T5 を用いた日本語の複雑な質問文に対する質問分解

小原 涼馬 [1]　秋元 康佑 [2]
[1] 北海道大学　[2]NEC データサイエンス研究所
obara.ryoma.s1@elms.hokudai.ac.jp　kosuke_a@nec.com

概要

質問応答システムで多段階の推論や数値演算などが必要な複雑な質問文を扱う 1 つの方法として、質問文を簡単な質問に分解する方法がある。本研究では深層学習による大規模言語モデルである T5 を用いて異なる設定で英語の訓練データを用いて日本語の質問分解を行う実験を行う。その結果多言語 T5 を用いるより、機械翻訳を組み合わせ単言語で学習を行う方がより適切な分解が行われることを確認した。

1　はじめに

質問応答 (QA) は自然言語処理の重要なタスクの 1 つである [1]。質問応答で一般的な問題設定である機械読解タスクのデータセットの 1 つに SQuAD[2] がある。しかし、SQuAD はモデルが与えられた文章を参照することで答えが得られる単純なファクトイド質問が多い。そのような単純な問題設定をより発展させた複雑な質問のデータセットとして HotpotQA[3] や DROP[4] などがある。HotpotQA は多段階の推論を必要とするマルチホップ質問 (multi-hop question) により構成されており、DROP は数値演算などを含む複雑な質問により構成されている。複雑な推論を必要とする質問を解くために、従来のエンドツーエンドの手法に加え、複雑な質問を複数の簡単な質問に分解するアプローチが提案されている [5]。エンドツーエンドの手法は内部がブラックボックスであり、解釈可能性が低いことや、質問分解によるアプローチは分解後の単純な質問（単なる計算や比較を含む）を処理するために既存の質問応答モデルやルールベースのシステムを利用できる利点がある。これらを踏まえ、本研究では質問分解によるアプローチに注目する。

BREAK[6] では複雑な質問を読解するために質問分解のための意味表現として QDMR(Question Decomposition Meaning Representation) が提案された。BREAK では QDMR を用いることで複雑な質問に対する質問応答の精度が向上することが示された。また、Guo ら [7] は BREAK のデータセットに対して T5[8] を用いて学習することで質問文から分解質問を出力することが可能であることを示した。また、モデルが出力した分解質問を用いて質問応答器を学習することで質問応答の精度が向上することを示した。

日本語での質問応答システムは研究されているが、日本語での質問分解のためのデータセットは存在していない。そこで、英語データセットである BREAK を用いて日本語での質問分解を行うことが考えられる。その時、多言語で学習する方法や機械翻訳を用いる方法で英語データと同じように質問分解ができるかまだ明らかでない。

そこで DeepL[*1]による翻訳を用いて単言語で T5 を学習する手法を 2 手法と多言語 T5 を用いる手法を実施し、その性能を評価した。人手評価の結果、単言語の学習の手法で 60 パーセントの精度で一定の基準以上の分解ができることを示した。

[*1] https://www.deepl.com

2 関連研究

2.1 BREAK データセット

BREAK[6] は複雑な問題に対してその分解を与えるデータセットである。BREAK では複雑な問題の読解のために QDMR(Question Decomposition Meaning Representation) を提案している。QDMR は質問に答えるために必要な分解のリストを構成する。例えば表 1 のような分解が与えられる。

BREAK の問題は HotpotQA や DROP を含む 10 個の QA データセットから構成されている。QDMR を得るためにクラウドソーシングを用いてアノテーションを行った。

Q	Return the keywords which have been contained by more than 100 ACL papers.
Q_1	papers
Q_2	#1 in ACL
Q_3	keywords of #2
Q_4	number of #2 for each #3
Q_5	#3 where #4 is more than 100

表 1　QDMR の例。Q_1 から Q_5 の分解質問を順に実行することで質問の答えが得られる。'#<数字>' は'<数字>番目の分解質問の答え' を表す。

2.2 T5

T5（Text-To-Text Transfer Transformer）[9] は、Google が 2020 年にリリースした大規模な自然言語生成モデルである。T5 は Transformer のアーキテクチャを使用した text-to-text モデルであり、文書要約や翻訳、質問応答、テキスト分類など、様々なタスクで高い精度を発揮することで知られている。また、Google Research による mT5(multilingual-T5)[8] は 101 の言語に対応し、その 101 言語で事前学習されたモデルも公開されている。

2.3 T5 による質問分解

Guo ら [7] は BREAK データセットに対して T5 を用いて学習することで質問分解が可能であることを示した。教師データの入力として質問文、出力として"<subQ>Q_1<subQ>Q_2<subQ>...Q_s" を与える。ただし Q_1,Q_2...Q_s は分解後の質問であり、<subQ >は特殊トークンである。評価として内省的評価と外省的評価を行っている。内省的評価では、BLEU[10] や ROUGE[11] を用いている。外省的評価では分解した質問を用いて質問応答パイプラインを構築し、その性能を評価している。結果としてベースラインを F1 値で 7.2 上回ることを示した。一方で質問分解の構造を考慮した評価は行われていない。

3 手法

3.1 問題設定

本研究は日本語質問文を入力として、その分解質問を日本語で出力することを目的とする。

3.2 提案手法

本研究では以下の 3 つの手法を提案する。

- 手法 1: 日本語 T5 を用いる。翻訳による日本語データを用いて学習を行う。
- 手法 2: 多言語 T5 を用いる。英語データを用いて学習を行う。[*2]
- 手法 3:T5 を用いる。学習は英語データを用いて学習を行う。推論時は日本語の入力質問を英語に翻訳して入力を行い、出力された結果を日本語に翻訳する。[*3]

モデルに対する入力は質問文、出力は分解質問をセミコロンで連結した文字列 "$Q_1;Q_2;...;Q_s$" である。[*4]推論時はモデルの出力をセミコロンで分割したものを分解質問とする。

[*2] mT5 のような多言語モデルに対して英語データのみで下流タスクの学習を行うことで、教師データの無い他言語に転移させることが可能であることが示されている。[8]

[*3] オリジナルの英語データが学習ができることや、将来的より大規模で高性能なモデルが利用できる可能性がある利点がある。

[*4] 分解質問の区切りは他の記号を利用することも考えられるが、本研究では BREAK のデータと同じセミコロンを用いる。

4 実験

4.1 実験設定

本実験では、データセットとして Hugging Face[*5] で提供されている BREAK データセット[*6]のうち、”QDMR-high-level”のサブセットのデータを用いた。また、そのうち”Reading Comprehension”のタスクである”CWQ”, ”DROP”, ”HOTPOT”の 3 種類のデータセットに由来するデータを用いた。BREAK が提供するテストデータは正解が与えられていないため、本実験では訓練データより 3,097 件をテストデータとして利用する。よって訓練データ 14,406 件、開発データ 3,130 件、テストデータ 3,097 件を用いて実験を行う。

日本語での質問分解を行うために BREAK データセットを DeepL を用いて翻訳した。またこの時、翻訳の精度を上げるため、翻訳前のデータに対して文頭の”return”を全て”please answer”に、”# <数字>”を”XXX <数字>”に置き換えてから翻訳を行なった。

T5 のモデルは手法 1 では `t5-base-japanese-web`[*7]、手法 2 では `mt5-base`[*8]、手法 3 では `t5-base`[*9]を使用した。学習時のパラメータは学習率 0.0001、バッチサイズ 16、エポック数 20 として開発データで最も性能の高いモデルを選択した。最も性能の高いモデルを選択する際は、metric として BERTScore[12][*10] を用いた。

4.2 評価指標

予測結果の評価は文の類似度による評価と、分解質問の構造による評価を行なった。文の類似度による評価は BLEU[10]、ROUGE[11]、METEOR[13]、BERTScore[12] を用いた。

さらに本研究では分解質問 $Q_1, ..., Q_n$ 中の参照構造を、各ノード i が分解質問 Q_i に対応し、Q_i 中に別の分解質問 Q_j に対する解答への参照が存在する場合に $i \rightarrow j$ のエッジを持つような有向グラフで表現する。そして生成された分解質問の有向グラフが正解のものと同型かどうかを二値で判定する評価指標を提案し、これを以後 isomorphic と呼ぶ。

4.3 実験結果

実験結果を表 3 に示す。ただしベースライン (baseline) は質問文をそのまま出力する手法である。多くの指標について、手法 1 が最も良い性能を示し、手法 3 も同等の性能を示した。一方で手法 2 は性能が大きく劣っていた。

	baseline	手法 1	手法 2	手法 3
BLEU	14.33	54.12	9.80	**54.13**
ROUGE1	49.23	**79.41**	45.82	79.36
ROUGE2	27.62	**63.39**	20.85	63.34
ROUGEL	42.13	**74.19**	38.77	73.98
METEOR	32.14	**73.88**	27.69	73.31
BERTScore	73.75	**89.50**	70.16	89.27
isomorphic	11.04	**70.13**	25.15	68.61

表 2　質問分解の実験結果。太字は各行で最もスコアの高いものを示す。

4.4 分析

4.3 節で示したように手法 2 の性能は他の手法と比べて大きく劣っていた。そこで手法 2 の出力を人手で確認したところ、出力中に英語と日本語が混在していたり、”return ...”などの英語データのフォーマットがそのまま出力されているなど、言語間の汎化が十分に起こっていないことがわかった。[*11]

5 人手評価

5.1 評価方法

4.2 の評価指標では、各分解質問文の妥当性や分解の構造が適切かの評価ができない。また、正解

[*5] https://huggingface.co/
[*6] https://huggingface.co/datasets/break_data
[*7] https://huggingface.co/megagonlabs/t5-base-japanese-web
[*8] https://huggingface.co/google/mt5-base
[*9] https://huggingface.co/t5-base
[*10] モデルは英語学習時は `roberta-large` 日本語学習時は `bert-base-multilingual-cased` を用いた。
[*11] 実際の出力例は付録も参照されたい。

データは翻訳データであるため、正解データが必ず適切であるとは限らない。そこで以下の基準に基づいて人手評価を行なった。

項目 1: 問題文で何を聞いているかわかるか
- 0: 何を聞いているかわかる
- 1: 少し不自然だが、何を聞いているかわかる
- 2: 質問文をなしていない・意味不明

項目 2: 分解により正しい答えが得られるか
- 0: 得られる
- 1: 分解質問は何を聞いているがわかるが、解釈次第で最終的な答えが得られる (解釈次第では得られないこともある)
- 2: 分解質問は何を聞いているかわかるが、最終的な答えを得るためには違うものを聞いている
- 3: 分解質問が何を聞いているかわからないため答えが得られない

項目 3: 項目 2 の理由
- 項目 2 で 1,2,,3 と判断したとき、どの分解質問がその根拠になったか

手法 2 は英語と日本語が混在した出力であり評価が難しいため、手法 1、手法 3 について評価を行なった。それぞれ 100 個の例 (ただし、”CWQ”, ”DROP”, ”HOTPOT”の問題から均等にサンプリングした) に対して著者ではない複数人の日本語母語話者（大学生、大学院生）による評価（各問題 1 名で評価）を行なった。

5.2 評価結果

評価結果を表 3 に示す。評価結果より、手法 1、3 どちらもある程度分解ができている (項目 2 が 0 か 1) ものが 60 個を超えているため、60 パーセントの精度である程度の分解ができていることがわかる。またこれらの分解に成功しているデータには複数の推論パターンが含まれていることも確認できた。[*12] 実際の分解例は付録に示す。

	数字	手法 1	手法 3
項目 1	0	69	81
	1	18	11
	2	13	8
項目 2	0	43	46
	1	17	20
	2	22	14
	3	18	20

表 3　人手評価の結果。各要素は個数を表す

5.3 エラー事例分析

項目 2 で 3 と判断されている場合にどのようなエラーが起こっているのかを調べた。その結果、「<名詞>についてお答えください。」のように分解質問が何を答えればいいか明確でないような事例が多く現れていた。前述した分解質問は原文の英語では’return <名詞>’ であり、BREAK ではその名詞のリストを返すことを想定している。このように BREAK では分解質問の意図（質問の種類）が必ずしも分解質問文に反映されておらず、本研究の単純な翻訳手順では本来の意図通りの日本語質問に翻訳されない課題がある。前述したエラーはこうしたデータセットの翻訳の課題が原因であると考えられる。

6 まとめ

本研究では、質問応答システムで日本語の複雑な質問文を扱うために、日本語の質問文をより簡単な質問に分解するモデルを学習できるかどうか試みた。その結果機械翻訳したデータセットを用いて日本語 T5 を学習する手法により、人手評価で 60% の分解精度を達成できることを示した。一方で翻訳に依らない言語間転移についての課題や、データセットが想定している質問の意図を考慮せずに翻訳することの問題点などが示された。今後は分解した質問を QA システムに入力することを考慮し、それに応じてどのような分解が適切であるのかの研究を行いたい。

[*12] BREAK の operator の種類に基づいて判断した。

謝辞
本研究は日本電気株式会社研究インターンシップで行ったものである。

参考文献

[1] Robert F. Simmons. Answering english questions by computer: a survey. *Commun. ACM*, 8:53–70, 1965.

[2] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing*, 2016.

[3] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

[4] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[5] Ruiliu Fu, Han Wang, Xuejun Zhang, Jun Zhou, and Yonghong Yan. Decomposing complex questions makes multi-hop qa easier and more interpretable. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

[6] Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198, 2020.

[7] Xiao-Yu Guo, Yuan-Fang Li, and Gholamreza Haffari. Complex reading comprehension through question decomposition. *ArXiv*, abs/2211.03277, 2022.

[8] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*, 2020.

[9] Colin Raffel, Noam M. 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. *ArXiv*, abs/1910.10683, 2019.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2002.

[11] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2004.

[12] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675, 2019.

[13] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, 2005.

付録

出力と人手評価の例

出力例を以下に示す。また、それに対する人手評価の結果を表 4 に示す。

出力例 1

質問文

台北を有する国では、どのような言語が話されているのでしょうか。

正解

' 台北がある国をお答えください。', 'XXX1 で使われている言語をお答えください。'

手法 1

' 台北がある国をお答えください。 ', 'XXX1 において使用されている言語をお答えください。'

手法 2

'return 言語が話されている 台北を有する国 ', 'return languages 話されている 台北を有する国 ', 'return languages in both #1 and #2'

手法 3

' 台北がある国をお答えください。', 'XXX1 において使用されている言語をお答えください。'

出力例 2

質問文

44 ヤードのゴールと比較して、26 ヤードのゴールは何本少なかったのでしょうか？

正解

'26 ヤードゴールのヤード数をお答えください。', '44 ヤードのゴールについてお答えください。', 'XXX1 と XXX2 の違いについて教えてください。'

手法 1

'26 ヤードだったゴールをお答えください。 ', '44 ヤードだったゴールをお答えください。 ', 'XXX1 の番号をお答えください。 ', 'XXX2 の番号をお答えください。 ', 'XXX3 と XXX4 の違いについてお答えください。'

手法 3

' 得点となった 26 ヤードのゴールをお答えください。', ' 決められた 44 ヤードのゴールをお答えください', 'XXX1 の番号をお答えください。', 'XXX2 の番号をお答えください。', 'XXX3 と XXX4 の違いを教えてください。'

出力例 3

質問文

2012 年 2 月 16 日から 10 月 28 日まで 27 回の公式マネーゴルフトーナメントも主催したオハイオ州コロンバスに本社を置く会社の従業員数は？

正解

' コロンバス (オハイオ州) に拠点を置く企業をお答えください。', '2012 年 2 月 16 日から 10 月 28 日まで 27 回の公式マネーゴルフトーナメントを主催した XXX1 についてお答えください。', 'XXX2 社の社員についてお答えください。', 'XXX3 の番号をお答えください。'

手法 1

'2012 年 2 月 16 日から 10 月 28 日まで 27 回のオフィシャルマネーゴルフトーナメントを開催したオハイオ州コロンバスに本社を置く会社をお答えください。 ', 'XXX1 社の従業員についてお答えください。 ', 'XXX2 の番号をお答えください。'

		項目 1	項目 2	項目 3
出力例 1	手法 1	0	0	-
	手法 3	0	0	-
出力例 2	手法 1	1	1	3,4
	手法 3	1	1	3,4
出力例 3	手法 1	0	3	2,3

表 4　出力例の評価