3D4-2in

Wikipedia 記事情報に基づく歴史学習問題の自動生成手法

Generating Exercises for History Learning based on Wikipedia Articles

田村 吉宏[*1] Yoshihiro TAMURA　山内 崇資[*2] Takashi YAMAUCHI　林 佑樹[*1] Yuki HAYASHI　中野 有紀子[*1] Yukiko NAKANO

[*1] 成蹊大学理工学部 Faculty of Science and Technology, Seikei University

[*2] 成蹊大学大学院理工学研究科 Graduate School of Science and Technology, Seikei University

Since knowledge-base for ITS is manually developed in most cases, it requires high cost to create large-scale and multi-domain knowledge base. Aiming at reducing the cost for developing knowledge base for ITS, this study constructs a data base for Wikipedia articles about historical people and events, and proposes a method for generating history quizzes using the data base. Moreover, we propose a method for assigning categories and the level of importance in terms of history education based on Wikipedia article information and its link structure. Finally, an evaluation study showed that about 50% of the generated quizzes were educationally appropriate, and 85% of educationally important articles can be successfully chosen using the proposed method.

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学習に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[舟生 2010, 菅沼 2005]．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掛かるという問題点がある．

そこで我々は問題生成のための知識ベースの一つとして利用されており[Higashinaka 2007]，学習に利用できる知識が格納されていると考えられる Wikipedia に着目した．Wikipedia を知識ベースとし，そこから問題文及び，教育に用いる難易度などの情報を自動で付与することができれば，システム構築におけるコストを削減でき，個別学習における様々な分野の学習を知的に支援できる家庭教師エージェントの開発を行うことも可能である．本研究ではその基礎検討として，歴史上の人物や出来事に関する記事情報を収めたデータベースを作成し，一問一答形式の日本史人物学習問題を自動生成することを目的とする．また，作成された問題を学習に利用するために，記事情報や Wikipedia 固有の情報を利用して難易度やカテゴリの情報を付与する手法を提案する．

2. 日本史人物問題用データベース

2.1 Wikipedia

Wikipedia は世界最大規模のインターネット百科事典である．誰でも自由に編集できるのが特徴であり，2010 年 8 月時点で全言語を併せると 1600 万以上もの記事が存在している．査読機構などが存在せず，記事の情報の正確性は保証されていないものの，多数の編集者に編集された結果，情報の正確性は近年向上しており，学習教材として利用するのは問題ないと考える．

本研究では日本語の記事(約 136 万記事)を対象とし，その中から歴史に関連する記事を抽出したデータベースに基づき，歴史学習用の問題の自動生成を行う．

連絡先：田村吉宏，成蹊大学理工学部，東京都武蔵野市吉祥寺北町 3-3-1，yoshihiro.tamura@hotmail.co.jp

2.2 データベース構造

日本史に関係のある人物の記事を，各時代の人物一覧ページに記載されたリンクに基づき抽出し，記事情報に加えて記事の重要度や時代区分情報などを保持する歴史人物問題用データベースを作成した．

データベースの構築には MySQL を用いた．図 1 にデータベースのテーブル構造を示す．テーブル間を結ぶ矢印は外部参照を表す．今回対象とする日本史の時代区分は，鎌倉，室町，南北朝，戦後，安土桃山，江戸，幕末，明治とし，各記事の時代属性は人物がどの一覧ページに載っていたかで判定する．複数の時代にまたがって存在していた人物には 2 つ以上の時代情報が付与されることもある．また，歴史上の様々な人物が主体となる出来事のうち，最も多くの人物が関わる出来事として，「戦い」に関する記事も人物と同様に一覧ページから収集し，データベースに格納した．person テーブルには 3986 レコード格納され，battle テーブルには 500 レコード格納されている．表 1 にデータベースが保持する各テーブルの内容を示す．era，person，battle の各テーブルは関連テーブルで関係付けられている．

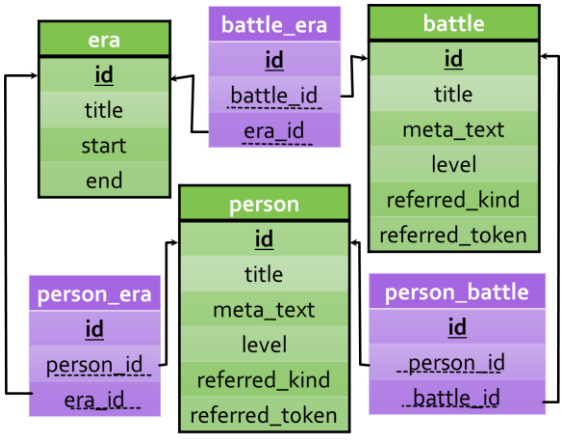

図 1. 日本史問題用データベースのテーブル構造

表 1. 各テーブルの内容

テーブル名	内容
era	時代区分情報
person	記事タイトル，記事内容(タグあり)，重要度，被引用数情報
battle	記事タイトル，記事内容(タグあり)，重要度，被引用数情報
person_era	person と era の関連テーブル
person_battle	person と battle の関連テーブル
battle_era	battle と era の関連テーブル

3. 人物問題文の生成手法

2 章で作成したデータベースを基に問題生成を行う. person テーブルに格納された meta_text を用いて問題生成を行う. meta_text は Wikipedia の編集用タグが残っているため，問題生成を行う際には削除する.

本文の中でも記事の先頭部分にあるアブストラクトは，人目に多く触れるため記事全体で最も正確性が高い部分であり，書き方にも一貫性が存在する[3]. また，書いてある情報も記事全体を概略している. そこで本研究では，問題文の雛型として記事のアブストラクト部分に着目し，タグ情報などを用いてアブストラクト部分を抽出し，記事タイトルの人物を問うための一問一答形式の問題文を生成する.

3.1 問題文生成アルゴリズム

問題生成にあたり，まず，今回解答となるタイトルの消去を行う. アブストラクトでは最初の文のみに記事のタイトルが含まれ，それに続く文の主語は省略されている. そこで最初の文からタイトルを削除できればタイトルが含まれない文を大量に生成することができる. タイトルの消去には Wikipedia 固有のタグを利用した正規表現を用いて削除を行った.

次に平叙文を疑問文に変換する. 疑問文を生成する際，日本語は「～ですか？」などの助辞「か」を含む助詞の集合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，それぞれに文末に

①「のは誰ですか？」
②「したのは誰ですか？」
③「は誰ですか？」

を追加することで，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」がある分節が 2 つ以上掛かる，もしくは助詞の「は」がある文節に他の助詞の「は」がある文節が掛かるといった，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[1]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. 以上の処理をまとめた問題生成フローチャートを図 2 に示す.

「徳川家康」の記事におけるアブストラクトを例に挙げて処理の流れを示す. 徳川家康のアブストラクト箇所の文を句点で分割すると以下の 7 つの文になる.

1. 江戸幕府の初代征夷大将軍
2. 三英傑の一人
3. 本姓は先に藤原氏、次いで源氏を称した
4. 家系は三河国の国人土豪・松平氏
5. 永禄 9 年 12 月 29 日に勅許を得て、徳川氏に改めた
6. 松平元信時代からの通称は次郎三郎
7. 幼名は竹千代

次に助詞の「は」を探索し，「は」が属している文節が最後の文末に係っていた場合は「が」に変換する. 今回の例では分割した文の 3, 4, 6, 7 番目の下線部に該当する「は」が変換対象となる. 最後に文末部分の形態素解析結果に応じて，上記①～③の変換パターンに合致する語句を補完する. 今回の例では，1, 2, 4, 6, 7 番目の文末が体言で終わっているため「は誰ですか？」を文末に，3, 5 番目の文末が助動詞で終わっているため「のは誰ですか？」をそれぞれ文末に付加する. 結果，生成された 1*～7*の問題文はどれも自然な疑問文になっていることがわかる.

1* 江戸幕府の初代征夷大将軍であるのは誰ですか？
2* 三英傑の一人であるのは誰ですか？
3* 本姓が先に藤原氏、次いで源氏を称したのは誰ですか？
4* 家系が三河国の国人土豪・松平氏であるのは誰ですか？
5* 永禄 9 年 12 月 29 日に勅許を得て、徳川氏に改めたのは誰ですか？
6* 松平元信時代からの通称が次郎三郎であるのは誰ですか？
7* 幼名が竹千代であるのは誰ですか？

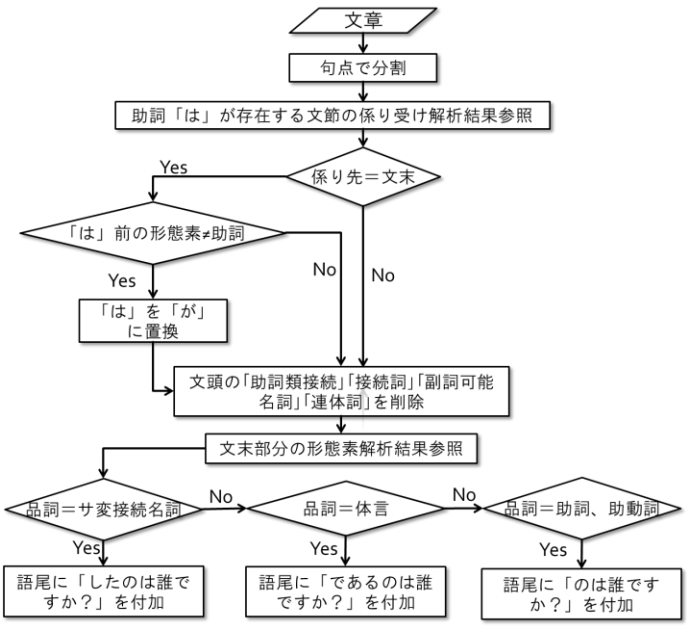

図 2. 人物問題生成のアルゴリズム

[1] CaboCha: Yet Another Japanese Dependency Structure Analyzer: https://code.google.com/p/cabocha/

4. 記事の重要度

4.1 Wikipedia 主要カテゴリ

記事の重要度を測定するために, Wikipedia カテゴリのリンク構造に着目した.「主要カテゴリ」と呼ばれる公式に設定された 9 つのカテゴリ(総記, 学問, 技術, 自然, 社会, 地理, 人間, 文化, 歴史)の歴史カテゴリから最も近い日本史関連のカテゴリは「日本の歴史」(第 5 階層:主要カテゴリ > 歴史 > 地域史 > 大陸別の歴史 > アジア史 > 日本の歴史)である. その「日本の歴史」カテゴリの中にある「日本の歴史」記事は, 同じ名前のカテゴリが主要カテゴリからもっとも近い日本史関連のカテゴリであることに加え,「日本史」と Wikipedia で検索した際にリダイレクトされるページであることから, この記事が Wikipedia 内で最も日本史全体をまとめた記事であるとした.

一般的に詳細な説明が載っている複数の記事をまとめて1つの概要記事を作ろうとしたとき, 各記事に載っている全ての情報をそのまま複写するのではなく, 情報を精査し, 重要な部分を優先的に記述しようとする. Wikipedia においても同様に記事を集約した記事には重要な情報しか残らないのではないかと考えられる. そこで, 本研究では「日本の歴史」ページにリンクが張られている人物を日本史における最重要人物であるとし,「日本の歴史」にリンクが直結している各時代説明記事にリンクされている人物は, その次に重要な人物であるとみなす. この情報は, 日本史人物問題用データベースの person テーブルの level に「日本の歴史」ページにある記事には 2 を, 各時代ページにある記事には 1 を, それ以外には 0 が格納される.

4.2 被参照リンク数

Wikipedia の階層構造をから得られた重要度(level)とは別に, ウェブ検索における重要な指標である被参照リンク数によりランク付けを行った. ここでは, 着目している人物の記事へのリンクが, 他の人物の記事内容からどれだけ参照されているかを算出した. 被参照数の計算方法には以下の二種類を考える.

1. 被参照リンクトークン数(referred_token)
2. 被参照リンク種類数(referred_kind)

どちらのほうが教育的に重要な記事をより上位にランク付けできるか不明瞭なため, 本研究では同記事から重複リンクをカウントした参照数(1)を person テーブルの referred_token に, 重複を 1 としてカウントした参照数(2) referred_kind に格納している. 記事重要度に関する評価は 5.2 節で行う.

5. 評価実験

5.1 人物問題文の評価

生成された人物問題文の妥当性を評価するために評価実験を行った. 評価用の問題文として, データベースから 100 個の人物記事を無作為に抽出し, 人物問題生成アルゴリズムを適用した. そして, 生成された問題文から更に 100 文を無作為に抽出した. 本実験では, (i)文法的に正しい問題であるか, (ii)学習問題として適しているか, という 2 つの評価基準を設け, 3 名の評価者を集い, 約 100 問ずつ評価させた.

評価結果を表 2 に示す. 評価(i)については 85%を超える評価を得られた. 不適切だと判断された問題文のほとんどは,「など」等の文末が副助詞のときに対するルール不足に代表される文末変換関係で発生しており, 改良の余地がまだあることを示している. 一方, 評価(ii)は, 半数以上の問題が教育的な観点から不適切だと判断されていた.「豊臣政権の五大老の一人であるのは誰ですか？」のように解答が複数存在する問題や,「幼名が竹千代であるのは誰ですか？」のような教育的に特に意味の無い問題が生成されてしまっており, 問題文を判定して評価する機構が無いため, そのまま出題してしまったことが不適切な問題と判定された問題を増やしてしまった主な要因である. 以上より, 概要部分に対する問題文生成手法は, 文法的には正しい質問を生成できるが, 教育的観点から正しい問題文を生成するためには, 問題内容を分析し, システムが歴史問題として適切かどうか判断しなければならないことが示された.

表 2. 人物問題の評価結果

	評価内容	結果
(i)	文法的な正しさ	86.5%(256/296)
(ii)	教育的な正しさ	48.0%(142/296)

5.2 記事重要度の検証

4.1 節で用意したカテゴリ構造を利用した記事重要度(level)と, 4.2 節で用意した二つの被参照リンク数(referred_kind, referred_token)を用いて重要度判定を行う. 被参照リンク数が多いもの上位にするランク付けと, level でまずランク付けを行い, level の高いものを上位にした後に同じ level 内でのみ被参照数を考慮する 2 手法をそれぞれ被参照リンク数で計算を行い計 4 つの結果を評価した. 検証方法は歴史教科書[山川日本史教科書 2009]に収録されている日本史 B を対象とした教科書 11 冊のうち, その単語が何冊に収録されているかを表す頻度数(1～11)を用いた. 掲載されていない単語は 0 とした. 頻度数が 5 以上である人物を歴史教育の上で重要な人物であると定義し, 前述の重要度検証方法 4 つそれぞれで抽出された上位 100 個のうちに何個が頻度数 5 以上の記事であったかをカウントし割合で評価した.

結果を表 3 に示す. 階層構造を用いず被参照数のみで重要度を測定した(Ⅰ)(Ⅲ)では, 約 65%以上の精度で教育的に重要な人物を抽出することができた. Wikipedia 階層構造も考慮すると, 非参照数のみの評価よりも約 10～15%程度抽出できた人物の割合が向上した.

表 3. 記事重要度検証の評価結果

	重要度測定手法	結果
(Ⅰ)	referred_kind	67%(67/100)
(Ⅱ)	referred_kind + level	84%(84/100)
(Ⅲ)	referred_token	74%(74/100)
(Ⅳ)	referred_token + level	85%(85/100)

6. まとめ

本研究では, Wikipedia を用いて歴史問題生成用のデータベースを構築し, Wikipedia 記事のアブストラクトから記事のタイトルが解答となる一問一答形式の人物問題の自動生成手法を提案した. 評価実験の結果, 約 5 割の精度で歴史学習に出題できる問題を自動生成することができた. 特に, 生成された問題文の 8 割以上は文法上正しい問題となっていたことを確認した.

また, Wikipedia の 2 つのリンク構造である, カテゴリによるリンク構造と関連記事からの被参照数で記事の重要度をランク付けし, ランキングの上位 100 を抽出して教科書に頻出する単語がどれくらい含まれているか分析することによって教育的に重要な記事を抽出できたかどうか評価した. その結果, 被参照数のみのランク付けでも 65%以上, 教育的に重要な単語が含まれて

いたが，階層構造も考慮することによって，10%~15%程度重要な記事が含まれている割合を増加させることに成功した．

今後の課題として，教育的に意義のある問題文を選別する手法の考案，並びに記事重要度の更なる妥当性の検討を行っていく予定である．

参考文献

[中山 2008] 中山浩太郎：自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築に関する一手法, DEWS2008, A3-2 (2008).

[山川日本史教科書 2009] 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).

[舟生 2010] 舟生日出男, 穐山雅史, 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).

[菅沼 2005] 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).

[Higashinaka 2007] 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).