

# NIE ワークシート自動生成に向けた Web ニュースからの問題文生成手法の検討

國方 海秀<sup>†</sup> 安藤 一秋<sup>‡</sup>香川大学院創発科学研究科<sup>†</sup> 香川大学創造工学部<sup>‡</sup>

## 1. はじめに

近年、小学校や中学校などの義務教育機関では、新聞を教育に活用する取り組み（NIE：Newspaper In Education）[1]が行われている。日本新聞協会や経済開発協力機構の調査結果[2]から、NIE は児童の読解力や学力の向上に効果的であることが示されている。しかし、NIE に関する教材準備やカリキュラムの調整が難しいなど、教師の負担が大きくなる実践課題がある。

NIE の実践を支援するため、新聞記事とそれに関する問題で構成される NIE ワークシートを配信している新聞社がある。しかし、全国で 23 社しか配信しておらず（2024 年 1 月時点）、配信していない地方新聞社も多い。また、教師が取り扱いたいテーマが存在しない場合は教師が独自に NIE ワークシートを作る必要があり、NIE 実践の支援には不十分である。そこで、本研究では NIE ワークシートの自動生成を最終目標とする。本稿では、その構成要素である問題文の自動生成手法として、構造がシンプルな問題文はテンプレートをを用いて生成、複雑な問題文は言語モデルを用いて生成する手法を検討する。なお、本研究で対象とする児童は小学校高学年とする。

## 2. 採用したテンプレート

本研究では、単調な問題文に対してはテンプレートをを用いて生成する。NIE で出題される問題文の傾向を把握するため、我々の先行研究[3]において、NIE ワークシートに採用された問題文の分析を進めてきた。その結果に基づき、疑問詞を文末に用いた基本的なテンプレート 6 種（「～は何ですか」「～は誰ですか」「～はどこですか」「～はいつですか」「～はなぜですか」「～はどのくらいですか」）に加えて、頻出構造に基づくテンプレート 2 種（「～について調べてみよう」「～はどんな N ですか」）を採用することにした。なお、各テンプレートの「～」部分には、記事内の単語・フレーズを挿入して問題文を生成する。

## 3. テンプレート型の問題文生成に向けた分析

まず、採用したテンプレートに対して、記事中のどのような単語・フレーズが挿入される傾向にあるかを確認するため、テンプレートに合致する問題文を分析する。対象とする問題文は全国の新聞社が提供する NIE ワークシートから収集した全 708 問の問題のうち、テンプレートに合致する問題文 109 問とし、対象文の～部分にどのような単語もしくはフレーズが挿入されているかを調査する。調査の結果、単語が挿入されている問題が 26 件（23.85%）、フレーズが挿入されている問題が 83 件（76.15%）あり、多くの問題においてフレーズが挿入さ

れる傾向を確認した。テンプレート別に見ると「～はどんな N ですか」のみ単語が挿入される割合が高くなった。また「～は何ですか」、「～について調べてみよう」では、フレーズが入る割合が高いが、他のテンプレートと比較すると単語が挿入される傾向があった。その他のテンプレートについては、単語が挿入される件数が 1 件以下となり、フレーズが挿入される傾向が見られた。

次に、記事中の単語・フレーズを挿入する際、変換の必要性を確認するため、テンプレートに合致する問題文を調査する。調査の結果、変換が必要な問題文のうち、単語が挿入されている問題文は全体の 1.8%、フレーズが挿入されている問題文は全体の 68.8%存在した。この結果から、単語を挿入する場合には変換は不要であるが、フレーズを挿入する場合には変換が必要であることを確認した。

続いて、挿入される単語やフレーズには固有表現が含まれる傾向があると考え、テンプレートに合致する問題文に対し、固有表現を含む問題文の件数を調査する。調査の結果、全体の 55.0%が固有表現を含んでいた。テンプレート別の内訳は「～はどこですか」のみ 30.8%と低い割合となったが、その他のテンプレートでは 50.0%以上となり、高い割合で固有表現を含む単語・フレーズが挿入されていた。「～はどこですか」のみ割合が低くなった理由としては、解答が地名つまり固有表現になることが多いことから、問題文には含まれにくい結果になったと考えられる。

また、「～はどんな N ですか」には、～部分と N 部分の 2 つの挿入部が存在する。該当する問題文を分析した結果、N には～部分の上位語となる語が挿入されることを確認した。

## 4. テンプレート型の問題文生成と簡易実験

前節の分析結果を踏まえて、テンプレートで問題文を生成する手法と簡易実験について述べる

問題文は、記事内の重要性を考慮して生成する必要がある。そこで、記事の要約に用いられる文は重要度が高いと考え、記事を要約した文にテンプレートを適用して、問題文を生成する。

まず、「～は誰ですか」、「～はどこですか」、「～はいつですか」の 3 テンプレートは記事中にそれぞれ人名、地名、時間を含む必要がある。そのため、それらの固有表現を含む文から、人名・地名・時間を削除してテンプレートに挿入する。これらのテンプレートで問題文を生成した結果、「～はどこですか」においては自然な問題文（例：全国で激減している赤トンボの代表種「アキアカネ」の自然羽化が確認されたのはどこですか）が生成された。しかし、「～はだれですか」、「～はいつですか」においては、問題と関係ないフレーズや固有表現を修飾する語が残り、不自然な生成が多く見られた。これらはフレーズを挿入しているため、対象を変換することで改善できる。

A Method for Generating Question Sentences from Web News for Automatic Generation of NIE Worksheets

<sup>†</sup> Kaishu Kunikata, Graduate School of Science for Creative Emergence, Kagawa University

<sup>‡</sup> Kazuaki Ando, Faculty of Engineering and Design, Kagawa University

また、「～は何ですか」、「～について調べてみよう」、「～はどんな N ですか」については、単語が挿入される割合が高い。そこで、固有表現を挿入して問題文を生成する。また、「～はどんな N ですか」における N 部分には RIKEN AIP 考案の拡張固有表現[4]のラベル名を挿入する。これらのテンプレートで問題文を生成した結果、「～は何ですか」、「～について調べてみよう」では自然な問題文（例：アマビエは何ですか、アマビエについて調べてみよう）が多く生成でき、固有表現を挿入する有用性を確認できた。しかし、「～はどんな N ですか」は、N 部分に挿入される単語が日常的に使用する語ではないため、違和感のある問題文（例：ツキノワグマはどんな哺乳類ですか）の生成が確認できた。よって、N に該当する上位語を探る手法の検討が必要である。

続いて、「～はなぜですか」は、接続助詞「から」に着目する。このテンプレートを利用した生成では、原因と結果を記事中から抽出する必要がある。原因と結果が別の文で表現されている場合、その因果関係を判定することは困難であるため、本実験においては、「から」が文頭や文末に出現しないパターンに限定して生成する。このパターンでは、「から」以前が解答である原因となり、「から」以降が結果となるため、「から」以降のフレーズをテンプレートに挿入する。なお、要約文には接続助詞「から」が出現しなかったため、このテンプレートは記事本文に適用する。このテンプレートで生成した結果、指示語を含む問題文（例：この作品はどの時代でも響くのはなぜですか）が生成され、問題文のみでは意図が掴みづらい結果が確認できた。この課題は、言語モデルを利用した因果関係抽出により改善できると考える。

最後に「～はどのくらいですか」は、寸法表現を問う問題であるため、記事中の寸法表現に着目した。しかし寸法表現には多様性があり、挿入すべきフレーズの特徴を発見できなかった。このテンプレートにおいては言語モデルと組み合わせた生成を検討する。

本節では、テンプレートを用いた問題文生成の簡易実験を実施した。「～はどこですか」、「～は何ですか」、「～について調べてみよう」テンプレートでは、NIE で活用できる問題文が生成できることを確認した。それ以外のテンプレートにおいては、係り受けや因果関係の判定に課題があり、テンプレートによる生成の限界を確認した。今後は、言語モデルを用いた生成手法と組み合わせる、記事内のフレーズを変換する手法などを検討する。

## 5. 言語モデルを用いた生成手法と簡易実験

構造が複雑な問題文は言語モデルを用いて生成する。本稿では、事前学習済みモデルとして GPT-2 を採用し、ファインチューニングすることで問題文を生成する。ファインチューニングには、静岡新聞社の NIE ワークシートに含まれる記事と問題文、その解答を用いる。ワークシート内の記事は画像であるため、同社の Web サイトでニュース記事を収集して、データセットを構築する。最終的に、700 件の Web ニュース記事とそれに付随する 2,274 件の問題および解答のデータセットを構築した。先行研究[5]を参考に、記事と問題、解答はトークンで連結して学習し、生成時には Web ニュース記事のみを与えて問題文を生成する。

本研究では、児童にそのまま出題するための問題文生成ではなく、教師が問題文を検討するためのサンプル問題の提示を想定している。そこで、簡易実験においては、

生成された問題文の語彙の多様性を測るため、distinct-N を用いて評価する。データセットに含まれていない NIE ワークシートの記事 5 件を用いて問題文を生成し、NIE ワークシートの問題文と生成した問題文の distinct-N を比較する。また、生成文を人手確認して主観的に評価する。

5 記事の平均を表 1 に示す。表 1 より、NIE ワークシートに採用される問題文は全体的に値が高く、多様性があることを確認した。生成文においても一定の多様性が確認できたが、実際の問題文より 0.1 程度低い値となった。生成した問題文の distinct-N が最も低い値となった記事は「今年の漢字」に関する記事である。これは静岡新聞でよく取り上げられる内容であり、類似した問題がデータセット内に複数存在したため、語彙の多様性が損なわれたと考える。

また、生成された全ての問題文を確認したところ、入力記事には含まれない内容を問う問題も確認できた。しかし、文中の一単語を変換する、児童に調査を促す一文を加えるなどの修正を前提とすれば、十分に活用できると考える。

今後は、データセットを下処理することで多様性の向上を図るほか、distinct-N 以外で評価する。

表 1 distinct-N の平均値

NIE ワークシートの問題文			生成した問題文		
N=1	N=2	N=3	N=1	N=2	N=3
0.636	0.907	0.972	0.534	0.747	0.843

## 6. おわりに

本稿では、NIE ワークシートの自動生成に向け、テンプレート形式と言語モデルを用いた問題文の生成手法を検討した。まず、テンプレートによる生成においては、「～はどこですか」、「～は何ですか」、「～について調べてみよう」では自然な問題文が多く生成できたが、それ以外のテンプレートでは違和感のある問題文が多く生成された。これらのテンプレートについては、変換した後に挿入する、言語モデルによる生成と組み合わせる等により、改善する必要がある。また、言語モデルを用いた生成手法として、GPT-2 をファインチューニングしたモデルを構築して生成した結果、生成された問題文には多様性が確認できたものの、改善の余地があることが確認できた。さらに主観評価の結果、問題の一部は修正が必要であるものの、問題の方向性や出題の意図を定める観点では活用可能であると考えられる。

今後は、本稿で提案した 2 手法に対する詳細を評価した後、テンプレートと言語モデルを組み合わせる手法を検討する。最後に NIE ワークシートの自動生成を目指す。

## 参考文献

- [1] NIE 教育に新聞を, <https://nie.jp/>
- [2] 2009 年度「NIE 効果測定調査」結果報告, [https://nie.jp/research/pdf/re5\\_201007.pdf](https://nie.jp/research/pdf/re5_201007.pdf)
- [3] 國方他, “小学校での NIE 支援に向けた新聞記事からのテンプレートによる問題文生成の検討”, 情報処理学会第 85 回全国大会講演論文集, 2023.
- [4] 拡張固有表現 – RIKEN AIP, <http://liat-aip.sakura.ne.jp/wp/拡張固有表現-2/>
- [5] 鈴木他, “Transformer モデルを用いた難易度調節機能付き読解問題自動生成手法”, 行動計量学会大会第 50 回大会抄録集, 2022.