嘘がなく、面白いクイズの自動生成

島田克行[1] 折原良平[1] 森岡慎太[1] 市川尚志[1]
[1]キオクシア株式会社
{katsuyuki1.shimada,ryohei.orihara,yasuhiro.morioka,takashi2.ichikawa}@kioxia.com

概要

本研究では、教育的効果や題材に対する興味の喚起を目的とした、回答モチベーションを高める早押しクイズ生成をテーマとした情報推薦に取り組む。具体的には、情報源からクイズ問題文に採用する情報を選定するプロセスを、人の手で作問されたクイズを教師データとした機械学習に基づく推薦システムにより再現する。

また、訓練した推薦システムの出力を用いてChatGPTによる問題文生成を行い、その「嘘のなさ」「面白さ」を評価する。

1 はじめに

配信サービスのTVerで「クイズ」で検索すると987件ものコンテンツがヒットする[1]。このように、クイズは日本人に浸透した娯楽であると言える。

また、QUIZ ROOM SODALITE[2]やクイズバースアール[3]といった、早押しクイズが体験できる施設の登場は、クイズが見るものからプレイするものに移り変わりつつあることを表しており、プレイヤーの人口は増えていくものと思われる。

また、クイズはその教育的効果にも注目が向けられている。例えばQuizKnockのサービスである「朝Knock」はクイズをエンターテイメントとして位置づけており、題材にした時事問題への興味を喚起し、自主的な学びを促進する効果を期待している。また、松田らは[1]で博物館における学習の促進を目的としたクイズの出題について検討している他、[2]では認知症に携わる医師がクイズの認知障害の予防への活用を示唆している。

クイズを活用するこれらのシーンでは人間を「正解者」と「不正解者」に分別するだけが目的ではないことは明らかである。しかし、どのようなクイズが面白いかは時と場所によって異なる。

また「面白さ」を評価する研究としては戎ら[3]、太田ら[4]、江連ら[5]があるが、面白さは多面的であり、これらが目指すものと、早押しクイズにおける面白さとは必ずしも一致するとは言いきれない。

プレイヤーの間で広く共有されている「面白いクイズ」の特徴の一つに、「前フリと後限定から構成される」というものがある。早押しクイズでは問題文が口頭で読み上げられることから、問題文が後半に差し掛かるにつれて次第に正答へと絞り込まれていく構成が好まれる。その結果、問題文前半では正答に関するあまり知られていない知識が、後半では誰でも確信できる情報が提示される。ここで問題文前半を「前フリ」、後半を「後限定」と呼ぶ。例えば伊沢[6]は早押しクイズの問題文は「前置き」「確定要素」「確定後補足」の3つの節からなるとしている。前フリは「前置き」、後限定は「確定要素」と「確定後補足」にそれぞれ対応する。

以下は理想的なクイズの例である。

- クレオパトラが酢に溶かして飲んでいたと伝えられる、海の宝石とも称される天然の美しい装飾品は何でしょう？（正答：真珠）

「嘘がなく、面白いクイズ」の作問は高度な作業であるため、作問者が限られる。そのため増大していく需要に対して十分な供給量が維持できているとは言えない。

GPTなどの大規模ニューラル文章生成モデルに個人が容易にアクセスできるようになった現在、GPTにクイズを生成させるという発想に至るのは自然なことである。しかし、GPTの文生成では、事実と異なる情報をもっともらしく生成してしまう、hallucination error(HE)と呼ばれる現象が生じる[7]。

[1] https://tver.jp/search/%E3%82%AF%E3%82%A4%E3%82%BA?genre=variety 2023年12月4日検索。
[2] https://quiz-sodalite.com/
[3] https://suahl.com/

質問応答タスクでの HE を避ける方法としては GPT が必要とする情報を追加で与える方法があり、Retriever-Reader モデルと呼ばれている [8]。

本研究では Retriever-Reader モデルを採用した自動作問を行う。特に、クイズを教師データとすることで、「クイズにした時に面白くなる情報」を上位に推薦する記事推薦モデルを訓練し、Retriever モジュールに用いる。また、Reader モジュールには ChatGPT を用いた自動作問のフローを構築し、その性能の評価を行う。

2 関連研究

クイズに関わる機械学習のテーマは「作問する」タスクと「回答する」タスクに大別される。

回答タスクの研究は Retriever-Reader モデルを採用したものがある [9] [10]。Retriever は質問文と関連が深い情報を記事集合から抽出するモジュールであり、Reader は質問文と Retriever によって抽出された情報に基づいて回答推定を行う。 [10]では BERT を用いた推定処理を行っているが、 [11]では chatGPT を用いたプロンプトエンジニアリングの方法も示されている。回答タスクでの性能向上のためには、Retriever モジュールには質問文と類似した関連情報を検索できることが期待され、類似度が高ければ高いほど Reader モジュールが正しい回答を推定できる可能性は高まる。そのため、Retriever モジュールの研究例は主に類似度が高い記事の抽出能力の向上に集中している。

一方で「クイズを作成する」タスクについての研究は前者と比べて盛んとは言いがたい。橋元ら [12]は早押しクイズのうち、「a は A ですが、b は何でしょう？」という文構造が特徴のパラレル問題に絞った自動生成について研究した。また [13]では時事問題を取り上げたクイズ生成について報告しているほか、第 4 回 AI 王[4]では初めて作問部門のコンペティションが開催されている。

3 実験

本研究では人間の作問プロセスが以下の 4 つのステップからなるとしており、後述の方法でそれぞれ適していると考えられる手段や値を選択した。

1. テーマの決定
2. 情報源の決定
3. 面白い情報の抽出
4. クイズ作問

3.1 テーマの決定

第 4 回 AI 王の作問コンペの予選ラウンドで提示された 20 個のテーマのうち、Wikipedia に登録されていた 19 個を採用した。

3.2 情報源の決定

クイズを作問する際、人間はあるテーマについて調べ、知りえたことから、クイズにしたときに面白い情報をピックアップすると考えられる。また、先述した通り「嘘がない」クイズを作問するには事実の誤りを含まない情報の集合を情報源として得る必要がある。そこで、そのテーマをタイトルにもつ Wikipedia の記事をセンテンスごとに分割したものを情報源として採用した。Wikipedia の記事は Wikimedia Downloads[5]から取得し、タグ等の除去を行った後、pySBD[6]によってセンテンスに分割した。なお、Wikipedia の該当記事の各センテンスはタイトルについて言及しており、その記述内容に誤りがないことを仮定している。

このようにして得られた 19 個のテーマと、それぞれのテーマが含む文の個数を表 1 に示す。

表 1　テーマとセンテンス数

テーマ	センテンス数
徳川家康	645
フランス革命	494
藤井聡太	335
セブン-イレブン	328
ノストラダムス	233
おにぎり	198
ChatGPT	194
サハラ砂漠	189
屋久島	182
自動販売機	166
乃木坂 46	109
灰原哀	100

[4] https://signate.jp/competitions/1234
[5] https://dumps.wikimedia.org/jawiki/ 2023 年 8 月に latest 版を取得。
[6] https://aclanthology.org/2020.nlposs-1.15/

フェルマーの最終定理	82
歯	75
真珠	73
カーボンニュートラル	70
旭川市旭山動物園	60
ホールインワン	21
7	9

3.3 面白い情報の抽出

「徳川家康」をテーマにする場合 645 のセンテンスから「面白い」ものを選定することになる。人間の作問者の情報選定を学習するため、以下の 2 つの大規模なクイズ大会の過去問を教師データとして用いた。abc/EQIDEN の過去問は AI 王の過去のコンペティションのデータセットとして提供されている。

- abc/EQIDEN[7]
- AQL[8]

収集したクイズに対しては下記の処理を行った。まず、クイズの問題/正答のペアのうち、正答をタイトルとする Wikipedia の記事があるものに絞ることで、人間の作問と同じフローを想定しての訓練を可能にした。続いて、各クイズは spaCy[9]および GiNZA[10]を用いて構文解析を行い、係り受け関係に基づいて前フリ部と後限定部に分割した。

教師データは 2 通りのクイズ大会に加え、クイズセンテンスは「全文」「前フリ部」「後限定部」と 3 通りあるため、全体では 6 通りとなる。それぞれの教師データのサンプル数を表 2 に示す。

表 2　教師データとサンプル数

大会	分割	問題数	センテンス数
abc/EQIDEN	なし	15132	1245636
abc/EQIDEN	前フリ	15132	1245636
abc/EQIDEN	後限定	8054	545943
AQL	なし	9425	236993
AQL	前フリ	9425	236993
AQL	後限定	5612	109612

情報源に存在するセンテンスがクイズに採用されたかは、問題文とセンテンスの類似度に基づいてスコア付けした。スコア付けは双方の文から名詞・動詞・形容詞のみを残し、単語単位でのマッチングによって計算した。

説明変数としては各センテンスの BERT による埋め込み$\boldsymbol{d}$、各正答の BERT による埋め込み$\boldsymbol{a}$と$\boldsymbol{a}$,$\boldsymbol{d}$ の cos 類似度s、そして各センテンスが記事の中で登場する相対位置locで、全体で 1538 次元である。

これらの説明変数・目的変数から、tensorflow-ranking [14]を用いて推薦システムの訓練を行った。

3.4 クイズ作問

クイズの作問には ChatGPT-3.5turbo 相当の環境を用いた。人間が作問した、前フリ・後限定からなるクイズを 1 問入出力例として示す One-shot learning を行ったうえ、3.3 節で訓練したモデルが上位に提示したセンテンス 10 件を箇条書きし、追加入力として用いた。

4 結果

結果を示す。

まずは推薦システムそのものの性能を測るための DCG(Discounted Cumulative Gain)である。DCG は下式で定義した。g(r)は人手で付加した 0〜5 の 6 段階のスコアである。

$$\mathrm{DCG} = \sum_{r=1}^{10} \frac{2^{g(r)} - 1}{\log_2(r+1)}$$

続いて、作問されたクイズの評価である。文中に矛盾や誤りがないかを検証し、誤りがある場合は「嘘がある」とした。また、「面白いか」は 0〜5 の 6 段階で評価した。これらはどのモデルから出力されたクイズかを伏せ、人手での評価を行った。

DCG および作問されたクイズの評価にあたっては、表 2 で示した 6 つのモデルと比較するベースラインモデルとして、センテンスと回答の BERT ベクトルの類似度が高いものから抽出する Similarity モデルを追加し、表 3 に示す 7 通りのモデルを扱った。

表 3　評価したモデル

モデル名	大会	分割
Similarity	-	-
ABC-nosplit	abc/EQIDEN	なし
ABC-split-F	abc/EQIDEN	前フリ
ABC-split-S	abc/EQIDEN	後限定

[7] https://abc-dive.com/portal/
[8] https://www.quizaql.com/
[9] https://spacy.io/
[10] https://www.megagon.ai/jp/projects/ginza-install-a-japanese-nlp-library-in-one-step/

AQL-nosplit	AQL	なし
AQL-split-F	AQL	前フリ
AQL-split-S	AQL	後限定

4.1 DCG

表 2　教師データとサンプル数 Similarity を基準とした、各モデルの DCG を Similarity と比較した勝敗を表 4 に示す。ABC-nosplit, ABC-split-S,AQL-nosplit の DCG が Similarity と比較して高い傾向があった。-nosplit は問題文全体との単語マッチングのために教師データに明確な特徴が付加されやすかったこと、-S は後限定であり、早押しクイズに特徴的な言い回しを多く含んでいたことが DCG を高めた要因と考えられる。

また、これらのモデルを推薦モデルとして用いて作問したクイズが、他のモデルやベースラインと比較して面白くなることが期待される。

表 4　モデルの勝敗

	Better	Even	Worse
ABC-nosplit	**10**	**3**	**6**
ABC-split-F	7	3	9
ABC-split-S	**10**	**2**	**7**
AQL-nosplit	**10**	**3**	**6**
AQL-split-F	7	4	8
AQL-split-S	8	4	7

4.2 クイズの面白さ

続いて、モデルが提示した文章に基づいて生成したクイズの面白さを評価し、図 1 に示す。推薦モデル採用の効果を評価するために、情報を与えずに生成したクイズを同時に採点し、図中には noInfo として掲載した。noInfo および Similarity と比較して人手で評価したスコアが高いことから、クイズを教師データとしたモデルから作問したクイズは面白い情報を抽出できているといえる。

ただし、4.1 節で言及した通りに ABC-nosplit, ABC-split-S,AQL-nosplit がもっとも面白いという結果にはならなかった。chatGPT による作問のさい、箇条書きで与えた独立した情報が一塊の説明文と認識されたり、主語を欠いたセンテンスの主語を取り違うといった現象が生じていたためであり、抽出した情報から適切に作問できれば十分面白いクイズが作問できていた可能性が高いと考えられる。

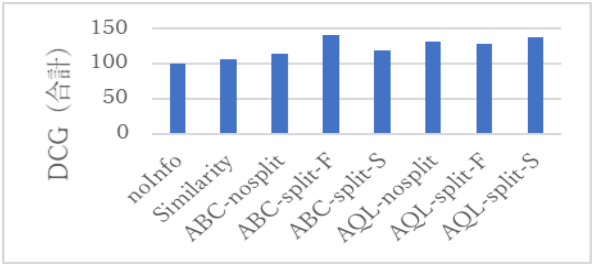

図 1　モデルごとのクイズの面白さ

4.3 クイズの嘘のなさ

続いて、クイズの本文が事実の誤認や誤りを含んだケースの総数を図 2 に示す。AQL-nosplit がもっとも少なく、19 問作問した中で誤りを含んだクイズは 4 問であり、情報を与えない noInfo の 11 問と比べて半分以下に減らせている。誤りを含む問題文が生成される頻度は少なくなっている。

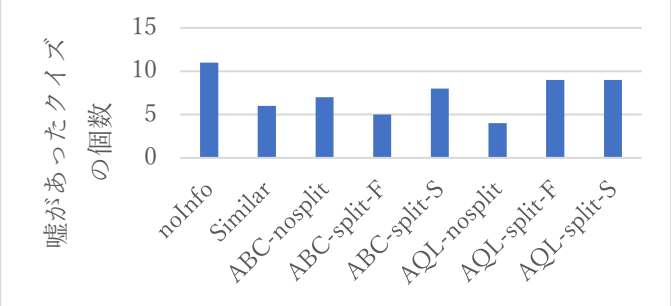

図 2　誤りを含んだクイズの総数

5 おわりに

大規模言語モデルを用いた「嘘がなく、面白いクイズ」の自動作問のための、Retriever に推薦システムを採用した Retriever-Reader モデルの構築と、その評価を行った。Retriever-Reader モデルを用いることで、生成される問題文に嘘や事実の誤認が含まれる頻度を抑えられることを示した。また、人間の作問したクイズを教師データとして訓練した推薦モデルにより、類似度だけに基づき情報抽出するよりも面白いクイズが作問されることを示した。

本研究ではある回答について作問する際、「その回答をタイトルとする Wikipedia 記事」に情報源を絞った点、記事の分割単位をセンテンスとした点、用いた埋め込みの次元数が小さい点に制約がある。例えば Wikipedia 全体をパラグラフ単位で分割し、RAG 等によって抽出した記事集合に対して推薦モデルを適用するなどの方法で、さらに嘘がなく、面白いクイズの精度が向上する可能性がある。

参考文献

1. 赤嶺有平,松田意仁,根路銘もえ子,. 博物館展示のための自然言語処理による質問文生成手法. 情報処理学会, 2021.
2. 精神科医・森隆徳が語る医療に活きるクイズ. (オンライン) 2019 年. https://www.qbik.co.jp/contents/mori_001/.
3. 戎達也,原尚幸. 大喜利における回答の面白さに関する定量的分析－お題と回答の意味的類似度からの考察－. 言語処理学会 第 27 回年次大会, 2021.
4. 太田聖三郎,野村理朗,河原大輔. 機械学習を用いた川柳の面白さの予測. 言語処理学会 第 29 回年次大会, 2023.
5. 江連三香, 内海彰. ユーモアを含む言語表現の解釈モデルに関する研究. 言語処理学会 第 5 回年次大会, 1999.
6. 伊沢択司. クイズ思考の解体. 朝日新聞出版, 2021.
7. 大野瞬,森脇恵太,杉山弘晃,酒造正樹,前田英作,. Transformer による hallucination error の事後修正. 言語処理学会 第 28 回年次大会, 2022.
8. 蓬田綾香,竹野貴法,村瀬文彦,平野徹,三谷陽,坂一忠,飯田哲也,岩堀恵介,. 技術ナレッジ活用に向けた Retriever-Reader モデルの検証. 言語処理学会 第 29 回年次大会, 2023.
9. 加藤拓真,鈴木潤,宮脇峻平,西田京介. オープンドメイン QA における DPR の有効性検証. 言語処理学会 第 27 回年次大会, 2021.
10. 初鹿憂,柴田千尋. オープンドメイン質問応答における文集合と位置情報を用いた抽出精度向上に関する検証. 言語処理学会 第 29 回年次大会 発表論文集, 2023.
11. 山田育矢,李凌寒,鈴木正敏,山田康輔. 大規模言語モデル入門. 技術評論社, 2023. ISBN 978-4-297-13633-8.
12. 橋元佐知,佐藤理史,宮田玲,小川浩平. 早押しクイズのパラレル問題の自動生成. 言語処理学会 第 28 回年次大会, 2022.
13. 折原良平,市川尚志,鶴崎修功,森岡靖太,島田克行狭間智恵. 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, 2022.
14. Pasumarthiet al.R.K. TF-Ranking: Scalable TensorFlow Library for Learning-to-Rank. Proc. of KDD'19, 2019. ページ: pp.2970-2978.