

Wikipedia を用いた選択肢付き問題の生成手法

田村 吉宏[†] 高瀬 裕[‡] 中野 有紀子[‡]

成蹊大学 理工学研究科 理工学専攻[†]

成蹊大学 理工学部 情報科学科[‡]

1. はじめに

近年、情報技術の発達により、e ラーニングシステムに代表されるコンピュータなどの情報機器を用いて個別学習を支援するシステムが登場している。そうした中でコンピュータに教師のような役割を担わせる ITS(Intelligent Tutoring System) 研究も盛んに行われている[3, 4]。しかし、そのようなシステムで利用される知識ベースや、学習問題の多くは人手で作成されているため、構築に大きなコストがかかるという問題がある。

そこで我々は問題生成のための知識ベースとして利用でき[5]、学習にも利用できる知識が格納されていると考えられる Wikipedia に着目し、記事のタイトルが正解となる一問一答形式の日本史人物問題を生成することに成功した[1]。文法的には 85%を超える正しきで問題文を生成できたものの、解答が複数存在していたり、問題内容がマニアックすぎて普通なら誰も正解できなかったりする問題が多数生成されてしまった。

そこで本研究では、一問一答形式で生成された問題に誤選択肢を付与し、誤選択肢を操作することでこのような不都合にも対応できると考え、Wikipedia を LOD (Linked Open Data) として公開している DBpedia Japanese を利用して誤選択肢を生成する

2. DBpedia Japanese

本研究では、誤選択肢生成に DBpedia Japanese を用いる。DBpedia とは Wikipedia からの構造化コンテンツの生成を目的とするプロジェクトである。2011 年 09 の時点で、英語の記事を主に対象としている DBpedia のデータセットは 364 万件以上の事物を扱い、うち 183 万件が一貫したオントロジーに分類されている。本研究では日本史を対象としているため日本語記事を対象とした DBpedia Japanese を用いる。

Generating Multiple-choice Quizzes for History Learning based on Wikipedia Articles

[†]Yoshihiro Tamura, Graduate School of Science and Technology, Seikei University

[‡]Yutaka Takase and Yukiko Nakano, Dept. of Computer and Information Science, Seikei University

DBpedia ではリンク構造や使用しているテンプレート、インフォボックスをトリプル構造として構造化している。このトリプル構造を利用すれば似た記事の収集や記事同士の関係を抽出することも可能である。本研究では、この特徴を利用して、誤選択肢の生成を行う。

3. 歴史人物問題生成

Wikipedia の人物一覧ページをもとに Wikipedia のダンプデータを抽出し、「問題作成用データベース」を構築した[1]。本研究では、このデータベースを用いて問題生成を行う。問題生成には Wikipedia の編集用タグを削除した記事のアブストラクトを使用する。

3.1 問題生成アルゴリズム

Wikipedia のアブストラクトを疑問文にするためには平叙文を疑問文に変換する必要がある。疑問文を生成する際、日本語は「～ですか？」などの助辞「か」を含む助詞を語尾に付与すれば疑問文になることに着目した。自然な疑問文にするためには、文末の表現によって繋ぎの役目を果たす語句を間に挟む必要がある。そこで、Wikipedia 記事のアブストラクト部の文末表現を調査し、どのようなパターンがあるかを分類した結果、文末が

①助詞・助動詞

②サ変接続の体言

③サ変接続以外の体言

の3つとなる場合が多いことがわかり、

① 「のは誰ですか？」

② 「したのは誰ですか？」

③ 「は誰ですか？」

をそれぞれ文末に追加することで、自然な疑問文を生成する。また、文末表現を置換することで、「幼名は竹千代であるのは誰ですか？」の様に1つの分節に助詞の「は」を含む分節が2つ以上係る、もしくは助詞の「は」がある文節に他の助詞の「は」がある文節に係るといった、日本語として不自然な表現となってしまう場合がある。この問題を解決するために、日本語係

り受け解析エンジン Cabocha 用いて、助詞「は」が存在する場合、その助詞を「が」に変換する機能を追加した。

3.2 問題生成の課題

本方式では生成した問題文の内容理解までは行っていない。そのため

- ・三英傑の一人であるのは誰ですか？
- ・戦国時代の武将・戦国大名であるのは誰ですか？

のように複数答えが存在する問題も多数生成された。

また、

- ・幼名が竹千代であるのは誰ですか？

のように質問の内容が一般的でないため、一問一答形式の問題文には適さない文も生成されている。

4. 誤選択肢生成手法

前章であげた解答が複数存在するという問題に対する解決策として、他の正解を除去し、誤選択肢を付与した多肢選択問題とすることで、解答を一つにすることができる。この機能を実現するために DBpedia のトリプル構造を SPARQL で検索する。

4.1 誤選択肢生成アルゴリズム

誤選択肢生成のためにまず、問題作成の対象となる記事(以降、対象記事)の記事名が主語(S)となるプロパティ(P)とオブジェクト(O)の組(以降、P0 セット)を全て検索する。そのようにして収集した各 P0 セットについて、同じ P0 セットを持つ記事を検索する。この方法で検索された記事群は、対象記事と同じ P0 セットを持っているわけだから、類似した要素を有する可能性がある。従って、このような 記事は、誤選択肢となりうる。

さらに、P0 セットの組み合わせを使って検索することにより、この組み合わせ条件にマッチした記事がえられる。全ての P0 セットの組み合わせについて検索を実行し、誤選択肢を生成することが理想的である。しかし現実には、組み合わせが爆発し、サーバが処理不能になるため、以下の方法で、組み合わせ対象とする P0 セットを限定することにした。

- ある P0 セットによる検索結果の 80%以上が 3 章で述べた問題生成用データベースに登録されている場合(一致率による制約)
- 問題生成用データベースに登録されている記事の 30 パーセント以上が含まれていた P0 セット(網羅率による制約)

以上の条件を満たす P0 セットを誤選択肢生成用

P0 セットとする。最大 3 つの P0 セットの組み合わせにより検索を実行し、検索結果の記事が 2 つ以上 50 以下のものを誤選択肢候補とした。

次に生成された問題文の名詞、複合名詞を全て抽出し、誤選択肢生成用 P0 セット中の 0 の名詞に照合し、同じ名詞が見つかった場合は、正解である可能性があるため、誤選択肢として採用しない。こうすることで複数の解答が存在する問題文に解答以外の適切な誤選択肢を付与することが可能になった。

5. まとめ

本研究では自動生成した問題文に適切な誤選択肢を付与する方法を提案した。自動問題生成の問題点の一つである「解答が複数存在する」という問題点に対して適切な誤選択肢を付与できた。

今後の課題として、生成した誤選択肢とそこで用いた P0 セットの更なる利用である。まず必要なのは、生成された誤選択肢群の評価である。どの P0 セットを用いると教育に効果的かを評価する必要がある。更に P0 セットを分析することで、どのような属性において各記事が近いかを判別可能である。教育的な重要度と属性の 2 つの側面においてが判別できれば、教育利用に際し、さまざまな要素で学習者が間違いやすい誤選択肢を生成することができる。

参考文献

- [1] 田村吉宏：Wikipedia 記事情報に基づく歴史学習問題の自動生成手法，JSAI2014，3D4-2in(2014)．
- [2] 全国歴史教育研究協議会(編)：日本史 B 用語集 改訂版，山川出版社 (2009)．
- [3] 舟生日出男，亀山雅史，平嶋宗：問題解決プロセスを利用した選択問題の誤選択肢及び解説の自動生成，電子情報通信学会論文誌 D，J93-D(3)，pp. 292-302 (2010)．
- [4] 菅沼明：学生の理解度と問題の難易度を動的に評価する練習問題自動生成システム，情報処理学会論文誌，Vol. 46，No. 7，pp. 1810-1818 (2005)．
- [5] Ryuichiro Higashinaka, Kohji Dohsaka and Hideki Isozaki: Learning to Rank Definitions to Generate Quizzes for Interactive Information Presentation, In Proc. of the ACL 2007 Demo and Poster Sessions, pp.117-120 (2007)