人間らしい予測処理機構を取り入れた質問応答モデルの提案：早押しクイズのパラレル問題を題材として

山下陽一郎[1] 原田宥都[2] 大関洋平[2]
[1]東京大学教養学部 [2]東京大学大学院総合文化研究科
{yamashita-yoichiro416,harada-yuto,oseki}@g.ecc.u-tokyo.ac.jp

概要

本研究では、早押しクイズの問題を題材として、不完全な質問文から解答を生成する質問応答モデルを提案する。提案モデルは、不完全な質問文に対して、その続きを予測して質問文の後半部を補完する前段と、そうして得られた質問文に対して解答を出力する後段からなる。このモデルの前段にはファインチューニングした GPT-2 を用い、後段には Dense Passage Retriever による質問応答モデルを用いる。結果として、質問文の前半部のみから解答を出力したモデルに比べて、予測機構を組み込んだモデルの方が高い精度の解答を出力できることがわかった。

1 はじめに

自然言語で表現された質問に対して適切な解答を与える技術は、質問応答と呼ばれている。特に、分野や領域が限定されていない質問応答はオープンドメイン質問応答と呼ばれる。その中でも単純な事実や出来事を問う質問はファクトイド型質問と呼ばれ、この技術は検索サービスなど身近な場面にも応用されている[1]。オープンドメイン質問応答のうちファクトイド型質問の研究では、クイズの問題が題材として扱われることが多く、SQuAD[2]や TriviaQA[3]といった英語のデータセットが存在する。2011 年には米国のクイズ番組 Jeopardy!において、IBM の開発した Watson というシステムが活躍した[4-5]。日本でもクイズの問題のデータセットである JAQKET が公開されている[6]。

さらに、クイズを題材とした質問応答研究においては、質問文の途中までを入力して解答を生成する、「早押しクイズ」モデルの開発も行われている [7-8]。これらの「早押しクイズ」モデルでは、主に不完全な質問文から「直接」解答となる単語を生成する手法が採用されているが、人間が早押しクイズを解く際は、まず、不完全な質問文からその続きを先読みし、その後に解答を導き出す。そのため、既存モデルのアプローチは「人間らしさ」の観点から妥当ではない、という可能性が残されている。よって本研究では、「問題文の先読み」に焦点を当て、「問題文の先読み＋解答の出力」という構造を持つモデルを提案する。人間のクイズプレイヤーと同様に、問題文の前半部からその後半部を予測し、完全な問題文を推測した上で正解を導き出すことを目指す。

2 関連研究

2.1 クイズを題材とした質問応答研究

「AI 王～クイズ AI 日本一決定戦～」（以下、「AI 王」）という日本語のクイズを題材とした質問応答コンペティションでは、22,000 問以上のクイズの問題が解答とともに公開されており、9 割を超える正解率を誇るシステムが開発されている。このコンペティションは、問題文全文から解答を導くという形式で、ベースラインモデルも併せて公開されている[9]。

2.2 早押しクイズの質問応答研究

Neural Information Processing Systems においては、Human-Computer QA という早押しクイズ形式の問題を題材としたコンペティションが開催された[7]。この大会は、逐語的に与えられる問題文に対し、モデルと人間のクイズプレイヤーが早押しクイズを競うというものであった。人間のプレイヤーを破り優勝したモデルは、与えられた問題文から正解となる単語の候補を列挙する Quiz Solver と、答えの型式を予測する Type Predictor によって構成されている[8]。このように、英語の問題文では、答えとなる単語が person, organization, product などのいずれのタイプに一致するかが早い段階で明示されるため、これを解答を導く際の手がかりとして利用できる。一

方で、日本語のクイズでは、このような答えの型が文末まで明示されない以上、クイズプレイヤーにとっては予測処理が重要だと考えられる。

3 提案モデルの構造

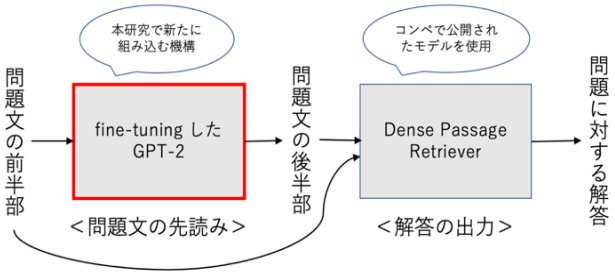


図 1 本研究の提案するモデル

本モデルは図 1 に示すように「問題文の先読み＋解答の出力」の 2 つの機構を組み合わせた構造である。モデルの前段では問題文の前半部の入力に対し、後半部を予測して出力する。モデルの後段では、問題文の前半部と前段が出力した後半部を合わせて完成した文を入力し、その問題の解答を出力する。

3.1 問題文の先読み: 本モデルの前段

前段では、問題文の前半部を入力として与え、その続きとなる後半部を補って出力させることで予測処理を実装する。実装には、Huggingface[10]上で rinna 社の公開している[11][i]GPT-2[12]をファインチューニングして用いる。

3.1.1 パラレル問題

早押しクイズの問題文には多様なタイプの問題が存在するが、本研究ではその中でも特に「パラレル」という形式の問題を扱う。パラレルの問題文は「ですが問題」とも呼ばれ、例えば

おおいぬ座のアルファ星はシリウスですが、こいぬ座のアルファ星は何？（答. プロキオン）

のように、「A は X ですが、A'は何？」という構造を持つ。パラレルの問題文では、本題は問題文の後半部に置かれるが、場合によっては問題文の前半のみを聞けば、その後半部を予測して解答できる。また、機械的な処理の観点からも、パラレルの問題文は、「ですが、」という特徴的な語によって問題文が明示的に前後半に分割できるため、前半部から後半部を予測するという目的に適っており、扱いやすいデータと言える。本研究では、問題文の先読みに焦点を当てるため、パラレルの問題文に絞って扱う。

3.1.2 パラレル問題の分類

本研究では、「ですが、」までを読んだ時に問題文の後半部が予測しやすいかどうかという観点から以下の 2 つのカテゴリに分類して分析する。[ii]

- カテゴリ 1：後半部を予測しやすい問題
  このカテゴリには、問題文の前後半で明らかな対比があり、後続する後半部が明らかなもの(1)、3 つの要素を対比させているもの(2)が含まれる。
  (1) 手の爪に施す化粧はマニキュアですが、足の爪に施す化粧は何？
  (2) 中国の三国時代において、魏の首都は洛陽、呉の首都は建業ですが、蜀の首都はどこ？
- カテゴリ 2：後半部を予測しにくい問題
  このカテゴリには、問題文の前後半の対比が明らかではなく、後続する後半部が一つに絞りきれないような問題が含まれる(3, 4)。
  (3) 日本の都道府県で、宮城県の県庁所在地は仙台市ですが、茨城県の県庁所在地は何市？
  (4) 砂糖やミルクを入れない紅茶をストレートティーといいますが、砂糖やミルクを入れないコーヒーを一般に何という？

3.1.3 データセット

データセットとして、以下の資料からパラレルの問題と解答を収集した。

- AI 王データセット[6]
  全 22335 問のうち、「ですが、/ますが、」を含むパラレルの問題 2167 問を抽出した。
- Quiz Works[iii]
  全 18477 問の各問題にタグが付されており、その中で「パラレル」のタグが付されているものを 850 問抽出した。
- クイズの杜[iv]
  二次利用がフリーになっている問題セットのみを使用し、その中から「ですが、/ますが、」を含むパラレルの問題 959 問を抽出した。

以上の資料からこのようにして得られた問題のうち、「ですが、/ますが、」を文字列として含むがパ

[i] rinna/japanese-gpt2-medium のモデルを使用した。
https://huggingface.co/rinna/japanese-gpt2-medium

[ii] 伊沢[13]は、早押しクイズの問題を解く際のクイズプレイヤーとしての視点から、パラレルの構造を持つ問題をさらに細かく 4 種類に分類している。

[iii] https://quiz-works.com/

[iv] https://quiz-schedule.info/quiz_no_mori/data/data.htm

ラレルの構造を持たない問題[v]を除外した結果、最終的に 3765 問のパラレルの問題とそれに対応する解答を得た。GPT-2 のファインチューニングに際しては、「ですが、/ますが、」の直後に区切り記号[SEP]を挿入して学習することで、与えた問題文の前半部に対して自然な問題文の後半部を出力できるようにファインチューニングした。

3.2 解答の出力: 本モデルの後段

後段では、前段で後半部を予測して補った問題文を入力として与え、その問題に対する適切な解答を出力する。実装には第 2 回 AI 王で公開されたベースラインモデル[7]を用いる。このモデルは、現在の質問応答の主流の技術の Dual Passage Retriever (以下、DPR)[14]を日本語の質問応答のために実装している。AI 王の評価用のデータセットを用いた正解率は 0.5865 である。

4 実験と考察

4.1 実験

rinna/japanese-gpt2-medium のモデルを、epoch 数を 10 としてファインチューニングを施した。パラレルの問題文を用い、以下の 4 条件で実験を行った。

a. 問題文全体を DPR に入力して解答を得る。
b. 問題文の前半部のみを DPR に入力して解答を得る。
c. 問題文の前半部をファインチューニングした GPT-2 に入力し、問題文の後半部を予測させた上で、前半部と結合させて DPR に入力して解答を得る。
d. GPT-2 をファインチューニングする際に、問題文の前半部で対比されている語を鉤括弧で括って学習した上で c 条件と同様に解答を得る。実際の早押しクイズでは、対比される語が強調されて読まれることが多いという慣例[13]に準えたものである。

a 条件は DPR の解答生成の精度をそのまま測るベースラインである。今回の実験の各条件での正解率は以下のようになると予測される。

a 条件 ＞c 条件 ＝d 条件 ＞b 条件

また、c 条件と d 条件の予測の精度を、「妥当な予測の割合」により評価する。「妥当な予測の割合」とは、GPT-2 が元の問題と同じ意味の後半部を正しく予測できたものに加え、問題文の前後半で対比構造を持ち、想定される解答が現実世界に存在するような問題文の予測を合わせた割合である。(5)のように、GPT-2 の予測が元の問題文とは異なるものの、入力された前半部と自然な対比構造を保っている場合は「妥当な予測」をしていると判定し、そうでない場合は予測は妥当でないと判定する。

(5)日本の初代内閣総理大臣は伊藤博文ですが、
元の問題文: アメリカの初代大統領は誰？
妥当な予測の例: 二代目の内閣総理大臣は誰？
妥当でない予測の例: 日本の初代大統領は誰？

4.2 結果

第 2 回 AI 王のデータセットで開発用として公開されている問題のうち、パラレルの問題 113 問を用いて実験を行った。各条件での正解率と「妥当な予測の割合」は表 1 の通りであった。

表 1 各条件でのモデルの正解率

正解率は元の問題文に対する正解率。c, d 条件の妥当な予測の割合の括弧内の値は、左から順にカテゴリ 1 , 2 の問題についての妥当な予測の割合。

条件	DPR の生成した解答の正解率	妥当な予測の割合
a. 問題文全体	0.354	--
b. 問題文前半部のみ	0.159	--
c. 後半部を予測生成	0.186	0.443 (0.62/0.36)
d. 後半部を予測生成（対比語を明示）	0.230	0.549 (0.70/0.46)

a 条件を見ると、DPR のパラレル問題の正解率が 0.354 と、開発データ全体の問題に対する正解率である、0.5865 を大きく下回っている。パラレルの問題は解きにくいタイプの問題だとわかる。

b 条件と c 条件を見ると、b 条件での正解率に比べて c 条件での正解率の方が高い。予測処理機構であるファインチューニングした GPT-2 が、正解率の向上に寄与しているとわかる。

d 条件の結果から、対比語を明示させた上で学習

[v] 例えば、「「東京湾」は 2 つの半島に囲まれていますが、それは房総半島と何半島でしょう?」という問題は「ますが、」を含むものの、パラレルの構造を伴っていない。これらの問題は学習データから除外した。

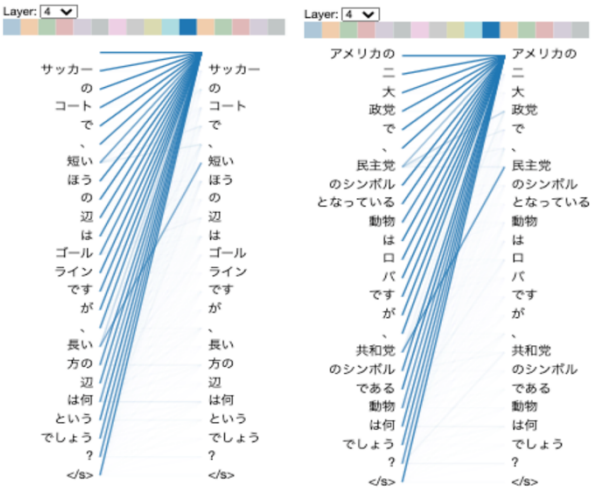


図 2 妥当な予測をした例の Self-Attention

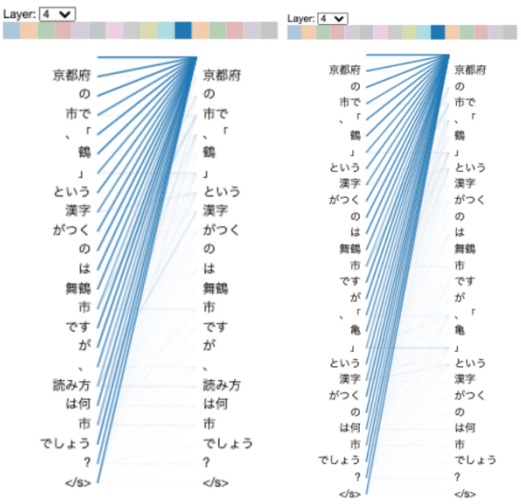


図 3 妥当でない予測をした例の Self-Attention

させると、正解率・妥当な予測の割合のどちらもが向上することが確かめられた。

また、d 条件における妥当な予測の割合は全体の 0.549 であった。問題のカテゴリに応じた妥当な予測の割合は、カテゴリ 1（予測しやすい対比）では 0.70, カテゴリ 2（予測しにくい対比）では 0.46 となり、クイズプレイヤーはカテゴリ 1 の問題を予測しやすいとする伊沢[13]の分類と整合する結果となった。

4.3 分析：Attention 重みの可視化

ファインチューニングした GPT-2 によるパラレルの問題文の予測で、モデルが実際に予測できた例とできなかった例のそれぞれについて、モデル内部の Self-Attention を可視化して分析を行った。

Attention 重みの可視化には BertViz[15][vi]を用いた。BertViz は、GPT-2 が文をエンコードする際の自己注意(Self-Attention)を可視化できる。

モデルが妥当な予測をしたいくつかの例について、特定の Attention-Head[vii]に注目した際の Self-Attention を可視化した結果が図 2 である。「長い」「共和党」という後半部の対比語をエンコードする際に、それぞれの前半部の対比語である「短い」「民主党」へ強く Attention を当てて対比関係を処理しているような Attention-Head が存在することがわかる。

一方で、図 2 と同じ Attention-Head に注目して、妥当でない予測を生成した例の Attention 重みを可視化した結果が図 3 である。予測した後半部の最初の語、すなわち「ですが、」の直後の語の Attention は、前半部の対比されている語に強く当たっていない。また、本来のパラレルの問題文を入力した際の Self-Attention についても、後半部の対比語である「亀」の Attention は、前半部の対比語である「鶴」に強く当てられていない。このように、問題文の前後半の対比が捉えられていないことがわかる。

5 おわりに

本研究では、クイズのパラレルの問題を題材として、予測処理機構を組み込んだ質問応答モデルを用いて実験を行った。結果として、質問文の前半部のみから解答を生成する際、予測処理機構は正解率の向上に寄与することが示された。さらに、問題文に特化した手がかりを明示的に示した上でモデルを学習することでも正解率が向上した。問題文の先読みという予測処理に限ると、5 割以上の問題で妥当な予測を行えることが確かめられた。また、モデルが予測しやすい問題の傾向は、クイズプレイヤーによる経験的な問題の分類と合致する結果となった。

モデルの予測の成否とモデルの Self-Attention に注目すると、問題文中の対比関係を反映しているような特定の Attention-Head の存在が明らかになった。モデルが妥当な予測をした例では後半部で対比されている語の Self-Attention が、それと対応する前半部で対比されている語に強く当たっているのに対し、妥当な予測をできなかった例ではその語に強い Attention が当たる傾向が見られないというように、当該の Attention-Head の振る舞いが異なることも確かめられた。

[vi] https://github.com/jessevig/bertviz

[vii] 実験に使用した GPT-2 のモデルは 24 層のトランスフォーマーの層からなり、各層に 16 の Attention-Head が存在する。今回の可視化に際しては、4 層目の特定の Attention-Head における Self-Attention に注目している。

謝辞

本研究は、JST さきがけ JPMJPR21C2 の支援を受けたものです。

参考文献

[1] 奥村学, 磯崎秀樹, 東中竜一郎, 永田昌明, 加藤恒昭. **質問応答システム**. コロナ社, 2009.

[2] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018. https://arxiv.org/abs/1806.03822

[3] Mandar Joshi, Eunsol Choi, Daniel Weld, , and LukeZettlemoye. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension, **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,** 2017.

[4] Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., & Welty, C. Building Watson: An overview of the DeepQA project. **AI magazine**, Vol. 31, No. 3, pp. 59–79, 2010.

[5] 金山博, 武田浩一. Watson: クイズ番組に挑戦する質問応答システム. **情報処理**, Vol. 52, No. 7, pp. 840–849, 2011.

[6] 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 回年次大会発表論文集, pp. 237–240, 2020.

[7] AI 王コンペティション実行委員. AI 王公式ベースラインシステム, 2019. https://sites.google.com/view/project-aio/baselines

[8] Escalera, S., and Weimer, M. **The NIPS'17 Competition: Building Intelligent Systems**. Springer International Publishing, 2018.

[9] Yamada, I., Tamaki, R., Shindo, H., and Takefuji, Y. Studio Ousia's quiz bowl question answering system. **The NIPS'17 Competition: Building Intelligent Systems**, pp. 181–194, Springer International Publishing, 2018.

[10] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., and Rush, A. M. Transformers: State-of-the-art natural language processing. **Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations**, pp. 38–45, 2020.

[11] 趙天雨, 沢田慶. 日本語自然言語処理における事前学習モデルの公開. 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 93 回 (2021/11), pp. 169-170, 2021.

[12] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, 2019.

[13] 伊沢拓司. **クイズ思考の解体**. 朝日新聞出版, 2021.

[14] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., and Yih, W. T. Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906, 2020.

[15] Vig, J. A multiscale visualization of attention in the transformer model. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019.

A 付録

A.1 GPT-2 が予測した問題文の例

ファインチューニングした GPT-2 に、パラレルの問題文の前半部を与えた際の出力の例を表 2 に示す。

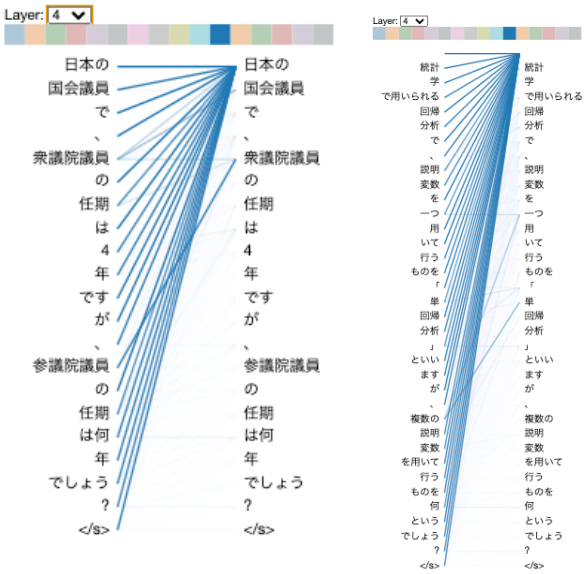
表 2 GPT-2 の予測した問題文

「ですが、/ますが、」までを入力として与えた際の GPT-2 の出力の例。妥当な予測な例と妥当でない予測の例をそれぞれ示す。出力に際しては、ビーム幅を 5 としてビームサーチを行った。

GPT-2 が予測した問題文	妥当な予測
アメリカの二大政党で、民主党のシンボルとなっている動物はロバですが、共和党のシンボルである動物は何でしょう?	○
xy 平面において、x と y がともにプラスなのは「第 1 象限」ですが、マイナスなのは何象限でしょう?	○
他人のものの方がよく見えることの例えで、赤く見えるのは隣の花ですが、青く見えるものは何でしょう?	○
ベートーベンの交響曲第 5 番は『運命』ですが、ピアノ協奏曲第 5 番は何でしょう?	○
音楽の楽譜で、最初にシャープが 1 つつくとト長調ですが、最後にフラットがつくと何長調でしょう?	
イギリスのエリザベス 1 世はチューダー朝の女王ですが、フランスのマリーアントワネットはどこの国の女王でしょう?	
昆虫が、卵からかえることを「孵化」といいますが、死んだあとに、その体から何が出てくることを「何」というでしょう?	
海の漁師にちなんだ「漁師風」という意味のパスタは「ペスカトーレ」ですが、豚の脂身を使った「ポーク風」という意味で、イタリア語で何というでしょう?	

A.2 可視化した Attention 重み

図 4　Attention 重みの可視化の例



モデルが問題文の妥当な予測をした例について、Attention 重みを可視化した結果を左に示す。本文セクション 4.3 の分析と同じ Attention-head に注目している。本文中の例と同様に、aa 問題文後半部で対比されている語「参議院議員」「複数の」の Self-Attention が、それぞれと対応する前半部の対比語「参議院議員」「一つ、単」に強く当たっていることが見て取れる。