FAQ 検索における言い換え生成を利用したデータ拡張手法

曹 羽隆 [1]　小川 泰弘 [1,2]　外山 勝彦 [1]
[1] 名古屋大学大学院情報学研究科　[2] 名古屋大学情報基盤センター
cao.yulong.d0@s.mail.nagoya-u.ac.jp
{yasuhiro,toyama}@is.nagoya-u.ac.jp

概要

FAQ 検索タスクは，ユーザクエリに対して，データセットの中からそのクエリに関連する QA ペア（質問文と回答文のペア）を出力するタスクである．クエリと QA ペアの関連性を捉えるために，BERT [1] などの自然言語モデルを適用する手法が提案されてきたが，ファインチューニング用の学習データが足りないという問題点がある．本研究は，QA ペアの質問文の言い換えを生成することにより，学習データを拡張する手法を提案する．また，LocalgovFAQ データセット [2] を用いて学習データを拡張し，評価実験を行った．

1　はじめに

Web の FAQ ページは，ユーザの疑問が解決するように，よくある質問（FAQ）を参照するためのページである．しかし，ページに記載された FAQ の件数が少ない場合，ユーザが知りたい情報が掲載されていない可能性がある．逆に，FAQ の件数が多い場合，知りたい情報を発見するまでの時間が長くなるという問題がある．よって，ユーザが回答をより便利に探せるようにするため，検索機能が必要になる．FAQ 検索は，自然言語によるクエリに応答するための効率的な方法を提供する．すなわち，ユーザクエリが与えられた際に，データセットからそのクエリに関連する QA ペアを抽出する．FAQ 検索は近年，企業のチャットボットや，製品についての質問応答，顧客向けの技術サポートなどでよく用いられる．

しかし，特定のドメインに対応するためのデータセットは，そのデータ量が十分でない．例えば，LocalgovFAQ データセット [2] には，尼崎市役所の Web サイトに掲載されている QA ペアと尼崎市役所へのクエリが含まれ，各クエリには，どの QA ペアに関連するかについてのラベルがついている．データセットに含まれている QA ペアは 1,786 対，クエリは 784 個だけである．

本研究では，LocalgovFAQ データセットに対して，言い換え生成を利用するデータ拡張手法を提案し，その有効性を検証する．

2　先行研究

Sakata ら [2] は，教師なし情報検索システム TSUBAKI を利用して，クエリと質問文の類似性を計算した．一方，クエリと回答文の関連性は，BERT モデルを使用して計算した．また，TSUBAKI と BERT を組み合わせたシステムを構築した．しかし，LocalgovFAQ データセットのデータ量はファインチューニングの実施のためには十分でないため，尼崎市以外の地方自治体の Web サイトに掲載されている QA ペア 2 万件を収集して使用した．

Mass ら [3] は，三つのリランカーによる教師なし手法を提案した．ユーザクエリ–質問文とユーザクエリ–回答文の間で BERT のファインチューニングを行い，GPT-2 [4] モデルを用いて質問文の言い換えを生成することにより，既存の教師あり手法と同等程度の性能があることを示した．

Sourav ら [5] は，TF-IDF スコアを利用して質問文の代表的なキーワードを取得し，それに基づいてユーザクエリと QA ペアの類似度を計算する手法を提案した．さらに，ニューラルネットワークを追加した変種を提案し，高性能の FAQ 検索を実現した．

堂坂ら [6] は，Sentence-BERT [7] モデルを使用し，BERT モデルをファインチューニングすることにより，類似文検索のための良質な文ベクトルを生成した．その結果，類似文検索タスクで高い性能を示した．

3　TSUBAKI+BERT モデル

本研究は Sakata らが提案した TSUBAKI+BERT モデル [2] を利用するため，その詳細について以下に

述べる．

3.1 TSUBAKI による q-Q 類似性

TSUBAKI [8] を利用してユーザクエリ q と QA ペアの質問文 Q の類似度を計算する．TSUBAKI は，OKAPI BM25 [9] をベースにした教師なし検索エンジンであり，単語だけでなく文の依存構造も考慮して，正確な検索を提供する．柔軟なマッチングを実現するために，辞書や Web コーパスから自動的に抽出された同義語も使用する．ここでは，各 QA ペアの質問文 Q を文書と見なし，クエリ q と質問文 Q の間の類似度を計算する．

3.2 BERT による q-A 関連性

BERT を利用してユーザクエリ q と回答文 A の関連度を計算する．学習データとしてクエリ q の代わりに質問文 Q を使用する．

データセットの各 QA ペアを $(Q_1, A_1), (Q_2, A_2), ..., (Q_M, A_M)$ とする．学習は全 QA ペア (Q_j, A_j) を正例とし、負例は Q_j に対して A_j 以外の回答文からランダムに選択し，正例と負例を学習データとして BERT により二値分類する．具体的には，Q と A の関連度を score(Q, A) とすると，正例 (Q_j, A_j) に対しては score(Q_j, A_j) が 1 になるように，負例 $(Q_j, A_{j'})$ $(j' \neq j)$ に対しては score$(Q_j, A_{j'})$ が 0 になるように学習する．

検索時は，ユーザクエリ q とすべての QA ペアの回答文 A_j の関連度 score(q, A_j) $(j = 1, ..., M)$ を計算し，その上位を検索結果とする．

3.3 TSUBAKI と BERT の組み合わせ

柔軟なマッチングを実現するために，TSUBAKI による q-Q 類似性と BERT による q-A 関連性を組み合わせている．

TSUBAKI のスコアが高い場合は，クエリと質問文の間で重複している語が多く，正解である可能性が特に高い．そこで，そのような QA ペアを優先してランキングし，他の QA ペアはスコアを単純に加算して統合する．

具体的には，TSUBAKI による検索結果上位 10 件の中にスコアが閾値 α 以上のものがあり，それが BERT によるスコアの上位 10 件にも含まれている場合，優先的に 1 位から順にランキングし，残りの候補は TSUBAKI と BERT のスコアを単純に加算し，ランキングする．スコア上位を検索結果とする．

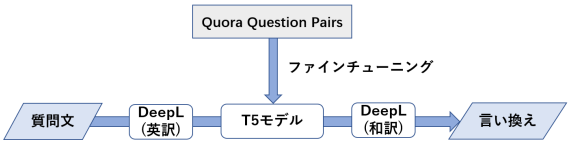

図 1 言い換え生成（手法 1）

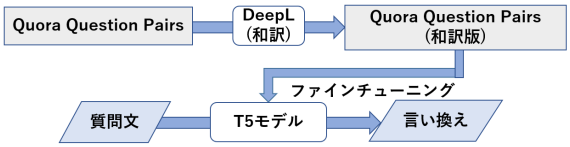

図 2 言い換え生成（手法 2）

4 提案手法

本研究では，BERT のファインチューニング用の学習データ不足の問題に注目している．Sakata ら [2] は地方自治体 21 市の Web サイトに掲載されている QA ペア計 2 万件を収集して，ファインチューニング用のデータに追加した．しかし，地方自治体から収集したデータ数は有限であり，また専門家によるデータ作成は手間がかかる．したがって，本研究では言語モデルを使用し，QA ペアの質問文に対して, その言い換えを生成する手法を提案する．

4.1 手法 1

日本語言い換えデータセットは利用できるものが少ないため，手法 1 を提案する．この手法では，質問文をまず英訳し，英語事前学習モデル T5 で言い換えを作成し [10]，最後にそれを和訳する. 生成の流れを図 1 に示す．

ここで，英語事前学習済み T5 モデルのファインチューニングは英語データセット Quora Question Pairs [11] を用いた．Quora Question Pairs には質問応答 Web サイト Quora から収集したユーザ質問約 40 万行が含まれる．各行は，質問文 2 文から構成され，この 2 文は言い換えであるか否かのラベルがついている．

4.2 手法 2

手法 1 では言い換え生成の過程で機械翻訳 2 回を行うため，重複した言い換えが作成されることがある．そこで，手法 2 を提案する．この手法では, 英語データセット Quora Question Pairs をまず和訳し, 和訳データを用いて日本語事前学習済み T5 のファインチューニングを行う．このファインチューニング済みの T5 で言い換えを作成する. 生成の流れを図 2 に示す．

表 1 BERT のハイパーパラメータ

手法	バッチサイズ	学習率	エポック数
ベースライン	32	2e-5	3
手法 1	32	1e-5	10
手法 2	32	1e-5	25

4.3 言い換えによる学習データ拡張

手法 1 と手法 2 で得られた言い換えを別々に用いて，BERT のファインチューニング用の学習データを拡張する．

具体的には，ある質問文 Q_j に対して，N 個の言い換え P_{ij} $(i = 1, ..., N)$ が得られたとする．そのとき，データ拡張前の正例 (Q_j, A_j) と負例 $(Q_j, A_{j'})$ に対して，正例 (P_{ij}, A_j) と負例 $(P_{ij}, A_{j'})$ を作成する．これらを BERT のファインチューニング用の学習データとして追加する．また，正例に対しては score(P_{ij}, A_j) が 1 になるように，負例に対しては score$(P_{ij}, A_{j'})$ が 0 になるように学習する．

5 実験

手法 1 と手法 2 の有効性を検証するために，評価実験を行った．

5.1 実験設定

LocalgovFAQ データセットを用いて実験を行った．ベースラインとして，言い換えによるデータ拡張前のデータを使用するものとする．言い換え生成では，各質問文に対して，10 文の言い換えを生成した．なお，BERT モデルと TSUBAKI+BERT モデルの評価指標は P@5，MAP，MRR を用いた．

BERT の 事前学習には，日本語 Wikipedia 事前学習済み [12] を使用した．ファインチューニングには，元々の QA ペア 1,786 対に言い換えで作成したデータを追加して用いた．入力テキストは形態素解析器 Juman++ [13] を用いて，形態素に分割した．3.3 節で説明した TSUBAKI 優先統合の閾値 α は 0.3 とした．また，BERT モデルのハイパーパラメータの設定を表 1 に示す．

5.2 実験結果

言い換え生成では, 質問文 1,786 文の各文に対して，10 文の言い換えを生成し，重複文を削除した結果，手法 1 では計 11,495 文が得られ，手法 2 では計 16,789 文が得られた．

言い換えにより拡張したデータを用いた実験結果を表 2 に示す．すべての手法において，TSUBAKI+BERT の結果は BERT より良い．これにより，TSUBAKI と BERT を組み合わせる手法の有効性を確認した．また，手法 1 と手法 2 はベースラインを上回り，言い換えを用いて学習データを拡張する手法の有効性を確認した．手法 1 と手法 2 の結果を比べると，手法 2 はわずかに上回っていることがわかる．手法 1 では，重複文が多く生成され，得られた拡張データは手法 2 より少ないことが原因であると考えられる．また，手法 1 と手法 2 はともに先行研究を上回ることができなかった．

表 2 言い換えにより拡張したデータを用いた実験結果

手法	システム	P@5	MAP	MRR
ベースライン	BERT	0.237	0.456	0.492
手法 1	BERT	0.262	0.491	0.532
	TSUBAKI+BERT	0.275	0.548	0.599
手法 2	BERT	0.277	0.508	0.549
	TSUBAKI+BERT	0.283	0.575	0.634
先行研究	BERT	0.333	0.576	0.631
	TSUBAKI+BERT	**0.357**	**0.647**	**0.705**

5.3 考察

手法 1 と手法 2 は先行研究を上回ることができなかった．その原因としては，先行研究は他の地方自治体の QA ペア 2 万件を用いており，その文数は今回の手法 1 および手法 2 で生成した言い換えより多いことが考えられる．また，本研究の言い換え生成は QA ペアの質問文を対象にしていて，回答文の言い換えを考慮しなかったため，回答文の多様性が先行研究が使用した QA ペアよりも劣っていることが考えられる．

手法 1 と手法 2 で得られた言い換えについて考察する．手法 1 では，和訳前の英文の言い換えは異なりで 17,824 文あるが，和訳後の言い換えは異なりで 11,495 文であった．手法 1 が重複文を生成した原因としては，言い換え生成の流れの中で機械翻訳を使用したためと考えられる．

また，手法 1 と手法 2 で得られた言い換えの分布を図 3 と図 4 にそれぞれ示す．各図の横軸は各質問文に対する言い換えの数を表し，縦軸は言い換えの元になった質問文の数を表す．手法 1 では質問文 1 文あたりに 6 文 ~8 文程度の言い換えが得られ，手法 2 では質問文 1 文あたりに 9 文 ~10 文程度の言い換えが得られた．

この分布から，手法 1 では質問文 1 文あたり 10 文

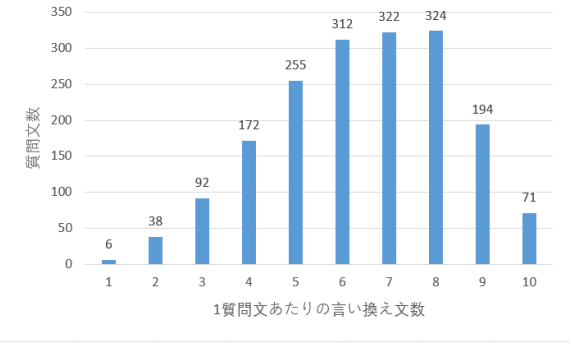

図 3　言い換えの分布（手法 1）

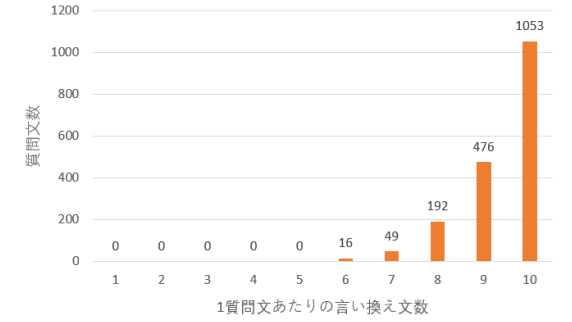

図 4　言い換えの分布（手法 2）

以上の言い換えを生成しても，重複文があるため，実際に得られた言い換えは 6 文 ~8 文程度になると考えられる．よって，手法 1 では 10 文以上の言い換えを得ることが実際には難しいと考えられる．また，手法 2 は，手法 1 に対して重複文が比較的少ないため，手法 1 より多くの言い換えを得られた．

また，手法 2 において，追加する言い換えの文数が与える影響を調べた．手法 2 では，16,789 文の言い換えが得られ，それらすべてを追加して学習データを拡張した．ここでは，より少ない文数を追加した場合と比較することにより，その有効性を検証する．具体的には，手法 2 で得られた言い換えから，質問文 1 文あたり 5 文の言い換えをランダムに抽出して追加した場合（5 文拡張）と，全部の言い換えを追加した場合（10 文拡張）を比較した．結果を表 3 に示す．ここで，質問文 1 文あたり 5 文としたのは，図 4 に示すように，質問文 1 文あたり少なくとも 6 文の言い換えが得られたからである．表 3 により，言い換えを 5 文だけ追加した場合と全部を追加した場合の間で結果に差がほとんどなく，単純に言い換え文数を増加しても，結果の改善には貢献しないことが分かった．

表 3　異なる言い換え文数で拡張したデータを用いた結果

手法	システム	P@5	MAP	MRR
手法 2（5 文拡張）	BERT	0.262	0.517	0.560
	TSUBAKI+BERT	0.282	**0.582**	0.628
手法 2（10 文拡張）	BERT	0.277	0.508	0.549
	TSUBAKI+BERT	**0.283**	0.575	**0.634**

6　まとめと今後の課題

本研究では FAQ 検索における言い換え生成を利用したデータ拡張手法を提案した．手法 1 では言い換え生成において機械翻訳 2 回を使用したので，重複文が多く得られた．手法 2 では英語データセット Quora Question Pairs を和訳することで，より多くの言い換えを得られ，手法 1 をわずかに上回った．今後の課題としては，質問文の言い換えだけでなく，回答文の言い換えも考慮することにより，回答文の多様性を拡張することが考えられる．

謝辞

本研究は JSPS 科研費 22H03901 の助成を受けたものである.

参考文献

[1] 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.

[2] 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] Yosi Mass, Boaz Carmeli, Haggai Roitman, and David Konopnicki. Unsupervised FAQ retrieval with question generation and bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 807–812, 2020.

[4] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[5] Sourav Dutta, Haytham Assem, and Edward Burgin. Sequence-to-sequence learning on keywords for efficient FAQ retrieval. *arXiv preprint arXiv:2108.10019*, 2021.

[6] 堂坂浩二, 金子和樹, 木村幸司, 伊東嗣功, 石井雅樹. 生活保護業務支援のための質問応答システムの開発と評価. 第 21 回情報科学技術フォーラム講演論文集 (FIT2022), Vol. 21, No. 2, pp. 235–238, 2022.

[7] 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. Association for Computational Linguistics, November 2019.

[8] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. Tsubaki: An open search engine infrastructure for developing information access methodology. *Journal of information processing*, Vol. 20, No. 1, pp. 216–227, 2012.

[9] 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.

[10] Hemant Palivela. Optimization of paraphrase generation and identification using language models in natural language processing. *International Journal of Information Management Data Insights*, Vol. 1, No. 2, p. 100025, 2021.

[11] Iyer Shankar, Dandekar Nikhil, and Csernai Kornel. First quora dataset release: question pairs (2017). *URL https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs*, 2017.

[12] 柴田知秀, 河原大輔, 黒橋禎夫. BERT による日本語構文解析の精度向上. 言語処理学会第 25 回年次大会発表論文集, pp. 205–208, 2019.

[13] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 54–59, 2018.