専門性の高いオープンブック質疑応答システムの構築と専門家添削による誤答抑制

後藤成晶[1] 上山道明[1] 須藤栄一[1] 清水司[1] 木村英彦[1]
[1]株式会社豊田中央研究所
{sg-goto, kamiyama, eiichi-s, shimizu, hdkimura}@mosk.tytlabs.co.jp

概要

専門家の知識の一部は、報告書などのテキストデータとして所属組織内に蓄積される。材料分析ドメインを対象に、テキストを参照して専門性の高い質問に自動回答するシステムを、オープンブック質疑応答を応用して構築した。専門性の高い質疑応答は回答抽出の難度が高いため、誤答の多発が懸念される。そこで、回答の正誤検証器を備えるシステムを構築し、その訓練手順を提案した。提案の訓練手順は、質疑応答ログを専門家が正誤添削するだけで良く、一般的なデータセット作成よりも作業負荷が少ない。実験から、訓練によって誤答の提示頻度が50%以上抑制されたことを確認した。

1 はじめに

産業分野などにおけるドメイン専門家の知識の一部は、報告書などのテキストデータとして所属組織内に蓄積される。その専門知識を再利用できるよう、膨大なテキストデータから情報抽出できる仕組みが求められる。その手段として、テキストデータを参照する質疑応答システムの研究が進められている。なお、そのようなシステムは、教科書を開きながら試験問題に答える「持ち込み試験」に例えてオープンブック質疑応答システムと呼ばれる。

本検討では、オープンブック質疑応答を「材料分析」ドメインに応用し、図１に示すようなシステムを構築している。ユーザが質問を入力すると、バックエンドでオープンブック質疑応答アルゴリズムがテキストデータを参照し、回答を抽出し提示する。また、参照したテキストのリストを参考文献として提示する。

材料分析のような専門性の高い分野への質疑応答システムの応用は、主に次の２つの難しさがある。

教師データの作成コストが高い　専門性の高い教師データ作成作業は、作業者に専門知識が必要で、クラウドソーシングを活用しにくい。そのためデータ作成コストが高くなる。

誤答の発生頻度が高い　非専門的な質問の例として「或る人物の出生地」を問う場合、出生地は１つしかないため誤答が起きにくい。一方で「材料の分析方法」を問う場合、分析方法は１つではなく、目的や条件によって適切な分析方法が異なるため、誤答が発生しやすい。

以上を踏まえ、本報告の内容は以下の通りである。

- オープンブック質疑応答を専門性の高い材料分析ドメインへ応用した。
- 一般的な教師データ作成より負荷の少ない「添削」による訓練手順を提案した。
- 訓練により誤答を抑制可能なことを実験から確認した。

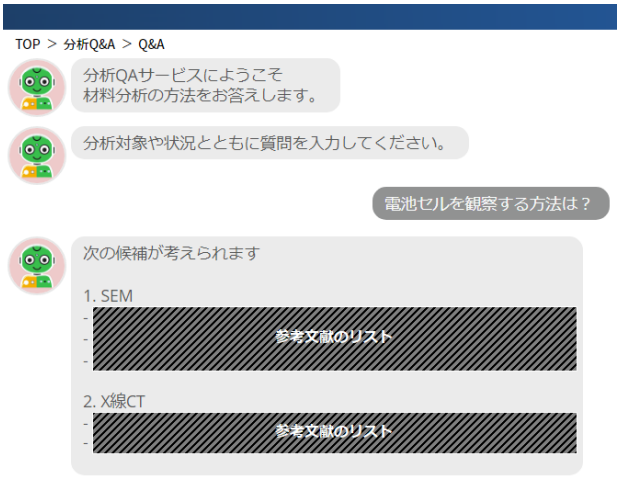

図 1　構築した材料分析の質疑応答システム

2 質疑応答システムの構成

本検討で構築した質疑応答システムの構成を図 2 に示す。文章検索(Retriever)と回答生成(Reader)を組み合わせる方式で、さらに正誤検証(Verifier)を設け

ることで誤答抑制を図る。以下に各要素の実装詳細を述べる。

文章検索(Retriever)　ベクトル近傍探索で実装した。文章のベクトル化には BERT[1]を用いた。訓練データとして質疑応答データセットを QNLI (Question Natural Language Inference)形式に変換したものを利用した。訓練コードとして Sentence Transformers ライブラリ[2]を利用した。

回答生成(Reader)　複数文章を一括入力して処理できる Fusion-in-decoder [3]のアルゴリズムを採用し、実装コードは内製した。モデルアーキテクチャは BART (Bidirectional Auto-Regressive Transformers) [4]を採用した。推論時にはジェネレータの生成数を１０などに設定することで､複数回答を出力させた。

根拠検証(Verifier)　正答/誤答の２クラス分類モデルとして BERT を用いた。「質問、回答、コンテキスト文章」を入力し、「正答/誤答」のラベルを推論する。誤答を除外したものが最終的な回答セットとしてユーザに表示される。

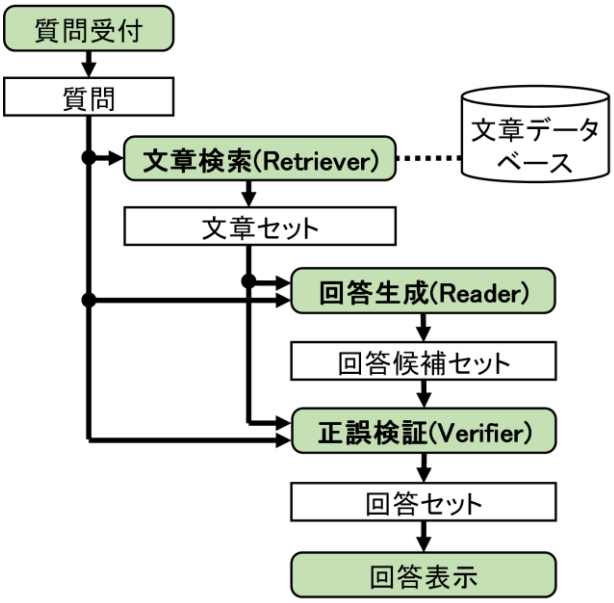

図 2　質疑応答システムの構成概要

3 訓練手順

3.1 全体の手順

本検討では、図３に示す手順で訓練を実施した。各作業の概要を以下に述べる。

事前学習　BERT および BART の事前学習を実施する作業である。なお本検討では「材料分析に関する質疑応答チャット」という目的に合致するよう、産業技術文章とソーシャルネットワーキングサービス文章を合計 30 GB 収集して利用した。

ファインチューニング-1　一般公開されている大量な質疑応答データセットによるファインチューニング作業である。なお本検討では SQuAD [5]および Natural Questions [6]を機械翻訳することで約十万件の日本語質疑応答データセットを作成して利用した。

ファインチューニング-2　対象ドメインの専門家が作成した質疑応答データセットによるファインチューニング作業である。

専門家添削　ファインチューニング-２で訓練された質疑応答システムによる質疑応答ログを、専門家が添削する作業である。

ファインチューニング-3　専門家が作成した質疑応答データと、専門家添削結果とを組み合わせたものを教師データに、ファインチューニングを実施する作業である。

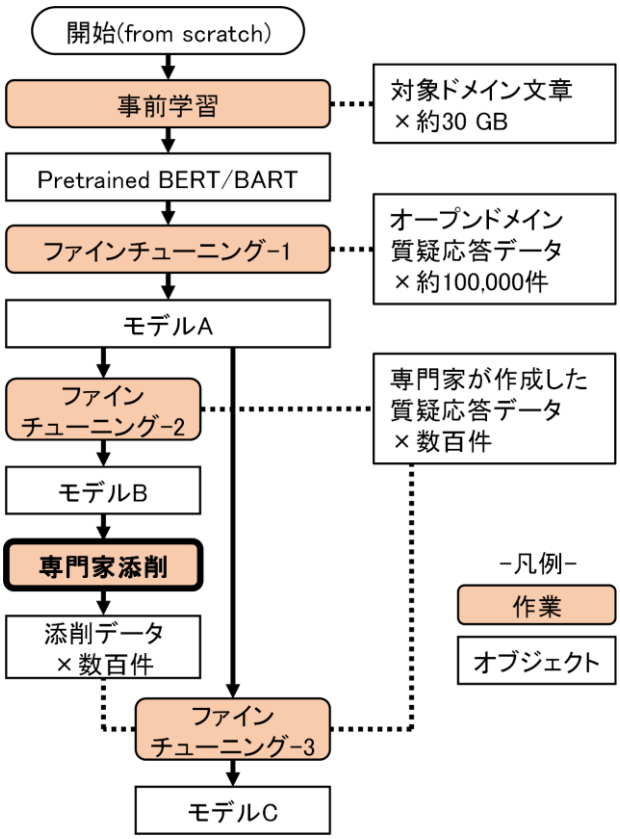

図 3　質疑応答システムの訓練手順

3.2 専門家による質疑応答データの作成

データ形式は SQuAD と同様で、「質問」「回答」

「コンテキスト」からなる。材料分析に関する社内報告書から抜粋した文章をコンテキストに利用し、そのコンテキストから回答可能な質問およびその回答を材料分析専門家が作成した。

質疑応答データセットの作成にあたり、データセットの偏りが無いよう注意した。例えば自由フォーマットの議事録をコンテキスト文章に利用することで、言い回しの偏りを予防した。また、材料分析ドメインを複数の小分野に分割し、それぞれの小分野に担当専門家を設定することで、分野の偏りも予防した。

3.3 専門家添削

質疑応答システムの回答ログを専門家が確認し、正答または誤答のラベルを付与する作業である。データ形式として、「正誤ラベル」「質問ログ」「回答ログ」に加えて「Reader が参照していたコンテキスト」も用いることで、質疑応答データセットと同様に扱えるようにした。

図４に、質疑応答データ作成と添削の作業フローを比較して示す。質疑応答データを作成するためには、１０００文字程度のコンテキスト文章を読解したうえで､そのコンテキストに沿った質問を作文し、回答フレーズを抜き出す作業が必要である。一方で添削作業は、コンテキスト読解作業が無く、さらに質問の「作文」作業が「読む」作業に負荷軽減されているため、より少ない負荷で作業可能である。

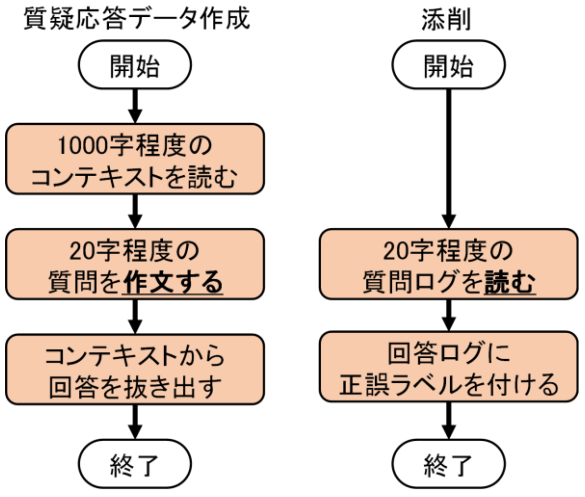

図 4　質疑応答データ作成と添削の作業比較

この添削作業は、実際の運用時には、専門家がログを確認して手動で実施することを想定している。ただし今回の実験では、簡易的に効果検証するために、質問文章とその正答リストからなる添削用質疑応答データを予め約２００件作成し、それを用いて自動添削を行った。具体的には、そのデータセットの質問文章をモデル B に入力し、提示された回答それぞれに対して、正答リストに含まれている場合は正答、そうでない場合は誤答であると、自動的にラベルを付与した。１質問あたり約３フレーズの回答を生成させ自動添削をかけることで、約６００件の添削データセットを準備した。

3.4 Verifier の訓練方法

Verifier の訓練データ概要を図５に示す。２クラス分類器である Verifier の訓練には正答(Positive)および誤答(Negative)のラベルが付与された両方のデータが必要であるが､SQuAD 等の質疑応答データセットはすべてが Positive データである。そのためファインチューニング-2 では、質疑応答データセットのうち､コンテキストをランダム(ただし回答フレーズが文中に出現するものに限定)に置き換えることで、Negative データを自動生成した。ファインチューニング-3 では、添削により Negative データが利用可能になっているため、自動生成された Negative データと添削された Negative データの両方を利用した。

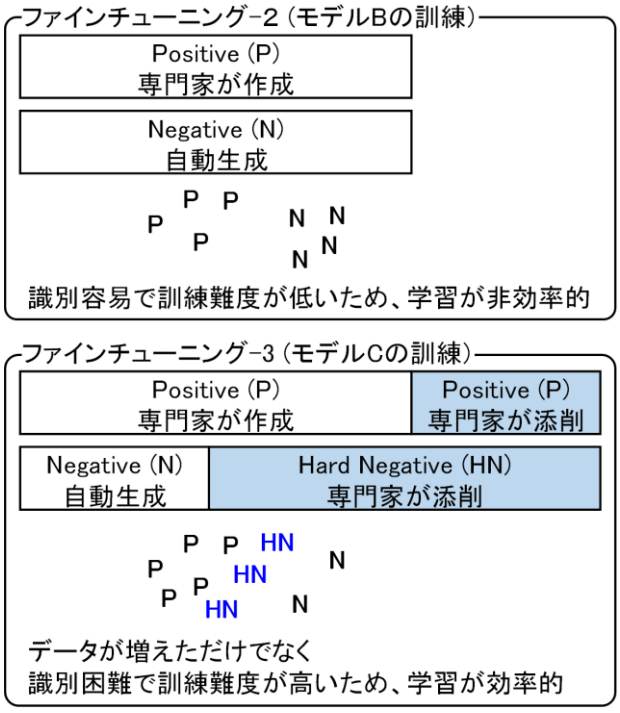

図 5　ファインチューニング-2 および 3 における Verifier 訓練データ

自動生成された Negative データは識別が容易で、Verifier の訓練効率が悪い。一方で添削データは、もともとモデル B が正答と誤認するほど識別困難な例(Hard Negative)であるため、Verifier の訓練効率が高い。よって添削による訓練は、作成工数が少ないにも関わらず、より高精度な Verifier を訓練できると期待される。

4 実験

材料分析に関する５０件の質疑応答テストデータを作成し、図３におけるモデル A~C の性能を評価した。参照する文章データベースには材料分析に関する報告書を約７０００パラグラフ利用した。なおテストデータは文章データベースの内容を知らされていない専門家が作文したため、回答不可能な質問も多く含まれる。

モデル A~C のテスト結果として、５０件のテストを合計した正答と誤答の総数を図６に示す。システムの提示する回答フレーズが、正答リストに登録されているフレーズのどれかと完全一致する場合は Exact Match すなわち正答、そうでなければ Not Exact Match すなわち誤答とする。なお、システムは１つの質問に複数の回答を提示するため、正答と誤答の合計数は、テストデータ件数より多い。左上にプロットされるほど優れていることを意味する。

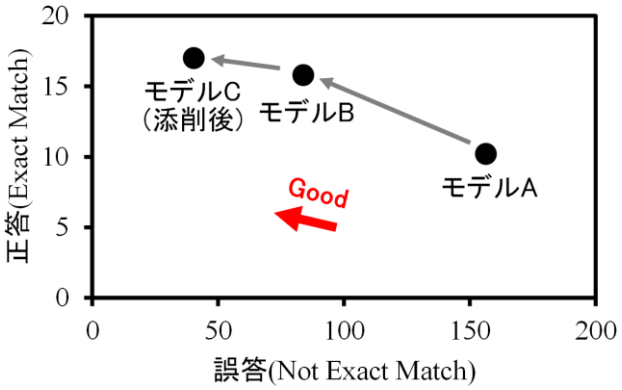

図 6　テストデータ全体を通した正答と誤答の表示総数

まず一般データセットだけで訓練したモデル A では、誤答数が１５０を超えた。次に、材料分析の質疑応答データセットで訓練したモデル B では、正答の増加と、誤答の減少が同時に見られた。最後に、添削データを加えて訓練したモデル C では、モデル B より更に誤答を５０％以上削減できている。以上の事から、添削を取り入れた訓練手順により、誤答を抑制できることを確認した。

なお、以下の理由から、図６の結果はモデル A~C を相対評価するための参考値であり、システム性能の絶対値を評価するものではないことを補足する。

- テストデータは、文章データベースの内容に依存せずに作成されたため、解答不可能な例も多く含まれる。文章データベースを充足することで、正答の増加が見込まれる。
- 事前に列挙していた正答セットに抜け漏れがある。そのため本来は正答であっても見過ごされ誤答に割り振られた回答もある。

5 関連研究

Shao ら[7]は、Verifier を持つオープンブック質疑応答システムを提案し、複数回答ベンチマークタスクにおいて State-Of-The-Art の性能を確認している。本検討のシステム構成と本質的に同じである。本検討は、添削から学習することによる誤答抑制効果を検証することが目的であり、Shao らのシステムでも本検討で検証した手順を適用できると考えられる。

鈴木ら[8]は、コンテキストを読解したうえで手動作成した Negative 訓練データが、質疑応答システムの性能を向上可能なことを実験から示している。一方で本検討は、コンテキストを読解せず簡易的に作成した添削データであっても、誤答抑制のための訓練データに有用なことを確認するものである。

Campos ら[9]は、質疑応答ログに正誤ラベルが付与された添削データを教師データに用いる事で、Wikipedia 等が由来の非専門的ドメインの質疑応答タスクにおいて、抽出型 Reader をファインチューニング可能なことを確認している。一方で本検討は、より難度の高い専門的ドメインにおいて、Verifier を含めた構成を添削データから訓練することで、誤答抑制に寄与できることを確認するものである。

6 おわりに

専門性の高い質疑応答システムの構築を行った。回答の正誤検証を行う Verifier を設け、専門家添削を取り入れた訓練手順を提案した。また実験から、数百件の添削データで訓練することにより、誤答の提示頻度を５０％以上抑制できることを確認した。今後は、構築したシステムを用いた質疑応答サービスの試行運用を行い、より多様な質疑応答ログを添削することで、回答品質をさらに向上可能か検証する予定である。

謝辞

本検討を進めるにあたり質疑応答サービスのアプリケーション作成に貢献頂いた株式会社豊田中央研究所の藤井亮暢氏、後藤邦博氏、奥村文洋氏、丹羽貴寛氏、加藤光樹氏に感謝申し上げる。

参考文献

1. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

2. Nils Reimers and Iryna Gurevych. 2019. 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), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

3. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

4. Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 874–880, Online. Association for Computational Linguistics.

5. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

6. Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. Transactions of the Association for Computational Linguistics, 7:452–466.

7. Zhihong Shao and Minlie Huang. 2022. Answering Open-Domain Multi-Answer Questions via a Recall-then-Verify Framework. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1825–1838, Dublin, Ireland. Association for Computational Linguistics.

8. 鈴木正敏, 松田耕史, 大内啓樹, 鈴木潤, 乾健太郎. 2021. オープンドメイン質問応答における解答可能性判別の役割. 言語処理学会 第 27 回年次大会 (NLP 2021)

9. Jon Ander Campos, Kyunghyun Cho, Arantxa Otegi, Aitor Soroa, Eneko Agirre, and Gorka Azkune. 2020. Improving Conversational Question Answering Systems after Deployment using Feedback-Weighted Learning. In Proceedings of the 28th International Conference on Computational Linguistics, pages 2561–2571, Barcelona, Spain (Online). International Committee on Computational Linguistics.