに有限会社を設立することはできなくなりました。**
USE+word (Proposal)	**会社法が施行されてからは、新たに有限会社を設立することはできなくなりました。従前の有限会社は特例有限会社として、概ね従前の有限会社法が適用されまる会社ですのでご安心下さい。**また株式会社以外にも合同会社 合資会社 合名会社などがあります。
	施行日現在の有限会社は、株式会社として存続はしますが、当該会社を『特例有限会社』と呼びます。特例有限会社は、そのまま存続できますし、一部例外を除き変更登記する必要もございません。特例有限会社から株式会社へ移行すること（商号変更）も出来ます。

謝辞

本研究では、国立情報学研究所の IDR データセット提供サービスによりヤフー株式会社から提供を受けた「Yahoo! 知恵袋データ（第 2 版）」を利用した。

参考文献

[1] 土居誉生, 石垣泰地, 小澤仁護, and 稲田徹. ″質問生成ソリューション 「なんでも Question」 の紹介.″ SIG-SLUD 5, no. 02 (2019): 123-124.

[2] 長谷川友治, et al. コールセンターにおける大規模質問応答データに基づく FAQ 作成支援システムの実装. *第 66 回全国大会講演論文集*, 2004, 2004. 1: 73-74.

[3] 友松祐太; 戸田隆道; 杉山雅和. AI チャットボットのためのチューニング支援システム. In: 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 90 回. 一般社団法人 人工知能学会, 2020. p. 08.

[4] 壹岐太一, et al. ヘルプデスクの対応記録からの QA リストの半自動抽出. *研究報告知能システム (ICS)*, 2018, 2018. 12: 1-8.

[5] PERONE, Christian S.; SILVEIRA, Roberto; PAULA, Thomas S. Evaluation of sentence embeddings in downstream and linguistic probing tasks. arXiv preprint arXiv:1806.06259, 2018.

[6] CER, Daniel, et al. Universal sentence encoder. *arXiv preprint arXiv:1803. 11175*, 2018.

[7] ERKAN, Günes; RADEV, Dragomir R. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 2004, 22: 457-479.

[8] ヤフー株式会社 (2011): Yahoo! 知恵袋データ（第 2 版）. 国立情報学研究所情報学研究データリポジトリ. (データセット). https://doi. org/10. 32130/idr. 1. 2