複数文書の読解を要する質問の自動生成と質問応答システムへの応用

小林俊介　河原大輔
早稲田大学理工学術院
{carlike787@toki., dkw@}waseda.jp

概要

オープンドメイン質問応答タスクは質問に対する文書の検索と読解で構成され、そのデータセットには、複数の文書を読解しなければ解答できない質問が多く存在する。本研究では、テキスト生成モデルをベースとした質問応答モデルである Fusion-in-Decoder [1] を用いて、複数文書の読解を要する質問を自動生成する。生成された質問を、オープンドメイン質問応答システムに応用した結果、正答率が約 2%向上し、質問生成の有効性が示された。

1　はじめに

自然言語処理における機械学習の利用においては、学習データの量が重要であり、多量かつ高品質なデータセットの取得が課題の 1 つになっている。データセットを作成する手法としてアノテーションが使われる。しかし、アノテーションを人手で行うとコストが高くなってしまうため、データ作成・拡張を自動で行う手法が用いられることもある。

本研究ではデータ拡張をオープンドメイン質問応答タスクに応用する。オープンドメイン質問応答タスクは質問に対する文書の検索と読解で構成され、そのデータセットには、複数の文書を読解しなければ解答できない質問が多く存在する。本研究では、テキスト生成モデルをベースとした質問応答モデルである Fusion-in-Decoder [1][1)](FiD) を用いて質問を自動生成する手法を提案する。提案手法の FiD は解答と複数の文書を入力とし、与えられた解答に対応する質問を生成する。FiD は元来、複数の文書を考慮した出力の生成が可能であり、生成される質問は複数の文書を参照しなければ解答できないものになることが期待される。

一方で、生成された質問は矛盾や不自然な表現を含む場合や、異なる解答を持つ場合があり、そのまま追加データとするには不適切な場合がある。このため、元の質問と表現が異なり、かつ解答が同一である質問を選別するため、生成された質問と元の質問との類似度を用いたフィルタリングを行う。

選別された質問をデータセットに追加して質問応答システムの学習を行った結果、データ拡張を行わずに学習を行った場合と比べて、正答率が最大で 2%改善し、自動生成された質問が学習に貢献することが示された。また、1 つの文書のみで生成した質問を加えた場合、複数文書で生成した質問で学習した場合よりも正答率が悪化し、複数文書で生成を行う提案手法が有効であることが示された。また、フィルタリングを行うと、正答率が最大で 0.6%向上し、フィルタリングも効果があることが示された。

2　関連研究

英語における質問応答タスク用のデータセットでは、関連文書を与えて質問に答える形式のものが多く、SQuAD [2]、TriviaQA [3]、Natural Questions [4] などがある。日本語における質問応答タスク用のデータセットには、SQuAD の形式を踏襲した運転ドメイン QA データセット [5] や、JGLUE [6] に含まれている JSQuAD、クイズ形式の質問文で、日本語 Wikipedia を関連文書とする JAQKET [7] などがある。SQuAD と TriviaQA では訓練用の質問が 100,000 件前後、Natural Questions では 300,000 件以上用意されている。一方で運転ドメイン QA データセットは 34,000 件強、JSQuAD は訓練用データで 64,000 件弱、JAQKET では評価用を合わせても 24,000 件前後と、英語データセットに対して一桁少ない。

自然言語処理の各タスクで大きな精度向上を達成したモデルが、Transformer に基づく汎用言語モデルである。その中でもテキストを生成できるモデルとして T5 [8] がある。

1）https://github.com/facebookresearch/FiD

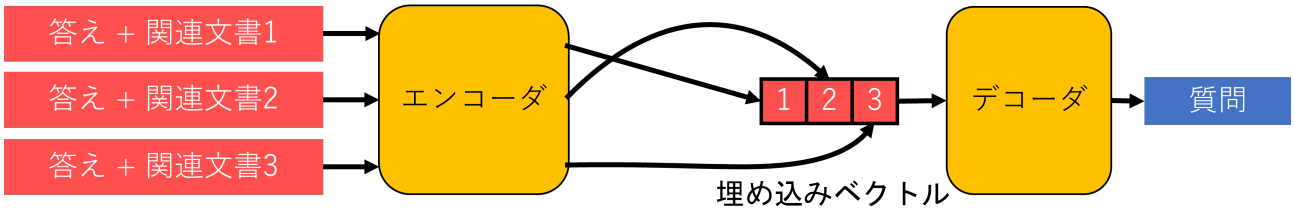

図 1　*N*=3 における質問生成の流れ

Izacard ら [1] は T5 をベースとした質問応答のモデルとして、複数の文書を入力に用いることができる FiD を提案した。このモデルでは入力される文書ごとに、質問文と結合してエンコードが行われ、得られた埋め込みベクトルは一括してデコーダに入力される。これにより、解答生成の際に複数の文書から情報を得られるという特徴があり、より精度の高い解答を生成できる。

質問の生成については、Du ら [9] が DNN による質問生成を行い、精度を改善させたことから、特にテキスト生成が可能な言語モデルによる質問生成が行われてきた。Murakhovs'ka ら [10] は、1 つの比較的長い文書を入力とし、異なる解答を持つ複数の質問を生成できる MixQG を提案している。既存モデルから 10%以上の精度向上を達成しているが、入力が 1 文書であり、複数の文書を参照した質問にはなっていない。日本語においても、折原ら [11] により、日本語 T5 モデルを用いて、ニュース記事からクイズを生成する試みが行われた。既存のクイズサービスに掲載されている質問を例に、生成されたクイズが単に正答を問うだけではなく、面白みのあるクイズになっているかについて考察している。しかし、ここでも入力する記事は 1 つだけであり、また使用された訓練データ数も 220 件と少ない。

3　複数文書による質問生成

3.1　質問の生成とフィルタリング

本研究の目標は、複数の文書を読解しないと解答することができない質問の生成である。この要素を満たす FiD を本研究での質問生成で用いるモデルとする。提案するモデルは FiD と同一の構造であり、入出力の形式が異なるのみである。通常、FiD の入力は、質問文と、読解すべき文書のタイトルおよび内容を入力するが、本研究では、質問文の代わりに解答と、解答を含む関連文書 *N* 個を入力する。*N*=3 の場合の質問生成の流れを図 1 に示す。

ただし生成された質問は、そのままでは矛盾や不自然な表現を含む場合や、異なる解答を持つ場合があり、そのまま追加データとするには不適切な場合がある。そこで、元の質問と表現が異なり、かつ解答が同一である質問を選別するために、フィルタリングを行う。フィルタリングは元の質問と生成された質問の類似性を測定することで実現し、その指標として BERTScore [12] と BLEU [13] を用いる。いずれも 2 つの文の類似度を測定する指標であるが、BLEU が N-gram 単位の表層的な類似度を測定するのに対し、BERTScore は BERT の出力する埋め込みベクトルに基づく意味的な類似度を測定する。よって、同じ解答で異なる表現の質問を得るには、BERTScore が高く、かつ BLEU のスコアが低い質問が適している。加えて、元の質問は事実に基づいたことが書かれているため、生成される質問に矛盾等が生じていれば、BERTScore の値は低くなると考えられる。以下の例 (1) はそのような質問のペアである。解答は「夜明け前」であり、元の質問 a のように書き出しは「木曽路はすべて山の中である」であるが、生成された質問 b は異なる書き出しになっており、矛盾が生じている。

(1)　a.「木曽路はすべて山の中である」という一文で始まる、島崎藤村の小説は何でしょう?
　　b.「だから、私はその家を飛び出した」という書き出しで始まる、島崎藤村の小説は何でしょう?

3.2　実験

3.2.1　実験設定

2020 年から質問応答タスクのコンペティション「AI 王」のデータセットとして JAQKET データセットが利用されている。本研究では、同コンペティションの第 2 回大会[2)]で提供されている、JAQKET をベースとしたクイズ形式のデータセット、および

2)　https://sites.google.com/view/project-aio/competition2

表 1 無作為抽出による質問 50 問の人手評価結果

	3 文書	1 文書
自然さ	18	13
複数文書の読解	7	0

日本語 Wikipedia 記事の文書集合[3]を用いた実験を行う。各質問には前処理として、Elasticsearch[4]を用いた、各質問の解答を含む関連文書[5]の抽出が行われている。このデータをもとに、FiD の学習で必要とされる質問、解答、関連文書の組を抽出した。

FiD で用いる日本語 T5 モデルは、Hugging Face Transformers ライブラリに存在するものを用いた[6]。学習で設定したハイパーパラメータを付録 A に示す。質問生成で使用する文書数は、関連文書のうち 3 つ、および、1 つのみとする 2 パターンとした。生成においては各入力に対し、beam search により生起確率の高い出力 7 つを取得した。

モデルの評価では、評価用データセットを用いて、各入力に対して質問の生成を行い、人手による評価を行った。自動生成そのものの結果を評価するため、フィルタリング前に無作為に抽出した 50 問を評価対象とし、評価の観点は矛盾や文法ミスがないなど、質問に自然さがあるかどうかと、複数文書の読解を要するか否かの 2 つとした。

3.2.2 実験結果と議論

質問の生成 以上の設定の下で、質問生成に対応した FiD のファインチューニングを行った。評価結果を表 1 に示す。以下に、3 つの関連文書を用いて生成された質問の例を示す。

(2) 1971 年に独立するまでは「東パキスタン」と呼ばれていた、首都をダッカに置く国はどこでしょう?

(3) 『斜陽』『人間失格』『堕落論』などの作品を残した、昭和を代表する作家は誰でしょう?

(4) 花見大根、桜島大根といえばどんな野菜の品種でしょう?

例 (2) は、質問文に矛盾や文法的誤りがなく、適切な生成が行われた例である。一方、例 (3) は「太宰治」を解答とする質問であるが、「堕落論」は坂口安吾の作品であり、事実に矛盾する例となっている。また、例 (4) は矛盾こそないものの、質問文中で解答の「大根」に言及してしまっているため、質問として全く意味を成していない。FiD がベースとする T5 の出力生成では、出力文全体の意味を考慮したデコードが難しいため、このような不適切な質問が生成されると考えられる。

次に、複数文書の読解を要する質問について、3 つの関連文書を用いて生成した例を (5) に示す。(6) は参照の必要がある文書の抜粋である。

(5) 「眠り猫」の銅像が有名な、徳川家康を祀った神社は何でしょう?

(6) a. 眠り猫 (ねむりねこ) は、栃木県日光市の日光東照宮の回廊にある建築装飾彫刻作品。
b. 日光東照宮 (にっこうとうしょうぐう) は、日本の関東地方北部、栃木県日光市に所在する神社。江戸幕府初代将軍・徳川家康を神格化した東照大権現 (とうしょうだいごんげん) を主祭神として祀る。

複数の文書の読解を要する質問は多く生成できなかったが、この原因として、訓練で利用した質問の多くが単一の文書の読解のみで解答できてしまうことが考えられる。オリジナルの質問を 50 問サンプリングして検証した結果、22 問が単一文書の読解で解答可能であり、これらのデータによる学習が複数の文書を参照することを妨げていると推測される。

また、関連文書 1 つのみを用いて生成した質問には、複数文書を読解しないと解答できない質問は存在しなかった。

質問のフィルタリング 次に、フィルタリングの効果を検証するため、BERTScore と BLEU[7]を計算し、3 文書で生成された自然な質問の割合がどのように変化するか調査した。その結果を表 2 に示す。なお、1 行目にはフィルタリングがない場合の結果を示す。表 2 から、自然な質問を多く残せていることが分かる。ただ、不自然な質問も多く残り、全体に占める自然な質問の割合は 40%程度にとどまった。フィルタリングによる質問除去は一定の効果があるものの、改善の余地も十分にあるといえる。

3) https://github.com/cl-tohoku/AIO2_DPR_baseline/blob/master/scripts/download_data.sh に記載のスクリプトでダウンロードできる。
4) https://www.elastic.co/jp/
5) 関連文書は日本語 Wikipedia 記事をパッセージの集合に分割したものである。
6) https://huggingface.co/megagonlabs/t5-base-japanese-web
7) 用いたしきい値は 4 節のしきい値に準する。

表 2 フィルタリング前後の質問の内訳と、自然な質問の割合 (BS は BERTScore を表す)

BS	BLEU	残存数	自然な質問数	割合 [%]
なし	なし	50	18	36.00
75	50	35	14	40.00
75	70	35	14	40.00
60	50	46	16	34.78

4 質問応答システムへの応用

4.1 質問応答システムの学習

生成された質問を、質問応答システムの学習に応用する。まず、訓練用データセットの質問および、その各質問から生成された質問に対し、質問応答システム用の訓練データの選抜を行った。具体的には、AI 王で提供されている DPR [14] のベースライン[8]を用いて、日本語 Wikipedia の記事から文書 100 個を抽出し、その中に解答が含まれないデータは訓練データから外した。

また、生成された質問と、生成元の質問を用いて、BERTScore と BLEU による類似度計算[9]を行った。3 節に従い、BERTScore を用いる選別では、一定のしきい値以上の質問を、BLEU を用いる選別では、一定のしきい値以下の質問を採用した。そして、元の訓練用データセットに、これらのしきい値双方を満たすように選別された生成質問群を追加した。この際、追加する質問は生起確率の高い質問 3 つまでに制限した上で追加を行い、システムの学習を行った。また、予備実験において、無制限に質問を追加しても精度の向上が見られなかったため、追加するデータ量は各実験設定で 6,000 問とした。

評価においては、評価用データセットで DPR による関連文書 100 件の抽出を行い、解答が存在する文書を抽出できた質問を入力し、解答を生成した。そのうえで、データに存在する想定解答と、生成された解答が完全一致しているかを基準とする Exact Match (EM) による評価を行った。この実験を各しきい値の組み合わせごとに、異なるシード値を用いて 5 回実施し、EM による正答率の平均を算出した。

8) https://github.com/cl-tohoku/AIO2_DPR_baseline

9) BERTScore は 0 から 1 まで、BLEU は 0 から 100 までの値をとる。いずれも類似度が高いときに大きい値になるが、全く関連のない文同士では BERTScore は 0.6 程度になる。

表 3 各実験設定の詳細と評価データセットでの EM 正答率。文書数は生成時に用いた文書数を示す。

文書数	BS	BLEU	EM [%]
なし	なし	なし	75.00
3	なし	なし	76.39
3	0.60	50	76.72
3	0.75	50	**76.99**
3	0.75	70	76.67
1	なし	なし	76.12
1	0.60	50	75.06
1	0.75	50	75.52
1	0.75	70	75.45

4.2 結果

表 3 に、元のデータを使用した場合と、生成されたデータを様々なしきい値によって選別した際の、EM 正答率を示す[10]。

生成された質問をシステムの学習に用いたことで、精度は最大で約 2%向上し、提案手法によって生成された質問が学習に貢献していることが確認された。生成時の文書数が 1 つだけの質問を追加した場合と比較すると、同じしきい値で 3 文書を使用したケースが最大 1.7%前後正答率で上回った。この結果から、複数の文書の読解を要する質問を生成することの有効性が確認できた。3 つの文書を用いて生成した質問を利用した実験において、類似度による選別を行わない場合と行う場合を比較すると、選別を行った場合は 0.3%から 0.6%の正答率向上が確認でき、フィルタリングに一定の効果があることが示された。

5 おわりに

本研究では、FiD を用いて、複数文書の読解を要する質問データの生成を試みた。質問生成の結果、同じ解答を違う表現を用いて導くような質問が生成できた。また、質問応答タスクのデータとして追加した結果、正答率が最大 2%向上した。加えて単一文書だけの読解で解答できる質問は、精度の改善幅が小さく、生成データの有用性が確認できた。今後の研究では、BERTScore や BLEU のフィルタリングで対応しきれない、意味は全く違うが求められる解答が同じ質問を自動で生成する方法を追求したい。

10) しきい値については組み合わせが多いため、関連文書 3 つの質問での検証における上位 3 つの結果を示し、関連文書 1 つの質問での検証でも同じ組み合わせを使用した。

謝辞

本研究はキオクシア株式会社の委託研究において実施した。

参考文献

[1] Gautier Izacard and Edouard Grave. 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**, pp. 874–880, Online, April 2021. Association for Computational Linguistics.

[2] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[3] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[4] 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. Natural questions: A benchmark for question answering research. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 452–466, March 2019.

[5] Norio Takahashi, Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. Machine comprehension improves domain-specific Japanese predicate-argument structure analysis. In **Proceedings of the 2nd Workshop on Machine Reading for Question Answering**, pp. 98–104, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[7] 鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也. JAQKET: クイズを題材にした日本語 QA データセットの構築. 言語処理学会第 26 回年次大会 (NLP2020) 発表論文集, pp. 237–240, Online, March 2020. 言語処理学会.

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[9] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1342–1352, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[10] Lidiya Murakhovs'ka, Chien-Sheng Wu, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. MixQG: Neural question generation with mixed answer types. In **Findings of the Association for Computational Linguistics: NAACL 2022**, pp. 1486–1497, Seattle, United States, July 2022. Association for Computational Linguistics.

[11] 折原良平, 鶴崎修功, 森岡靖太, 島田克行, 狭間智恵, 市川尚志. クイズビジネスにおける作問作業支援. 言語処理学会第 28 回年次大会 (NLP2022) 発表論文集, pp. 1401–1405, Online, March 2022. 言語処理学会.

[12] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[14] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics.

A　学習時のハイパーパラメータ

表 4 に、FiD の学習で用いたハイパーパラメータを示す。

表 4　学習時のハイパーパラメータ

最大エポック数	5
合計バッチサイズ	100
学習率	1.0×10^{-4}
トークン数上限 (質問生成)	512
トークン数上限 (reader)	200
入力文書数 (reader)	100