1N2-2

Wikipediaを用いた質問応答と多肢選択問題による歴史学習

Question-answer and Multi-choice Quiz Generation using Wikipedia for History Learning

田村 吉宏[*1] Yoshihiro Tamura　鶴﨑泰斗[*2] Taito Tsurusaki　高瀬 裕[*2] Yutaka Takase　中野 有紀子[*2] Yukiko Nakano

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

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

With a goal of building an interactive educational system for history learning, this paper proposes a QA system and a multi-choice quiz generation mechanism. First, we created a database for historical persons and events from Wikipedia. Then, we proposed a response generation mechanism from the summary table of a Wikipedia article. We also proposed a method for providing additional information related to the user's question. Second, we proposed a mechanism for generating multi-choice quizzes using property-object pairs in DBpedia. We also proposed a method for detecting invalid choices in order to avoid creating a quiz that has multiple correct choices.

1. はじめに

近年の情報検索技術の発展に伴い，Wikipedia から情報を取り出し，ユーザからの質問に応答する情報案内システム[翠 2007] 等，文書検索を用いたインタラクティブなシステムの研究が盛んにおこなわれている．これらの技術を応用したユーザの質問に対話的に応答する質問応答システムの研究は，IBM の WATSON[1] をはじめ，大きな成果を挙げている．しかし，従来の質問応答システムではユーザの質問に答えることが目的であり，ユーザの知識を広げるような質問を行ったり，関連のある情報を提供するといったことを目的としたものではない．

また，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情報機器を用いた個別学習を支援するシステムが普及しつつある．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[菅沼 2005] [舟生 2010]．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．そこで本研究では，Wikipedia[2] を知識ベースとして利用し，歴史上の人物について学習を行うシステムの実現を目指し，データベースの作成，ユーザ質問への応答手法の提案，および一問一答式の問題の自動生成手法の提案を行う．

2. 応答生成データベース

2.1 データベース構造

データベースの構築には MySQL を用いる．時代，人物，戦いの一覧ページを Wikipedia から取得し，そのページ中のリンクから，歴史に関連のある記事を 4486 件抽出し，記事情報や記事参照量を保持するデータベースを構築した．人物，戦い・出来事に焦点をあて，「鎌倉，室町，南北朝戦後，安土桃山，江戸，幕末，明治の時代の人物一覧」「日本の合戦一覧」に記載された記事をデータベース化した．表1にテーブルの詳細を示す．

表 1 テーブルの詳細

カラム名	型	内容
id	int	記事 ID， オートインクリメント
type	text	戦い，人物，出来事 いずれかの記事の分野
title	varchar(64)	記事タイトル
meta_text	long text	編集者が記した記事本文
referred_num	int	記事の被参照数
カラム名	型	内容

3. 応答生成手法

Wikipedia 記事本文は，編集者により書き方や表現にばらつきがあるため，そこから応答内容を抽出するのは容易ではない．そこで，記述形式がテンプレート化されている Wikipedia 記事の表と，記事冒頭文（これらを「記事情報」と呼ぶ）から応答に用いる情報を抽出する方法を考案した．

3.1 質問文解析

まず，係り受け解析器 Cabocha[3] を用いて質問文を係り受け解析し，その結果から以下のアルゴリズムを用いて主語と述語を推定する．

① 係り先がない句を述語とする
② 係り先が述語となっている句を主語の候補とする
③ 候補の中で末尾が係助詞となっているものを主語とする

次に，質問文中に「いつ」「どこ」「誰」「何」のいずれかの疑問詞を含んでいるか否かを，入力質問文と各疑問詞間の部分一致により判定し，一致した疑問詞を抽出する．

3.2 応答文生成

3.1.で推定した，主語から名詞を取り出し，それと一致するタイトルを持つ記事情報をデータベースに問い合わせる．例えば，

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

[1] IBM WATOSON: http://www.ibmwatson.com/
[2] Wikipedia：http://ja.wikipedia.org/
[3] CaboCha: Yet Another Japanese Dependency Structure Analyzer: https://code.google.com/p/cabocha/

「織田信長はいつ死にましたか？」という質問では，3.1.の手順で主語「織田信長は」が検出されるので，「織田信長」の記事情報をデータベースより得る．次に，表 2 に基づき，推定した述語の動詞と疑問詞に応じて記事情報の抽出箇所を決定し，記事情報から，抽出箇所に対応する情報を取り出す．例えば，上記の例では，疑問詞「いつ」，述語「死にました」が検出されているので，「織田信長」というタイトルをもつ記事の表の没年部分を取り出し答えを生成する．

表 2 抽出箇所決定ルール

疑問詞	述語の動詞	抽出箇所
何，なに	-	冒頭文
誰，だれ	-	冒頭文
何時，いつ	起きる	表（年月日）
	死ぬ	表（没年）
	生まれる	表（生誕）
何処，どこ	起きる	表（場所）

4. 関連情報生成手法

ユーザの求める知識と関連のある情報を提供するために，本研究では，質問の対象あるいは質問対象の分野（戦いか人物か）に応じて以下のように提供すべき関連情報を決定する．

(1) 連続で 3 回以上応答生成に用いた記事が同じ場合，その記事タイトルに興味があると推定する．

(2) 3 回以上連続で応答生成に利用した記事の分野が同じ場合，その分野に興味があると推定する．

(2)の場合には，質問対象と同じ分野でありかつ，直近の質問対象が本文内に含まれている記事を関連情報として提供する．(1)の場合は，その質問対象が本文に含まれている記事を関連情報として提供する．例えば，「織田信長は誰ですか」，「織田信長はいつ生まれましたか」，「織田信長はいつ死にましたか」と織田信長についての質問が 3 回以上連続したときに，織田信長に興味があると推定し，「織田信長」という単語を本文に含む記事をユーザに提供する．

4.1 関連情報提供優先度

上記の方法で関連情報を提供する際，条件を満たす記事が複数ある時は優先度の高いものを提供する必要がある．提供する関連情報に優先度をつけ，有名かつ直近の質問の対象と関連の深い順に関連情報を提供するために，以下の数値と式を用いて，優先度を付与した．

・被参照量（被リンク数）

・質問対象一致数

優先度＝(質問対象一致数)^2*(被参照量)

被参照量:Wikipedia 記事には，他の記事へのリンクが含まれていることがある．被参照量とは，ある記事が他の記事にリンクされている数であり，その記事の有名度を示す．

質問対象一致数:直近の質問の検索対象名が本文内に出現する回数を検索対象一致数とし，その記事と検索対象の関連度とする．

4.2 構築システム動作例

応答生成手法と関連情報生成手法を合わせてシステムを構築し，実際に動作をテストした，その時にシステムと行った対話のログは以下のようになった．

U:織田信長とは誰ですか？
S:織田信長は、戦国時代から安土桃山時代にかけての武将・戦国大名．三英傑の一人
U:彼はいつ生まれましたか？
S:天文3年5月12日（1534 年 6 月 23 日）天文3年5月28日など諸説あり
U:彼はいつ死にましたか？
S:天正10年6月2日（1582 年 6 月21日）
◆◆◆
豊臣秀吉が織田信長と関連がある可能性があります
U:応仁の乱とは何ですか？
S:応仁の乱は・・・

このように目指した機能が正常に動作された事が確認されたので評価実験をおこなった．

5. 質問応答，関連情報提供の評価実験

無作為に質問を 100 件入力し，その結果からシステムの応答精度を算出した．結果を表 3 に示す．

表 3 システムの精度実験結果

正答	67 件
誤答	2 件
見つからない	31 件

応答の正答率は 67%，誤答は 2%，答えが見つからない場合が 31%であった．これは記事に例外的な表現があること，没年等の抽出対象となる情報や，それが含まれる表自体がない記事が存在することが原因であると考えられる．

また，11 名（男性 6 名女性 5 名）にシステムを利用してもらい，アンケートを行った結果，「学習の意欲がわいた」「新しい知識を得られた」といった意見が多く，今後の改良によっては学習支援としての効果が期待できる．また，関連情報提供機能については，有名順に関連情報を提供できているという意見がある一方，一般的でない情報も存在したという意見もあり，ユーザの知識量も考慮して関連情報を選択する必要性が示唆された．

6. 選択肢生成手法

前章では，ユーザ手動で学習を進めるための質問応答機能について述べたが，本章では，システムが主導する教育的インタラクションとして，2.1 で作成したデータベースを用いて，問題を自動生成する手法について述べる．まず，[田村 2014]で提案された問題文生成手法について概要を述べ，次にそこでの問題点への取り組みについて述べる．

6.1 質問生成機構概要

[田村 2014]では，記事冒頭のアブストラクト部分を句点で分割して生成した文から問題文生成を行った．具体例として徳川家康の Wikipedia 記事を用いて問題文を生成する例を示す．

まず，徳川家康のアブストラクトを句点で分割すると

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士豪・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

という7つの文が生成される.

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. この例では分割した文の 3, 4, 6, 7 番目の下線部の「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, 文末表現変換ルールに合致する語句を補完する. このようにして生成した 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士豪・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？

しかし, この方式により生成された問題文には, 自由記述方式で解答させるには, 内容が細かすぎたり, 正解が複数存在したりしてしまう等, 修正を要する問題文も多数生成されてしまった.

そこで選択肢を問題文に付与することによって自由記述方式の問題文から多肢選択問題に変換することにより, これらも適切な問題となるのではないかと考えた. 正解が 1 つになるように選択肢を設定できれば, 自由記述のままでは, 複数正解が存在してしまう問題を, 正解が 1 つである問題に変更することができる. また, 設定する選択肢を動的に変更できるようになれば, 難しい問題に対しては, 誤選択肢が明白であるような選択肢を付与するなど, 同じ問題文に対して様々な難易度の出題ができる可能性もある.

6.2 DBpedia

前節での議論に基づき, DBpedia Japanese を用いた選択肢生成手法を提案する. DBpedia とは Wikipedia から情報を抽出して LOD (Linked Open Data)として公開するコミュニティプロジェクトである. LOD 上では RDF(Resource Description Framework)として情報を提供する. RDF は主語(Subject), 述語(Predicate), 目的語(Object)のトリプル構造でリソースの関係情報を表現している. 図 1 にリソースの具体例を示す. 今回は具体例として,「徳川家康の子供は徳川秀忠」という情報をトリプル構造で表した例を挙げる.

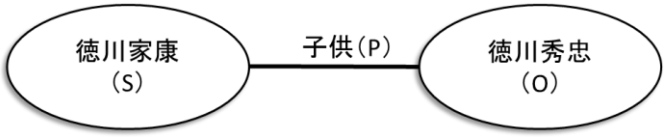

図 1. トリプル構造の具体例

この情報は URI(Uniform Resource Identifier)で識別が行われ, 実際の情報は

- http://ja.dbpedia.org/resource/徳川家康
- http://dbpedia.org/ontology/child
- http://ja.dbpedia.org/resource/徳川秀忠

という形式で格納されている.

6.3 選択肢生成アルゴリズム

選択肢生成アルゴリズムのフローチャートを図 2 に示す. まず始めに記事のタイトルが主語となる述語と目的語の組み合わせの一覧を取得する. そうして取得した述語と目的語のペア(以後 PO セット)で検索をかけることで, 同じトリプル構造を持つ主語の集合を取得する. ここで取得できた主語群は問題を生成した人物と何かしらの共通点を持っており選択肢として有用であると考えられる. しかし, 単一 PO セットで検索を行うと, 取得される主語の数が膨大になるので 2 つの PO セットから得られる主語群の積集合を収集することにより, 有用な主語のみを抽出することを試みた. 理想的には, 全ての PO セットの組み合わせから主語群を抽出すべきだが, 莫大な時間とサーバへの負荷が掛かってしまうので, 本稿では以下の 2 つの制約を設け, 積集合を取得する PO セットを絞り込んだ

- ある PO セットによる検索結果の 50%以上が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 5%以上が検索結果に含まれている PO セットである場合(網羅率による制約)

こうして絞り込んだ PO セット(以後, 選択肢 PO)を用いて 2 重ループを作成し, 検索結果が 2 以上 50 以下の選択肢 PO の組み合わせとその検索結果を格納し, 選択肢候補群とした.

こうして生成した選択肢候補群は記事の重要度などによる重み付けは行っておらず, 教育的に適切であるかどうかの判断はできない. そのため, 選択肢としての妥当性を評価していく必要がある.

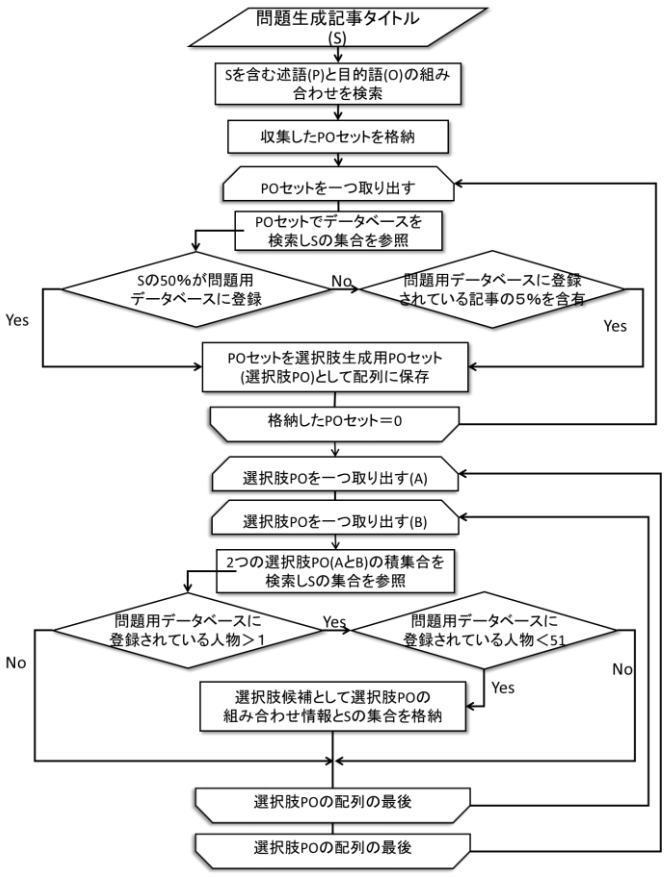

図 2 選択肢生成のアルゴリズム

6.4 選択肢妥当性判定機構

前章で生成した選択肢 PO の中から，不適切な選択肢を排除する機構を提案する．まず，自動生成した問題文を形態素解析し，2 文字以上の名詞のみを取り出す．次に選択肢 PO の O の URI の末尾が取り出した全ての名詞のいずれかと一致するか否かを判定する．名詞が一致した場合にはその PO セットで検索できた人物群を選択肢候補から除外する．

例として「徳川家康」記事から出題される

・三英傑の一人であるのは誰ですか？

という問題文について述べる．

この文章の解答は「徳川家康」「織田信長」「豊臣秀吉」が当てはまり，このままでは解答が一意にならない．そこで，この問題文に形態素解析を行い 2 文字以上の名詞を取り出すと，「英傑」という単語を取り出すことが出来る．

次に 6.3 節で提案した手法により取得した選択肢 PO における O の末尾と「英傑」とのマッチングを行う．その結果，以下の選択肢の O の末尾が「英傑」であることがわかる．

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/MilitaryPerson>
<http://dbpedia.org/ontology/wikiPageWikiLink>
<http://ja.dbpedia.org/resource/Category:三英傑> .

この選択肢 PO で取得できる主語は．「織田信長」「豊臣秀吉」である．従って，この 2 つを選択肢候補から除外すれば，「徳川家康」のみを正解とする選択肢群が生成できる．この選択肢妥当性判定機構と前章の選択肢生成アルゴリズムを「徳川家康」の記事で生成した問題文に対して適用した結果を図 3 に示す．どの問題文の選択肢も正解ではない（つまり，徳川家康のみが正解となる）ことが確認できる．

徳川家康
戦国時代から安土桃山時代にかけての武将・戦国大名であるのは誰ですか？
徳姫 足利義持 足利義詮
江戸幕府の初代征夷大将軍であるのは誰ですか？
後藤基次 長宗我部盛親 水野勝成
三英傑の一人であるのは誰ですか？
武田信玄 伊達政宗 佐野房綱
本姓が先に藤原氏、次いで源氏を称したのは誰ですか？
真田昌幸 徳川秀忠 松平清康
家系が三河国の国人土豪・松平氏であるのは誰ですか？
織田信長 豊臣秀頼 松平忠吉
1567年2月18日に勅許を得て、徳川氏に改めたのは誰ですか？
豊臣秀吉 松平忠正 松平広忠
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？
鵜殿長照 里見義堯 於大の方
幼名が竹千代であるのは誰ですか？
豊臣秀吉 佐野房綱 徳川家定

図 3 不適切な選択肢の削除後の選択肢生成結果

7. 選択肢，問題生成と対話システムの相互利用

3 章4章で開発したシステムと 6 章で提案した問題生成，及び選択肢生成機構の相互利用法を議論する．

質問応答システムでは抽出箇所決定ルールに従い，ユーザは「何」「誰」「いつ」「どこ」の 4 属性の質問をシステムに行っていく．この時に，どの属性の質問をどの記事にたいして行ったかを記録しておく．この情報を利用すれば，その記事から問題生成を行なった際に，ユーザが興味を示した属性から生成された選択肢などが判定可能になるため，新たな属性への興味を促すために選択肢には用いないなどの利用が可能になる．

また，関連情報提示手法を応用し，関連情報提示手法で上位にランキングされる記事を含む PO セットは，有用な選択肢であるとして重み付けをすることが可能である．

更に，現在の質問応答システムでは質問を行うのはユーザのみであるが問題生成と関連情報提示機構を組み合わせることにより，システムから質問をすることも可能である．その際に選択肢生成機構を応用し，解答が一意となる質問文の判定を行えれば，教育効果を期待できるシステムとなる．

8. まとめ

歴史学習支援のための質問応答システムを構築し，興味推定によりユーザの求める知識と関連の深い情報を提供する手法を提案，実装した．また，問題生成機構において，選択肢生成を行う機構を提案した． 今後の課題として，データベース，質問形式を拡張し，対応できる質問を増やし精度を上げること，応答の精度を上げつつ，検討している選択肢属性付け，重み付け手法の確立を行い，教育効果を検証する実験を行なう必要がある．

謝辞：本研究は，科学技術振興機構(JST)戦略的想像研究推進事業(CREST)「実践知能アプリケーション構築フレームワーク PRINTEPS の開発と社会実践」の支援によって実施した．

参考文献

[翠 2007] 翠 輝久，河原 達也，正司 哲朗，美濃 導彦：質問応答・情報推薦機能を備えた音声による情報案内システム,情報処理学会論文誌, Vol.48, No.12, pp.3602—3611(2007).

[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).

[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).

[田村 2014] 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, JSAI2014, 3D4-2in (2014)