対話型学習のための 質問応答を用いた高品質な質問文の自動生成

瀬井　雄太†　　牛尼　剛聡††

† 九州大学芸術工学部　〒 815–8504 福岡県福岡市南区塩原 4-9-1
†† 九州大学大学院芸術工学研究院　〒 815–8504 福岡県福岡市南区塩原 4-9-1
E-mail: †sei.yudai.373@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　現在，学習者がシステムと会話して学びを深めるための対話型学習システムが注目されている．対話型学習システムにおける重要な課題の一つに質問生成がある．従来の自動質問生成手法は必ずしも正しい質問を生成するとは限らないため，生成した質問候補を評価し，正しい質問を抽出する処理が重要である．正しい質問を抽出するために，自動生成された質問に対して質問応答モデルを適応し，その解答と確信度に基づいて質問の正しさを評価することが考えられる．本研究では，質問応答モデルの学習に適した負例を自動生成することにより，質問応答モデルによる検証の精度を向上させる手法を提案する．評価実験において，提案手法を用いることで，より高品質な質問を抽出できることが示された．

キーワード　質問生成，質問応答，対話システム，学習支援

1　はじめに

近年，教育に ICT 技術を取り入れる動きが加速している．令和 3 年に文部科学省が発表した学校における教育の情報化の実態等に関する調査結果 [1] によると，教育用コンピュータ一台に対する児童の数はこの数年で急激に減少しており，児童や生徒一人一人が自分の教育用コンピュータを使えるようになってきている．文部科学省が掲げる GIGA スクール構想 [2] では，児童一人が一台のコンピュータを持ち，「多様な子供たちを誰一人取り残すことなく、 子供たち一人一人に公正に個別最適化され、資質・能力を一層確実に育成できる教育 ICT 環境の実現」を目標としている．また，近年，ChatGPT[1]など，人間がコンピュータと対話できるようなサービスや技術が開発されている．これらの対話生成の技術を利用することで，学習者個人に適した教育を行う対話型学習システムの実現が期待されている．

文部科学省が公示している高等教育学習指導要領 [3] では，主体的・対話的で深い学びを実現するための授業改善について述べられている．Nestojko ら [4] は，他人に教えることを前提として学習することで，学習効率を高めることを明らかにした．また，ベネッセ総合教育研究所のレポート [5] によると，グループ検討により，ある議題に対する認識の改善や視点の多様性の向上という結果が出ている．また Schroeder ら [6] は，教育エージェントに関する 43 の研究のメタ分析を行い，教育エージェントが学習に効果をもたらすことを明らかにした．

このように，対話的な学習および教育エージェントを利用した学習は一定の効果があることが示されているが，対話的な学習を実践するにあたって，いくつかの課題が考えられる．例えば学習者が対話する相手をすぐに見つけられない場合がある．たとえ，学習者が友達，家族，教師などと気軽にコミュニケーションを取れるような関係を築いていたとしても，相手と都合が合わないと対話的な学習を行うことはできない．多様な視点を得るという点では，対話的な学習に参加する人数は多い方が良いと考えられるが，多くの人を集めるのは困難である．また，対話的な学習を実践するのが困難な場合として，他の参加者の意欲が低い場合がある．たとえ参加者が集まったとしても，議題に対して興味関心がなく，対話的な学習への意欲がなければ，効果的な学習を行うことは困難であると考えられる．

上記の理由から，学習者はいつでも，自分のしたいときに対話的な学習を実践できるわけではない．対話型学習システムがあれば，学習者は実在の人間の都合に合わせることなく，いつでも，好きなときに対話的な学習を行うことができる．そのため，対話型の学習システムの有用性が高いと考えられる．

本論文では，対話的学習システムにおける重要な機能の一つである質問文生成機能に注目する．システムが生成する質問を工夫することにより，テストを行って学習効果を検証できたり，理解を深めたり，新たな知識を身につけたりすることが期待できる．本論文の貢献は以下の通りである．

- 高品質な質問文の抽出に適した負例を自動的に生成する手法を開発した．
- 質問応答モデルのファインチューニングを負例のあるデータセットと負例のないデータセットで行い比較した結果を示した．
- 既存の質問文生成モデルが生成した質問文に対して，質問応答モデルを用いて評価を行い，正しい質問文を抽出できることを示した．

本論文の構成は次の通りである．2 章で関連研究について述べる．3 章で質問生成機構の全体像を述べる．4 章で提案手法

1：https://chat.openai.com/

について述べる．5 章で実験について述べる．6 章でおわりについて述べる

2 関 連 研 究

これまでにも，質問文生成に関する研究は行われている．

田村ら [7] は日本の歴史上の人物の Wikipedia の記事からルールベースで質問文を生成する手法を提案している．この研究では，記事概要部の文末表現を調査し，パターンに合わせたフレーズを追加することで，疑問文を生成している．この方法で生成された疑問文の 86.5%が文法的に正しいという評価を得ている．しかし，ルールベースでの質問文生成は，生成アルゴリズムの設計やその調査にコストがかかる点や，ルールから外れた平叙文を質問文に変換できず質問の種類の多様性が制限される点が問題である．

ルールベース手法の問題点を解決するために，近年では機械学習モデルを用いた質問文生成手法が提案されている．Du ら [8] は，アテンション機能を加えた Encoeder-Decoder モデルによる質問文生成モデルを構築している．この研究では，文章から指定する答えを含む一文を抽出して，モデルの入力とすることで質問文を生成している．また，提案手法は文法的な自然さと質問の難易度でルールベースによる質問生成手法を上回っていると報告されている．その後，Zhao ら [9] や Kriangchaivech ら [10] は一文だけではなく段落を入力とする質問生成モデルを提案している．

しかし，機械学習モデルを用いた質問生成では，与えられた答えが正解となるような意味的に正しい質問が生成されるとは限らない．Zhu ら [11] は BERT [12] をベースに構築された質問応答モデルが正解できるかを報酬として強化学習を行ない，意味的に正しい質問文を生成する手法を提案している．この研究では encoder-decoder モデルを質問生成モデルのベースラインとして構築し，そのモデルを強化学習によって調整している．強化学習の報酬を流暢性，類似性，解答可能性，関連性に大別し，それぞれでいくつかの評価機構を構築している．この研究の実験では生成した質問に対して，自動評価と人間による評価を行っている．その結果，解答可能性と類似性による報酬で調整を行ったモデルの評価が高かったという結果を得ている．この研究の BERT 質問応答モデルの学習に使われている SQuAD1.0 [13] は，Wikipedia の記事から人間によって作られた 10 万以上の質問と答えペアを持つデータセットである．SQuAD1.0 の質問の全てに，対応する Wikipedia の記事から解答可能であるというラベルが付けられている．このようなデータセットで学習した BERT 質問応答モデルは質問が与えられた文章から解答可能でない場合や質問が文法的に正しくない場合でも解答を出力する．そのため，不適切な質問であっても質問応答モデルによる評価が高くなる可能性がある．

上記を踏まえて本論文では，不適切な質問を負例として加えたデータセットで学習させた BERT 質問応答モデルで質問を評価することで，より高品質な質問文を生成する手法を提案する．

3 質問生成機構

本研究における質問生成機構の全体像を，図 1 に示す．この機構は主に質問生成モデルと，質問応答モデルから構成される．質問生成モデルが出力した質問文に質問応答モデルが解答することによって質問文を評価し，低品質な質問文をフィルタリングする．以下に，質問生成モデル，質問応答モデルについて述べる．

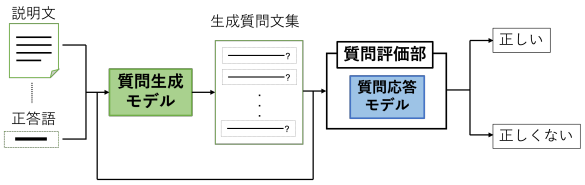

図 1 提案手法の全体図

3.1 質問生成モデル

質問生成には，T5 [14] をベースに園部が開発した，T5 日本語質問生成モデル [2]を用いる．T5 は質問生成，文書分類，質問応答，要約，翻訳などの自然言語処理のタスクを 1 つモデルで解くことができる汎用的な大規模言語モデルである．T5 日本語質問生成モデルは園部が構築した日本語 T5 事前学習済みモデル [3]を，SQuAD1.1 を日本語に翻訳して不正なデータをクレンジングしたデータセットを用いて，ファインチューニング行って，開発されたものである．T5 日本語質問生成モデルは，文章と文章中に含まれる単語を入力することで，その単語が答えとなるような質問文と生成確信度を出力する．

ここで，本研究で扱う用語について以下のように定義する．質問生成モデルに入力する文を**説明文**と呼ぶ．質問生成モデルに入力する文中に含まれる単語を**正答語**と呼ぶ．質問生成モデルが出力した質問文を**生成質問文**と呼ぶ．

ここで，説明文集合を $\mathbf{E}$，正答語集合を $\mathbf{A}^{True}$ を与えられたものとする．説明文 $e \in \mathbf{E}$ と正答語 $a^{True} \in \mathbf{A}^{True}$ のペア (e, a^{True}) に対して，質問集合を返す関数 QG を式 (1) で定義する．

$$QG(e, a^{True}) = \{qg_1, qg_2, ...\} \tag{1}$$

なお，それぞれの質問 qg_i は質問文 q_i と生成確信度 c_i^{QG} のペア $qg_i = (q_i, c_i^{QG})$ として定義される．

例として，Wikipedia の「オランダ風説書」の記事の冒頭を説明文として，オランダ風説書を正答語として T5 質問生成モデルに入力して質問を生成する様子を図 2 に示す．

3.2 質問応答モデル

質問解答には，BERT をベースとした質問応答モデルを用いる．BERT 日本語質問応答モデルは，文章と文章に関連する質問文を入力すると，質問文に対する解答語と解答確信度を出力

2：https://huggingface.co/sonoisa/t5-base-japanese-question-generation
3：https://huggingface.co/sonoisa/t5-base-japanese

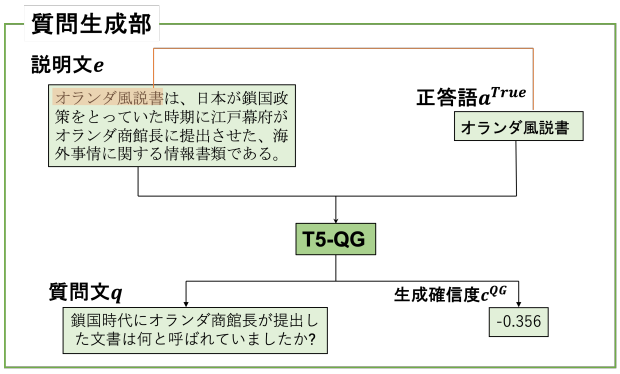

図 2　質問生成の例

する．解答の根拠となる説明文 e と質問文 q に対する質問応答モデルの解答 qa を生成する関数 QA を以下の式 (2) で定義する．

$$QA(e,q) = qa \tag{2}$$

なお，解答 qa は解答語 a^{QA} と解答確信度 c^{QA} のペア $qa = (a^{QA}, c^{QA})$ として定義される．

例として,「オランダ風説書」に関する説明文と質問文のペアを BERT 質問応答モデルに入力して解答を生成する様子を図 3 に示す．

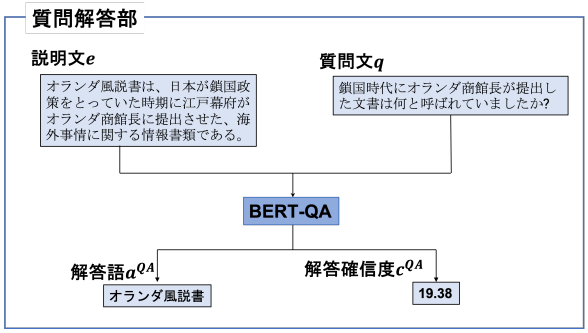

図 3　質問応答の例

ここで利用する質問応答モデルは解答抽出型のモデルである．図 4 に解答語を抽出するまでの流れを示す．本モデルでは，入力文章から解答語の先頭のトークンと解答語の末尾のトークンを指定し，その範囲が解答語として抽出される．図 4 のように，特殊トークン [CLS]，質問文，特殊トークン [SEP]，説明文で構成される一連の文章を入力文章とし，これをトークンに分割したものを $\mathbf{x} = \{x_1, x_2, ..., x_l\}$ とする．$\mathbf{x}$ に対する BERT の最終層の各トークンの隠れベクトル集合を $\mathbf{h} = \{h_1, h_2, ..., h_l\}$ とする．このとき，解答語の先頭である可能性集合 $\mathbf{p}^s$，および解答語の末尾である可能性集合 $\mathbf{p}^e$ を，

$$Linear(\mathbf{h}) = (\mathbf{p}^s, \mathbf{p}^e) = ((p_1^s, p_1^e), (p_2^s, p_2^e), ..., (p_l^s, p_l^e)) \tag{3}$$

と表すことができる．

ここで，説明文 e，質問文 q に対する質問応答モデルの解答 $QA(e,q)$ の解答確信度 c^{QA} を，解答語の先頭である可能性集合 $\mathbf{p}^s$，および解答語の末尾である可能性集合 $\mathbf{p}^e$ を用いて，

$$c^{QA} = max(\mathbf{p}^s) + max(\mathbf{p}^e) \tag{4}$$

と定義する．

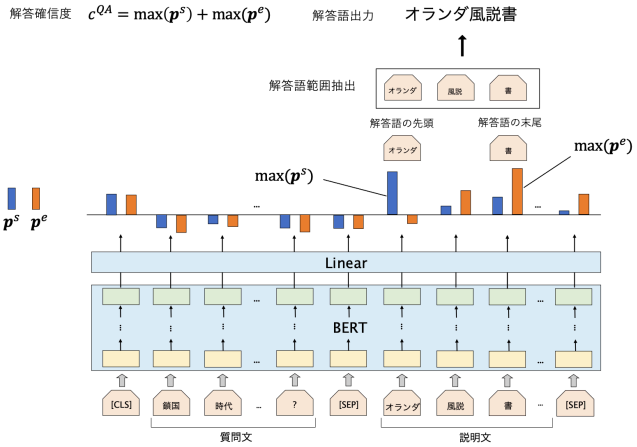

図 4　質問応答モデルの解答語と解答確信度の出力

4　提 案 手 法

4.1　概　　要

本論文で提案する高品質な質問文を生成する手法の処理は以下の通りである．

（1） 質問生成モデルにより候補となる質問文集合を生成する．

（2） 生成されたそれぞれの質問文に対する解答を質問応答モデルを利用して生成する．

（3） 質問応答モデルの解答語と解答確信度に基づいてそれぞれの質問文の品質を評価する．

（4） 評価の高い質問文をユーザーに提示する．

4.2　質問文の評価

質問文が正しいと考えられるには以下の 3 つの条件を満たすこと必要である．

（1） 質問文は意味を捉えることができる．

（2） 質問文は説明文を根拠として解答することができる．

（3） 質問文は正答語が答えとなる．

(1) は質問文の意味を捉えることができなければ学習に用いることができないためである．(2) は質問文が説明文を読んでも解答できない場合は学習者に解答の根拠を示すことができず，学習に向かないためである．(3) は指定した正答語が質問文に対する解答語と一致しない場合は，生成の意図にそぐわないためである．

質問文を評価するために，まず，生成質問文に対する質問応答モデルの解答語が正答語と一致しているかを調べる．生成質問文 q を生成する際に質問生成モデルに入力した正答語 a^{True} と生成質問文 q に対する質問応答モデルの解答語 a^{QA} について以下の関数を定義する．

$$match(a^{True}, a^{QA}) = \begin{cases} 1 & (a^{True} = a^{QA}) \\ 0 & (a^{True} \neq a^{QA}) \end{cases} \quad (5)$$

ここで，質問応答モデルを用いた質問文の評価おいて，2 つの生成質問文 q_i，q_j に対して，q_i が q_j よりも適切であると評価するとき，その関係を

$$q_i \geqq_{prop} q_j \quad (6)$$

と定義する．

生成質問文 q_i において，正答語 a_i^{True}，質問応答モデルの解答語 a_i^{QA}，解答確信度 c_i^{QA} を定め，生成質問文 q_j において，正答語 a_j^{True}，質問応答モデルの解答語 a_j^{QA}，解答確信度 c_j^{QA} を定める．このとき q_i と q_j の適切さの評価の順序関係 $\geqq_{prop}$ を以下のように定める．

$$\begin{cases} q_i \geqq_{prop} q_j & (match(a_i^{True}, a_i^{QA}) = 1 \\ & \text{かつ } match(a_j^{True}, a_j^{QA}) = 1 \\ & \text{かつ } c_i \geqq c_j) \\ q_i \geqq_{prop} q_j & (match(a_i^{True}, a_i^{QA}) = 1 \\ & \text{かつ } match(a_j^{True}, a_j^{QA}) = 0) \\ q_i \geqq_{prop} q_j & (match(a_i^{True}, a_i^{QA}) = 0 \\ & \text{かつ } match(a_j^{True}, a_j^{QA}) = 0 \\ & \text{かつ } c_i \leqq c_j) \end{cases} \quad (7)$$

式 (7) は質問文に対する解答語と正答語が一致することを重要視して定義されている．式 (7) の一番上の条件式は，解答確信度が大きい方が，実際に質問文に対する解答語と正答語が一致する可能性が高いという考えに基づいている．式 (7) の一番下の条件式は，解答語と正答語が一致していない時は解答確信度が小さい方が，実際には質問文に対する解答語と正答語が一致する可能性が高いという考えに基づいている．

ここで，式 (7) の関係を満たすようなスコア関数を設計する．スコア関数 $SCORE(q_i)$ と $SCORE(q_j)$ は以下の関係を満たす．

$$q_i \geqq_{prop} q_j \text{のとき} \quad SCORE(q_i) \geqq SCORE(q_j) \quad (8)$$

ここで，シグモイド関数を以下のように定義する．

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \quad (9)$$

式 (7) のように，生成質問文 q に対するスコアは，生成質問文 q を生成する際に質問生成モデルに入力した正答語 a^{True}，生成質問文 q に対する質問応答モデルの解答語 a^{QA}，および式 (4) で定義される解答確信度 c^{QA} を用いて表すことができる．よってスコア関数 $SCORE(q)$ を

$$\begin{aligned} SCORE(q) &= SCORE(a^{True}, a^{QA}, c^{QA}) \\ &= sigmoid(c^{QA} \times (-1)^{match(a^{True}, a^{QA})}) \\ &\quad + match(a^{True}, a^{QA}) - 1 \end{aligned} \quad (10)$$

と定義する．図 5 に解答確信度の値によるスコア関数の値の変化を示す．

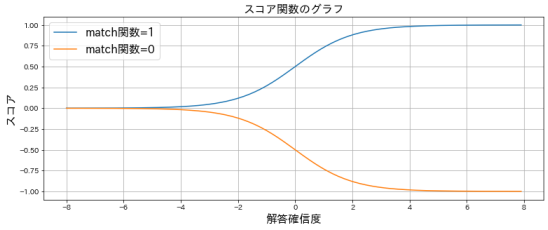

図 5　解答確信度とスコア関数

4.3　負例を加えたデータセット作成

質問応答モデルのファインチューニングに用いる JSQuAD [15] は正しい質問文のみのデータセットである．JSQuAD は Wikipedia の記事の段落ごとにいくつかの質問文とその答えが与えられている．質問文は該当する段落から解答できるように人手で作成されている．このデータセットでファインチューニングを行った質問応答モデルは，質問文に対して，文章中から解答となりそうな範囲を抽出するだけである．そのために，生成質問文が不適切である場合でも，正答語と解答語が一致してしまう問題が起こる可能性が高い．この問題に対応するために，質問応答モデルが正しい質問（正例）と正しくない質問（負例）を区別して，負例に対しては解答しないようにする必要がある．本研究では，正例のみのデータセットである JSQuAD から，負例を加えたデータセットを 3 種類作成した．

4.3.1　swap データセットの作成

負例の作成の流れを図 6 に示す．負例の作成方法として，まず JSQuAD から同じタイトル，同じ答えで異なる段落の質問文を 2 文抽出する．そして，この 2 問の質問文を入れ替えて，負例としてデータセットに加える．質問文は該当する段落から解答できるようになっていて，該当しない段落から解答できない可能性が高いと考えられるため，入れ替え後の質問文は負例となり得る．ここで作成されたデータセットを swap データセットと呼ぶ．swap データセットの作成の際，入れ替え後の質問文が実際には正しい質問である場合も許容する．

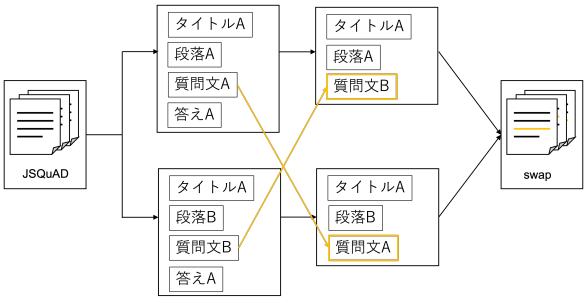

図 6　swap 負例データセット作成

4.3.2　JSQuAD+T5 負例データセットの作成

負例の作成の流れを図 7 に示す．負例の作成方法として，まず JSQuAD から文章と答えを抽出して，質問生成モデルに入

力し，質問文を n 問生成する．そして，n 問それぞれの生成確信度の符号を反転した値を重みとしてランダムサンプリングして，負例とする．生成確信度は負の値をとるため，符号を反転させることで，生成確信度が低い質問ほど大きい重み付けがされる．JSQuAD+T5 負例データセットの作成の際，サンプリングされた質問文が実際には正しい質問である場合も許容する．

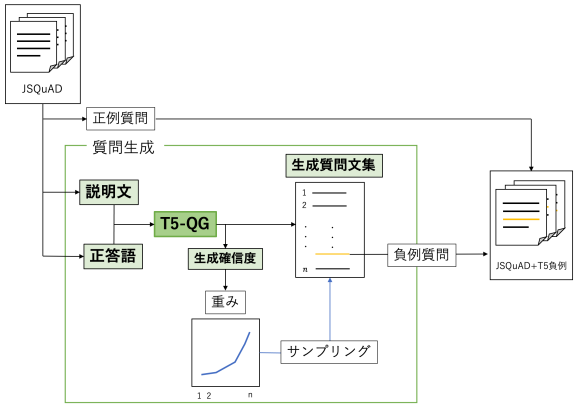

図 7　JSQuAD+T5 負例データセット作成

以上の手法で，JSQuAD に負例を加えて，JSQuAD+T5 負例データセットを作成した．生成した負例質問文の例を表 1 に示す．

表 1　T5 負例質問文の例

説明文	正答語	T5 負例質問文
造語 [SEP] 造語（ぞうご）は、新たに語（単語）を造ることや、既存の語を組み合わせて新たな意味の語を造ること、また、そうして造られた語である。新たに造られた語については、新語または新造語とも呼ばれる。	造語	「新語」という言葉は英語で何を意味しますか
通称 [SEP] 通称（つうしょう）は、正式な名称ではないが、特定の人や物、事象に対する呼び名として世間一般において通用している語のことである。別名（べつめい）とも、俗称（ぞくしょう）ともいう。	つうしょう	通常、世間一般で一般的に使われている通称は何ですか

4.3.3　swap+T5 負例データセットの作成

swap データセットに，JSQuAD+T5 負例データセットで加えた負例を全て加えて swap+T5 負例データセットを作成した．

5　実　　験

5.1　実験内容

5.1.1　質問生成

山川出版社の日本史用語集 [4] と倫理用語集 [5] に含まれる見出し語から，ランダムに単語を 91 単語をサンプリングし，それぞれを正答語とした．次に，Wikipedia API を利用して，それぞれの正答語を検索クエリとして Wikipedia の記事を取得し，概要にあたる冒頭の 1 段落を説明文とした．正答語と記事の概要を 1 セットとして，T5 日本語質問生成モデルに入力し，生成質問文を 1 セットにつき 10 件生成した．すなわち，全生成質問文は 910 件である．

5.1.2　アンケート作成のための質問解答

910 件の生成質問文を JSQuAD でファインチューニングした質問応答モデルに解答させた．検証のため，1 セットの全ての質問文の解答語が正答語と一致したもの，および全ての質問文の解答語が正答語と一致しなかったものを除き，クラウドワーカーの負担を減らすため，説明文の長さが比較的短いもの中から，12 セットの生成質問文をアンケート対象質問文とした．

5.1.3　主観評価

クラウドソーシングサービス [6] を利用して，アンケート対象質問文の主観評価を行い，29 人の有効な回答を得た．アンケートでは，説明文，生成質問文，正答語を提示し，生成質問文に対して適切，不適切いずれかを選択させた．選択の基準として，説明文を読んで，質問文に答えるとき，正答語が答えになる場合は適切を選択し，説明文を読んでも答えられない，正答語が答えにならない，質問文の意味が分からない場合は不適切を選択するように指示した．なお，質問文の答えの候補が複数あるときでも，その候補に正答語が含まれている場合は，適切を選択するように指示した．

5.1.4　質問応答モデル

東北大学乾研究室が構築した，訓練済み日本語 BERT モデル [7] をファインチューニングして，BERT 日本語質問応答モデルを開発した．ファインチューニングに用いたデータセットは，JSQuAD，swap，JSQuAD+T5 負例，swap+T5 負例で，これらでファインチューニングしたモデルをそれぞれベースライン，提案手法-swap，提案手法-T5，提案手法-swap-T5 とする．JSQuAD+T5 負例データセットの作成時は $n = 30$ として，JSQuAD の説明文と正答語を入力として，30 番目までに生成された質問からサンプリングして負例質問文とした．そして，このときの負例質問文を swap に加えて，swap+T5 負例データセットを作成した．図 8 に負例質問文として加えられた質問文の生成順位とその数の関係を示す．なお，JSQuAD はテストデータが公開されていないため，検証データを分割してテストデータを作成した．表 2 にデータセットの内訳を示す．表 3 にベースラインと提案手法のテストデータに対する正解率を示す．

4：日本史用語集改訂版 A・B 共用
5：倫理用語集第 2 版
6：クラウドワークス https://crowdworks.jp/
7：https://github.com/cl-tohoku/bert-japanese

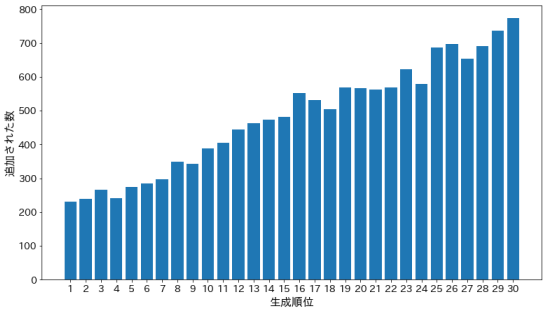

図 8　負例質問として加えた質問文の生成順位と数

表 2　データセットの内訳

	JSQuAD	swap	JSQuAD +T5 負例	swap+T5 負例
訓練データ				
総質問数	62859	62859	77330	77330
正例質問数	62859	50352	62859	50352
負例質問数	0	12507	14471	26998
検証データ				
総質問数	2257	2257	2782	2782
正例質問数	2257	1834	2257	1834
負例質問数	0	423	525	948
テストデータ				
総質問数	2185	2185	2686	2686
正例質問数	2185	1719	2185	1719
負例質問数	0	466	501	967

表 3　テストデータに対する正解率

	全体	正例のみ	負例のみ
ベースライン	86.77		
提案手法-swap	84.30	83.36	87.77
提案手法-T5	86.04	83.75	96.01
提案手法-swap-T5	85.96	81.21	94.42

5.2　実験結果

アンケートの結果から有効回答者数の k 人以上が適切と評価した質問文を真に正しい質問文，適切と答えたのが k 人未満であった質問文を真に正しくない質問文とした．表 4 にそれぞれの k における正しい質問文の数と正しくない質問文の数を示す．ここで，ある質問文を適切と評価したクラウドワーカーの人数をその質問文の**適切度**と定義する．

表 4　各 k のときの真に正しい/正しくない質問文数

	$k=15$	$k=20$	$k=25$
真に正しい質問文数	77	66	37
真に正しくない質問文数	43	54	83

図 9 に各適切度の質問文数を示す．この図を見ると，適切度が 20 代の質問文が全体と比べてやや多くなっているが，適切度が小さい質問文も一定数存在しており，質問文の適切度には散らばりがあることが分かる．

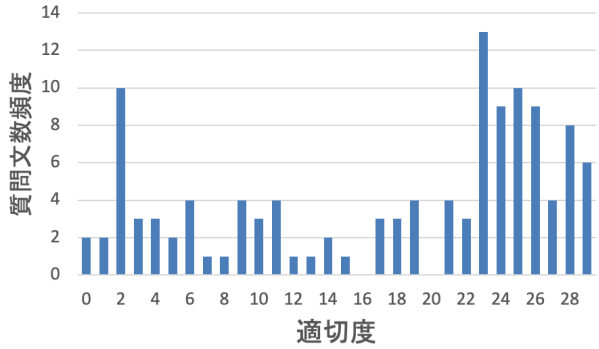

図 9　各適切度の質問文数

4.2 節で述べた方法で質問文のスコアを計算し，ベースラインと提案手法のそれぞれで生成質問文をランキングした．そして，$k=15$，$k=20$，$k=25$ それぞれのときで，Precision@r, Recall@r を計算した．ここで r は順位を表し，Precision@r はランキングした質問の上位 r 番目までを適切な質問文と評価した時の適合率を表し，Recall@r はランキングした質問の上位 r 番目までを適切な質問文と評価した時の再現率を表している．$k=15$，$k=20$，$k=25$ それぞれのときの PR 曲線を図 10，図 11，図 12 に示す．それぞれの図のベースラインと提案手法において，実線部分は正答語と解答語が一致したところで，点線部分は正答語と解答語が一致しなかったところである．

$k=15$，$k=20$，$k=25$ のいずれの場合においても，提案手法-swap-T5 の AUC がベースライン，提案手法-swap，提案手法-T5 の AUC を上回る結果になった．

ベースライン，提案手法-swap，提案手法-T5，提案手法-swap-T5 のそれぞれの生成質問文のランキングから，Average@r を計算した．Average@r は，ランキングの上位 r 番目までの質問文の適切度の平均である．ベースライン，提案手法-swap，提案手法-T5，提案手法-swap-T5 それぞれの Average@r を図 13 に示した．ほとんどの順位で提案手法-T5 や提案手法-swap-T5 の平均適切度はベースラインの平均適切度を上回る結果となった．

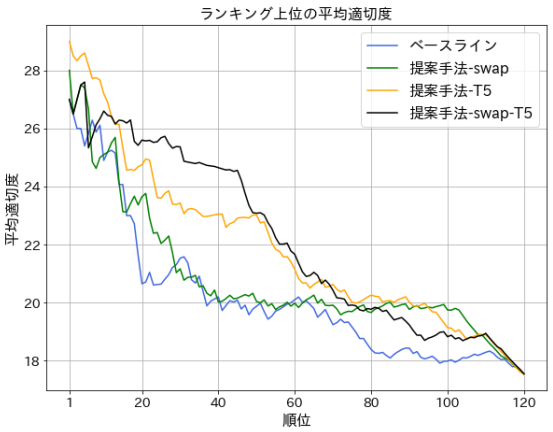

図 13　提案手法とベースラインの平均適切度

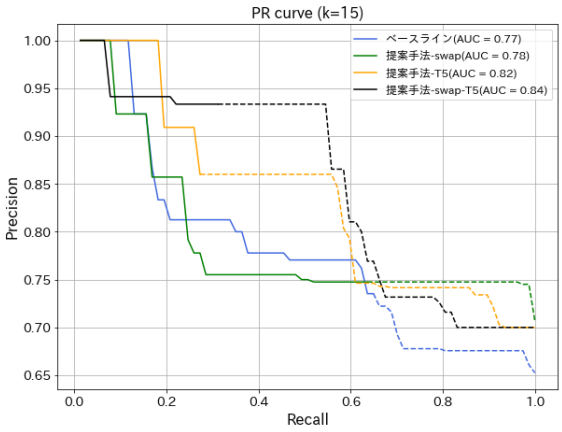

図 10　PR 曲線 ($k = 15$)

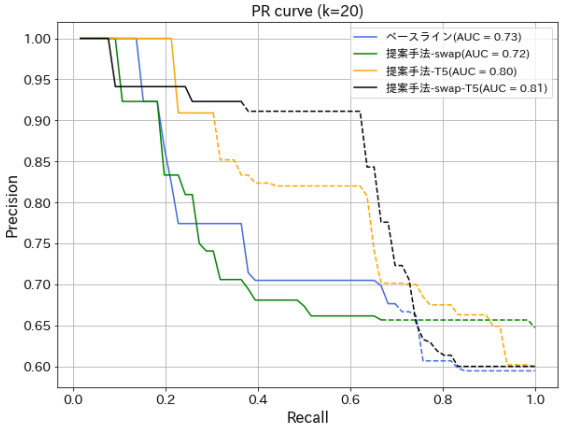

図 11　PR 曲線 ($k = 20$)

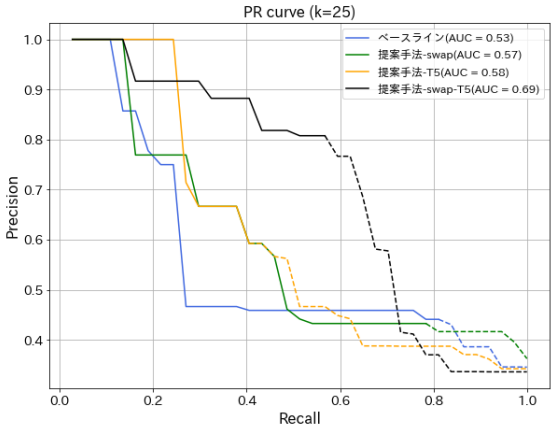

図 12　PR 曲線 ($k = 25$)

5.3　考　　察

5.3.1　質問評価の必要性

5.2 節の実験結果で示した表 4 を見ると，$k = 25$ を目安とすると真に正しくない質問文は全体の 7 割ほど生成されることが分かる．そのため，生成された質問文をそのまま学習者に提示するのではなく，質問文を評価して正しい質問文を抽出し，提示することが必要である．

5.3.2　高品質な質問文の抽出

図 13 から，提案手法-T5 や提案手法-swap-T5 はベースラインよりも適切度に基づいてランキングを行なっていることが分かる．適切度が高い質問文は多くの人に適切であると評価されているので，適切度が低い質問文に比べて高品質であると言える．そのため，この結果は提案手法-T5 や提案手法-swap-T5 がベースラインより高品質な質問文を抽出できていることを示している．

5.3.3　負例質問文生成

図 14 に，生成質問文の生成順位と各生成順位の平均適切度を示す．真に正しい質問と正しくない質問を分ける一つの境界であった 25 を平均適切度が上回る生成順位は存在しなかった．生成順位と平均適切度の相関係数は-0.68 であった．この相関係数から，生成確信度を用いて，生成された順番によってサンプリングする手法は妥当であったと考えられる．

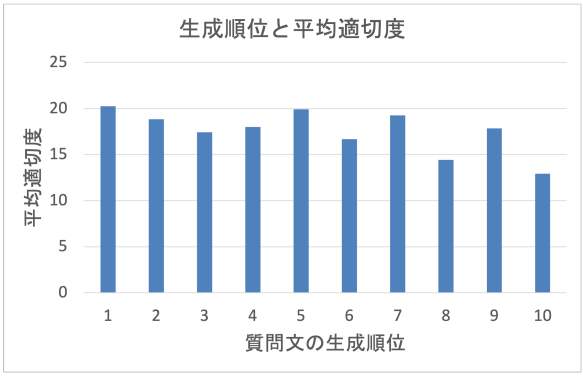

図 14　生成順位と平均適切度

真に正しくない質問文は，4.2 節の質問文が正しい条件から考えて，以下の 3 つのいずれかを満たすものである．

（1）　質問文は意味を捉えることができない

（2）　質問文は説明文を読んで解答することができない

（3）　質問文は正答語が答えとならない

このうち，質問応答モデルが正しくない質問文を見分ける際の判断材料となり得るのは (1) と (2) であるため，この 2 つについて考察する．swap により生成される負例は，元々人間が作成した質問文であることから，(2) の条件のみ該当すると考えることができる．T5 により生成される負例は (1)，(2) の両方の条件の少なくとも 1 つに該当すると考えられる．また，アンケート対象の質問文で正しくないと評価された質問文は (1) や (2) の両方の条件の少なくとも 1 つに該当する．そのため，提案手法-T5 は提案手法-swap より適切に質問を評価できたと考えられる．

今回の実験では，swap と T5 負例の 2 種類の負例を加えたデータセットでファインチューニングしたモデルは他のモデルと比較して，適切な質問評価を行うことができたという結果を得た．実験結果の表 3 を見ると，swap の負例は T5 の負例よりも見分けるのが難しいということが分かる．swap+T5 負例では，swap や JSQuAD+T5 負例よりも負例の多様性が広がったと考えられるため，より高度な学習を行えた可能性がある．

6 おわりに

本論文では，既存の質問生成モデルが出力する質問文を質問応答モデルを用いて評価することで，適切な質問文を抽出できることを示した．さらに，質問応答モデルのファインチューニングのデータセットに自動生成した負例を加えることで，より高品質な質問文を生成できることを示した．今回は，質問に対する解答が一意に定まらない質問でも正しい質問として実験を行ったが，解答の一意性は質問の品質の重要な要素であると考えられる．今後，解答の一意性を検証する手法の検討が必要である．また本研究では扱えなかった対話部分の機構，つまり学習者の発話を入力として，説明文を検索したり，正答語を決定したりする機構の提案も今後の課題となる．学習者の発話から話を深めたり，広げたりする方向性の質問生成を行うことができれば，従来のテスト形式の学習だけではなく，対話型の学習を実践できると考えられる．

謝　辞

本研究は JSPS 科研費 19H04219 の助成を受けたものです．

文　献

[1] 文部科学省. 令和元年度学校における教育の情報化の実態等に関する調査結果 (概要). pp. 8–10, 2020. https://www.mext.go.jp/content/20200909-mxt_jogai01-000009573_050.pdf,(2023.1.24 取得).

[2] 文部科学省. GIGA スクール構想の実現へ. 2020. https://www.mext.go.jp/content/20200625-mxt_syoto01-000003278_1.pdf,(2023.1.24 取得).

[3] 文部科学省. 高等学校学習指導要領平成 30 年告示. 東山書房, 2019.

[4] John F. Nestojko, Dung C. Bui, Kornell Nate, and Elizabeth Ligon Bjork. Expecting to teach enhances learning and organization of knowledge in free recall of text passages. *Memory —& Cognition*, Vol. 42, No. 7, pp. 1038–1048, 05 2014.

[5] ベネッセ教育総合研究所. 協働から個の思考を深める学習モデル実証研究レポート. （株）ベネッセホールディングス ベネッセ教育総合研究所, 2017.

[6] Noah L Schroeder, Olusola O Adesope, and Rachel Barouch Gilbert. How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research*, Vol. 49, No. 1, pp. 1–39, 2013.

[7] 田村吉宏, 山内崇資, 林佑樹, 中野有紀子. Wikipedia 記事情報に基づく歴史学習問題の自動生成手法. 人工知能学会全国大会論文集 第 28 回 (2014), pp. 3D42in–3D42in. 一般社団法人 人工知能学会, 2014.

[8] 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.

[9] Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3901–3910, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[10] Kettip Kriangchaivech and Artit Wangperawong. Question generation by transformers. *arXiv preprint arXiv:1909.05017*, 2019.

[11] Peide Zhu and Claudia Hauff. Evaluating BERT-based Rewards for Question Generation with Reinforcement Learning. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 261–270, 2021.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] 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.

[14] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, Vol. 21, No. 140, pp. 1–67, 2020.

[15] 栗原健太郎, 河原大輔, 柴田知秀. JGLUE: 日本語言語理解ベンチマーク. 自然言語処理, Vol. 29, No. 2, pp. 711–717, 2022.