深層学習を用いた難易度調整機能付き読解問題自動生成手法

Deep Learning Based Difficulty Controllable Automated Question Generation for Reading Comprehension Test

鈴木 彩香 *1，宇都雅輝 *1
Ayaka Suzuki*1, Masaki Uto*1
*1 電気通信大学
*1The University of Electro-Communications
Email: {suzuki_ayaka, uto}@ai.lab.uec.ac.jp

あらまし： 近年，任意の長文に関連する読解問題を深層学習を用いて自動生成する読解問題自動生成手法が注目されている．最先端手法では，与えられた長文と整合性がある自然な問題文を生成できるが，生成される問題の難易度は考慮できなかった．そこで本研究では，任意の難易度の読解問題を自動生成する手法を開発する．具体的には，Transformer ベースの事前学習済み深層学習モデルを用いた読解問題自動生成手法に対して，項目反応理論を利用して推定される問題難易度を組み込んだ入力データを与えることで，所望の難易度に合わせた問題を生成できる技術を提案する．

キーワード： 読解問題，問題生成，深層学習，言語モデル，項目反応理論，言語生成

1 はじめに

読解問題自動生成とは，与えられた長文からそれに関連する問題を自動生成する技術であり，教育分野において読解力を育成・評価するアプローチの一つとして活用が期待されている．読解問題自動生成手法として，従来は人手で設計したテンプレートを利用するルールベースの手法が主流であったが，近年では深層学習を用いた手法が多数提案されている [1]．最先端の深層学習ベースの手法は，人手でのテンプレート作成を行うことなく，柔軟で高品質な問題生成を実現している．

一方，既存の問題自動生成手法では，読解対象の長文（以降では「読解対象文」と呼ぶ）と整合性があり，文法的に正しい問題を生成することを目標としており，生成される問題の難易度などの特性は考慮されていない．しかし，読解力の効率的な育成支援のためには，学生の読解力のレベルに合わせた問題生成が必要と考えられる．

そこで，本研究では，任意の難易度の問題を自動生成する手法を提案する．具体的には，Transformer ベースの事前学習済み深層学習モデルを用いた読解問題自動生成手法に対して，項目反応理論（Item response theory：IRT）を利用して推定される問題難易度を組み込んだ入力データを与えることで，所望の難易度に合わせた問題を生成することを目指す．

2 提案手法

2.1 難易度を含んだデータセットの作成

提案手法では，問題の文章情報に加えて，それらの問題の難易度も訓練データとして使用する．各問題の難易度は，IRT を用いて以下の手順で推定する．

1. **各問題に対する正誤反応データの収集：**データ中の各問題を実際に出題して正誤反応データを収集する．ただし，本研究では人間の解答者を QA（Question Answering）システムで代用する．
2. **IRT を用いた難易度推定：**最も単純な IRT モデルである次式のラッシュモデルを利用して，正誤反応データから各問題の難易度を推定する．

$$p = \frac{\exp(\theta - b)}{1 + \exp(\theta - b)} \tag{1}$$

ここで，b は問題難易度，θ は解答者の能力値を表すパラメータである．

3. **推定された難易度を含んだデータセットの作成：**IRT で推定された難易度を用いて，提案手法の訓練データセット $\boldsymbol{C}$ を作成する．データセット $\boldsymbol{C}$ は，読解対象文 $\boldsymbol{w}$，答え単語列 $\boldsymbol{a}$，問題文 $\boldsymbol{q}$，難易度値 b の集合となる．ここで，$\boldsymbol{w}$，$\boldsymbol{a}$，$\boldsymbol{q}$ は単語の系列として，$\boldsymbol{w} = \{w_n \,|\, n \in \{1, \cdots, N\}\}$，$\boldsymbol{a} = \{a_m \,|\, m \in \{1, \cdots, M\}\}$，$\boldsymbol{q} = \{q_o \,|\, o \in \{1, \cdots, O\}\}$ とし，N，M，O はそれぞれの単語数，w_n，a_m，q_o はそれぞれの添字に対応する位置の単語を表す．

このデータセット $\boldsymbol{C}$ を用いて，提案手法では，1）読解対象文と指定した難易度から答えを抽出するモデルと，2）抽出された答えと読解対象文，および指定した難易度から問題を生成するモデル，の 2 段階モデルで問題生成を実現する．以降で各モデルの詳細を説明する．

2.2 難易度調整可能な答え抽出モデル

難易度調整可能な答え抽出モデルでは基礎モデルに BERT（Bidirectional Encoder Representations from Transformers）を用いる．BERT は，Transformer ベースの深層学習モデルを，33 億個以上の単語を含むデータセットで事前学習したモデルである．文書の分類や回帰タスクをはじめとして，系列ラベリングや抽出型文章要約のような文章からの要素抽出タスクにも広く利用されている [2]．そこで，本研究では，BERT を読解対象文から答えを抽出する基礎モデルとして利用し，それを問題の難易度を調整できるように拡張する．

具体的には，読解対象文 $\boldsymbol{w}$ と難易度 b を特殊トークンで連結したデータ（[CLS] b [SEP] $\boldsymbol{w}$）を入力，読解

対象文 $\boldsymbol{w}$ における答え単語列 $\boldsymbol{a}$ の開始位置と終了位置を出力とする学習データを用いて BERT モデルをファインチューニングすることで，読解対象文と難易度が与えられた時に答えを抽出できるモデルを得る．

2.3　難易度調整可能な問題生成モデル

難易度調整可能な問題生成モデルでは基礎モデルに GPT-2（Generative Pre-trained Transformer 2）を用いる．GPT-2 は，15 億以上のパラメータを持つ Transformer ベースの深層学習モデルを，800 万以上の文書データで教師なし学習することにより柔軟な文章生成を可能にした事前学習言語モデルである．問題生成タスクを含む様々な文章生成タスクで広く利用されている．本研究では，GPT-2 を用いた問題自動生成手法 [3] を，問題の難易度を調整できるように拡張する．

具体的には，読解対象文 $\boldsymbol{w}$ と答え単語列 $\boldsymbol{a}$，問題文 $\boldsymbol{q}$，難易度値 b を特殊トークンで連結した以下のデータを学習に用いる．

$$b <\mathrm{QU}> \boldsymbol{w}_1 <\mathrm{AN}> \boldsymbol{a} <\mathrm{AN}> \boldsymbol{w}_2 <\mathrm{G}> \boldsymbol{q} \quad (2)$$

ここで，<QU>，<G>，<AN>はそれぞれ読解対象文，問題文，答えの開始を表す特殊トークンである．また，$\boldsymbol{w}_1$ は答え以前の，$\boldsymbol{w}_2$ は答え以後の単語列を表す．予測の際には，上記のデータで訓練された GPT-2 モデルに<G>以前までの入力を与えることで，指定した難易度・読解対象文・答えに対応する問題文が生成できる．

3　提案手法の有効性評価実験

提案手法の有効性を評価するために次の手順で実験を行った．1）質問応答・問題生成タスクで広く利用される SQuAD データセットの訓練データを用いて，精度の異なる 5 つの QA システムを構築した．2）5 つの QA システムに SQuAD のテストデータ中の各問題を解答させ，正誤反応データを収集した．3）得られた正誤反応データを用いて，式 (1) のラッシュモデルで各問題の難易度を推定した．得られた難易度値は，-3.96，-1.82，-0.26，0.88，2.00，3.60 のいずれかとなり，値が小さいほど簡単な問題であることを意味する．4）得られた難易度値と SQuAD のテストデータの情報を使用し，提案手法で使用するデータを作成した．5）SQuAD の訓練データを用いて BERT と GPT-2 を難易度を考慮せずにファインチューニングしたのち，手順 4 で作成した提案手法のための訓練データを 90% と 10% に分割し，90% のデータで難易度を考慮したファインチューニングを行なった．6）残り 10% のデータを用いて所望の難易度に応じた出力が行えたかを「抽出された答えの難易度別平均単語数」と「生成された問題の難易度別正答率」の 2 つの観点で評価した．なお，正答率の評価には 2 つの QA システムを使用し，先行研究と同様に 2 つの QA システムが共に正解した問題のみを正答として扱った．

生成された問題の難易度別正答率を図 1 に示す．また，以前の著者らの研究 [4] では，答えを抽出型ではなく GPT-2 による生成型で行い，上記手順 5）の前半で述べた 1 度目のファインチューニングも行っていなかったため，これらの影響分析のために以前の結果を点線で示した．図より，どちらの結果も難易度が高いほど，生成された問題の正答率が減少する傾向が確認できる．また，先行研究 [4] に比べて本手法では正答率が大幅に向上しているが，これは，答え生成を抽出型にしたことで読解対象文との対応が明瞭になり，さらに 1 度目のファインチューニングにより，より適切な問題が生成できたためと考えられる．次に，抽出された答えの難易度別単語数を図 2 に示す．図より，難易度が高いほど，抽出された答えの平均単語数が増加する傾向が確認できる．また，先行研究 [4] では，文章内に含まれない答えが多数生成されていたが，本手法では，そのような不適切な生成を回避できた．以上から提案手法は，指定した難易度を適切に反映した答えや問題を出力していることがわかる．

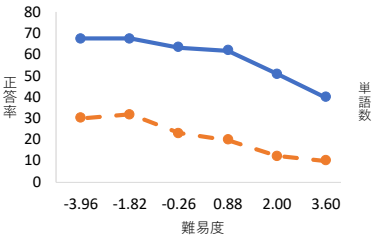

図 1：各難易度の正答率

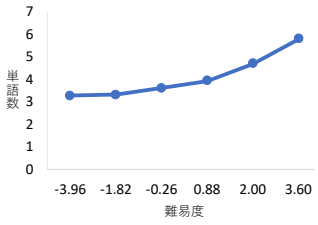

図 2：抽出された答えの単語数

表 1：生成された問題と答えの例

難易度	-3.96
問題	Where is much of the work of the Scottish Parliament done?
答え	committee
難易度	3.60
問題	What is the purpose of the chairman and member of the committee?
答え	take evidence from witnesses, conduct inquiries and scrutinise legislation

表 1 に出力された問題と答えの例を示す．表から，低い難易度の場合には単一の用語を答えとする簡単な問題が生成されたのに対し，高い難易度の場合には長めの用語を答えとする難しい問題が生成されたことがわかる．

4　まとめと今後の課題

本研究では，任意の難易度の問題を自動生成する手法を提案し，実験から提案手法の有効性を示した．今後は，QA システムではなく人間を対象にしたデータ収集と評価実験を行なっていきたい．

参考文献

[1] X. Du, J. Shao, and C. Cardie. Learning to ask: Neural question generation for reading comprehension. In *Proc. Annual Meeting of the Association for Computational Linguistics*, pp. 1342–1352, 2017.

[2] A. Srikanth, A. Shankar Umasankar, S. Thanu, and S. Jaya Nirmala. Extractive text summarization using dynamic clustering and co-reference on bert. In *In Proc. International Conference on Computing, Communication and Security*, pp. 1–5, 2020.

[3] M. Srivastava and N. Goodman. Question generation for adaptive education. In *In Proc. Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*, pp. 692–701, 2021.

[4] 鈴木彩香, 宇都雅輝. 難易度調整機能を持つ gpt-2 に基づく読解問題自動生成手法. 教育システム情報学会学生研究発表会, 2022.