F-07

知識獲得手段のための Wikipedia 中の説明文からの質問文生成

Quiz Generation from Explanation Sentences on Wikipedia for Acquiring Knowledge

佐藤 正隆†　　山西 良典†　　福本 淳一†
Masataka Satoh　Ryosuke Yamanishi　Junichi Fukumoto

1. はじめに

近年，「ご当地検定」と呼ばれる地域に関する文化や歴史などの知識量を測る試験がある．ご当地検定では，ある場所に関して，解答者の興味を惹くような質問が出題されている．例えば，ある場所についての知識を獲得手段として，単に説明を提示することよりも質問に解答させることで人により強い興味を持たせることができると考える．解答者は，与えられた質問に対して解答できた場合には充実感を得られる一方で，解答できなかった場合（つまり，質問内容についての知識が乏しい場合）には，新しい知識への好奇心をもつと考えるためである．

本稿では，Wikipedia 中の説明文から自動的に質問文を生成する手法を提案する．Wikipedia と特定した理由は情報量および項目数の観点から最も妥当な知識源であると判断したからである．提案手法では，説明文と質問文の形態的特徴の分析から得られた変換パターンを適用し，説明文を質問文へと自動変換する．ここで，質問として問う対象は，説明文の中の主題とする．説明文の記述には様々なパターンが存在するため，説明文を質問文に変換するためには，説明文のそれぞれのパターンに適した処理を行い，質問文へと変換する必要がある．

2. 質問文の特徴の分析

質問文，および Wikipedia 中の説明文の特徴について分析した．

2.1　質問文の分析

適切な質問文としての条件を分析するため，人手で Wikipedia 中の表題語句を含む 1 文中の表題語句を（　　）にし質問文を作成した．ここで，表題語句は二条城と設定した．二条城を（　　）にした質問文を人手で評価した結果，適切な質問文とは，質問対象と強い関連をもつ情報が含まれていることがわかった．

図 1 に，良い質問と悪い質問の例を示す．図 1 の良い質問では，「二条亭」は質問対象である二条城と強い関連をもつ単語である．「二条亭」は固有名詞であり，1 文中に質問対象と共起する固有名詞は質問対象と強い関連をもつ単語と考えられる．一方で，図 1 の悪い質問では．1 文中に固有名詞が存在しない．固有名詞が質問文中に存在することで表題語句を区別することができると考えられる．

2.2　Wikipedia ページの形態的特徴の分析

良い質問
幕府は（　　）と称したが、朝廷側はこれを二条亭と呼んだ。

悪い質問
この行幸が（　　）の最盛期である。

図 1．良い／悪い質問例

Wikipedia 中の説明文について，形態的特徴を分析した．Wikipedia のページ中で主題の説明文では，主題を表すための係助詞[は]が存在し，かつ主題以外の固有名詞または複合名詞が存在する．主題が難読な漢字であれば主題の直後に括弧を用いてふりがなを記載することもある．上記の形態的特徴を含む 1 文を説明文と定義し，以下のパターンに従って 1 文中の主題を特定する．

(1) [固有名詞 or 複合名詞]＋助詞の[は]

(2) [固有名詞 or 複合名詞]＋[（]＋...＋[）]＋助詞の[は]

3. 提案手法

説明文から質問文へと変換する提案手法の処理手順を以下に示す．また，図 2 に本稿の提案手法の処理手順の図を示す．

1. Wikipedia のテキストを入力データとする
2. Wikipedia のテキストを 1 文毎に分割する
3. 分割した文中に説明文の 2.2 で定めたパターンが含まれている場合，1 文の主題を抽出する
4. 主題についての Wikipedia ページが存在しない場合は，質問表現を[～はなにか？]，[～のはなにか？]とする
5. 主題についての Wikipedia ページが存在する場合，主題の Wikipedia ページから主題の上位概念を取得
6. 4．が存在した場合，質問表現を[“主題の上位概念”～は誰か？]，あるいは，[“主題の上位概念”～はどこか？]とする

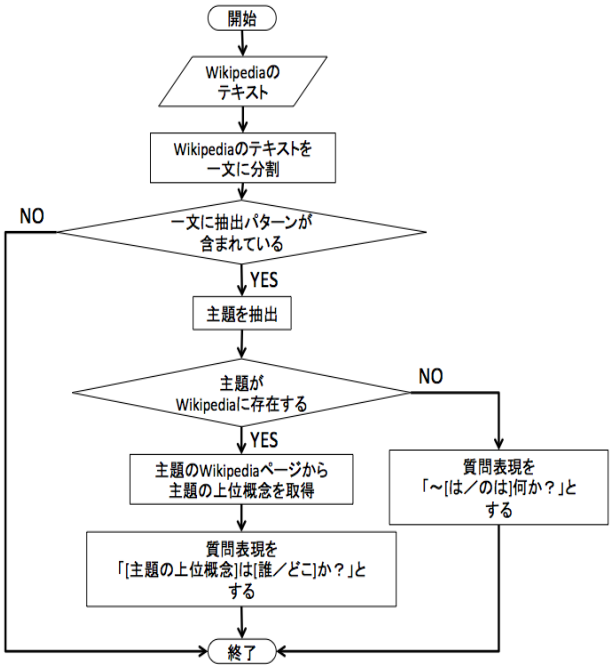

図 2．提案手法の処理手順

†立命館大学，Ritsumeikan University

質問表現では，質問文に変換するために質問表現として説明文の末尾の品詞が名詞の場合は「～は何か？」を追加する．それ以外の品詞である場合は，「～のは何か？」を追加する．特に人名や地名・建築物を問う質問では質問表現が「誰か」「どこか」といった属性を指定する代名詞が記載されることで，より自然な質問文になると考える．そこで，表題語句が人名，地名，建築物である場合には質問表現を属性を持つ代名詞に変換する．下節では，説明文の抽出および主題の抜き出し方法，人名・地名・建築物の質問表現の生成について述べる．

3.1 説明文の抽出及び主題の抜き出し

Wikipedia のテキストから入力した 1 文を MeCab[2]を用いて形態素解析を行い，形態素と品詞情報を取得する．品詞情報を手がかりに説明文を質問文に変換する．ここで，形態素解析用の辞書には，Wikipedia の項目名を固有名詞として追加した．

説明文の抽出パターンに該当する 1 文である場合，主題を文中から削除する．主題の前が名詞である場合，主題が複合名詞となっていると判断し，主題の前の名詞部分も主題に含める．図 3 中，説明文 1 に例を示す．説明文 1 の下線部分は説明文の主題であり，網掛け部分は主題の直前にある名詞である．

主題の直前が助詞の[の]であれば，主題を抜き出した場合助詞の[の]が残り，文法的に不適切となるために，助詞の[の]以前の文字列を削除して質問文を提示する．図 3 中，説明文 2 にその例を示す．説明文 2 の下線部分は説明文の主題であり，網掛け部分は主題に対する連体句である．

説明文 1　Wikipedia 中の金閣寺を対象としたページ中の説明文

室町幕府 8 代将軍足利義政は、祖父の義満が建てた舎利殿に倣い、造営中の東山山荘に観音殿（近世以降銀閣と通称される）を建てた。

説明文 2　Wikipedia 中の清水寺を対象としたページ中の説明文

近世の清水寺は「三職六坊」と呼ばれる組織によって維持運営されていた。

図 3. 主題の前に助詞の[の]があるときの質問文生成の例

3.2 人名・地名・建築物の質問表現の生成

説明文を質問文に生成するために質問表現として「～は何か？」「～のは何か？」を追加する．特に主題が人名，である場合には「～は誰か？」，地名，建築物である場合には「～はどこか？」を属性を表す代名詞を追加する．このとき，主題を表題語句とする Wikipedia ページが存在すれば，「http://ja.wikipedia.org/wiki/“主題”」の形で URL を生成し，主題の Wikipedia ページから 1 文目を取得する．取得した 1 文に対して形態素解析を行い，助詞を含む 1 文で，文末が名詞であれば名詞を取得する．

質問表現の属性の判別には，固有表現のタイプを分類可能である固有表現抽出ツール iNExT[3]を用いた．属性の判別では，予め iNExT に含まれる[誰か]と聞く対象が格納されている辞書データ及び，[どこか]と聞く対象が格納されている辞書データを参照した．

person 辞書は iNExT に登録されている野球選手や首相といった職業など人を表す言葉が含まれており，仏師や僧など iNExT に登録されていない固有表現を追加し作成した．locate 辞書は図書館や大学といった場所を表す言葉が含まれており，建造物や仏教寺院など iNExT に登録されていない固有表現を追加し作成した．

4. Wikipedia 中の説明文から質問文への自動生成実験

本実験では，Wikipedia 中の清水寺・金閣寺・八坂神社・仁和寺・東福寺・西本願寺・平等院・東寺を対象として説明文からの質問文の自動生成を行った．自動生成された質問文は全部で 182 文であった．実験者が質問文を文法的に主観評価したところ，以下の結果が得られた．

- 適切だと判断した質問文 62.64%（114/182）
- 不適切だと判断した質問文 37.36%（68/182）

下節では，提案手法によって自動生成された質問文について考察する．

4.1 文法的に適切な質問文の考察

4.1.1 主題が Wikipedia 中に存在しない場合の質問文生成

説明文の主題が Wikipedia 中に存在しない場合，質問表現として，文末が名詞である場合は「～は何か？」文末が名詞以外である場合は「～のは何か？」に変換した．その結果，文末が名詞である場合は名詞の次に係助詞[は]があるため，文法として適切な質問文が生成されたと考えられる．それぞれの出力例を図 4 に示す．

質問文 1 では，「神社」という名詞のあとに係助詞[は]があり，自然な質問文に見える．質問文 2 では，[た]という名詞以外の品詞のあとに形式主語の[の]がおかれているため，係助詞[は]は形式名詞[の]にかかり文法的に適切となった．

質問文 1　文末が名詞である説明文からの質問文生成例
官幣大社で、現在は神社本庁の別表神社は何か？

質問文 2　文末が名詞ではない説明文からの質問文生成例
「三職六坊」と呼ばれる組織によって維持運営されていたのは何か？

図 4. 主題が Wikipedia 中に存在しない場合の質問文生成例

4.1.2 主題が Wikipedia 中に存在する場合の質問文生成

説明文の主題が Wikipedia 中に存在する場合，主題の Wikipedia 中の第 1 文から上位概念語を抽出し，「～[上位概念語]は[誰／どこ]か？」とした．また，誰・どこの判定は person 辞書，locate 辞書を参照した．その結果，上位概念語を示すことにより，質問文としての聞く対象を特定することができ，より適切な質問文が生成されたと考える．その出力結果を図 5 に示す．

質問文 1 では，解答の祇園社についてのページが Wikipedia 中に存在したため，上位概念語である「神社」が取得できた．また，「神社」は locate 辞書に登録していたため地名の代名詞である「どこ」に質問表現を変更した．その結果「～神社はどこか？」という質問表現とな

り質問文として聞く対象を特定できたため，より適切な質問文であると考える．

質問文 2 では，解答の「杲宝」についてのページが Wikipedia 中に存在したため，上位概念語である「学僧」が取得できた．また，「学僧」は person 辞書に登録したため，人名の代名詞である「誰」に質問表現を変更した．その結果「～学僧は誰か？」という質問表現となり質問文として聞く対象を特定できたため，より適切な質問文である．

質問文 1　解答が地名・建築物である質問文生成例
当初は興福寺の配下であったが、10 世紀末に戦争により延暦寺がその末寺とした神社はどこか？
解答：祇園社

質問文 2　解答が人名である質問文生成例
現在国宝となっている「東宝記」という東寺の創建から室町時代に至る寺史をまとめた学僧は誰か？
解答：杲宝

図 5．主題が Wikipedia 中に存在する場合の質問文生成

4.1.3　主題が複合名詞である場合の質問文生成

提案手法で主題を抽出する際，主題以前が名詞である場合，その名詞も主題に含み質問対象とした．その結果，複合名詞の主題を取ることができたため，適切な質問文及び解答が生成されたと考えられる．その出力結果を図 6 に示す．

説明文　Wikipedia 中の金閣寺を対象としたページ中の説明文
室町幕府 8 代将軍足利義政は、祖父の義満が建てた舎利殿に倣い、造営中の東山山荘に観音殿（近世以降銀閣と通称される）を建てた。

質問文　解答が複合名詞である質問文生成例
祖父の義満が建てた舎利殿に倣い、造営中の東山山荘に観音殿（近世以降銀閣と通称される）を建てた将軍は誰か？
解答：室町幕府 8 代将軍足利義政

図 6．主題が複合名詞である場合の質問文生成

図 6 の質問文は，主題が「室町幕府 8 代将軍足利義政」という複合名詞であるため，文中から複合名詞を主題として抜き出し質問対象とすることができた．

4.2　文法的に不適切な質問文分析

提案手法により不適切だと判断された質問文は 182 文中 68 文であった．今後の精度向上のために不適切な質問文の特徴を考察する．提案手法により文法的に不適切だと判断した質問文には，以下の特徴が存在した．

1. 質問表現が不適切である
2. 係り受けが不適切である

4.2.1　質問表現が不適切な質問文

主題が人物，地名，建築物であるが質問表現が属性を持つ代名詞ではない質問文が存在した．例として Wikipedia 中の清水寺を対象としたページ中の説明文と質問表現が不適切な質問文の出力結果を図 7 に示す．説明文中の下線部分は，主題となる固有名詞である．質問文中の下線部分は，質問文に質問表現を成形した結果である．

図 7 は，解答が「大西」という人であるため，人を解答とする質問表現である「～は誰か？」が適切であるが，そのような出力はされなかった． 不適切な質問文が得られた原因として，本来説明文の主題は[大西良慶]であるが 2.2 のパターンにより主題を[大西]として認識したためと考えられる．この解決案として， [大西]を[大西良慶]と補完することにより解決できると考えられる．

説明文　Wikipedia 中の清水寺を対象としたページ中の説明文
大西は昭和 40 年（1965 年）に法相宗から独立して北法相宗を開宗、初代管長となった。

質問文　質問表現が適切ではない質問文
昭和 40 年（1965 年）に法相宗から独立して北法相宗を開宗、初代管長となったのは何か？
解答：大西

図 7．質問表現が不適切な質問文

4.2.2　係り受けが不適切な質問文

主題を説明文から削除した際に，主題に対する修飾部が主題の次の名詞にかかり受けてしまい，質問文の内容が変化してしまう質問文が存在した．例として Wikipedia 中の清水寺を対象としたページ中の説明文と係り受けが不適切だと判断した質問文の出力結果例を図 8 に示す．図 8 の下線部は修飾部であり，説明文中の網掛け部分は，主題である． 質問文中の網掛けは下線部の修飾部をかかり受けた名詞句である．図 8 は修飾部が係り受けていた主題を質問文にする際に抜き取ってしまうため，修飾部が主題を抜いた次の名詞句[修行中の賢心]にかかってしまい適切な質問文ではなくなると考えた．係り受けに関しては，連体形を連用形に変換することで適切な質問文へと変換することができると考えられる．図 8 の質問文では[音羽山に入り込んだ]を[音羽山に入り込み]という連用形に変換することで適切な質問文にできると考えられる．

説明文　Wikipedia 中の清水寺を対象としたページ中の説明文

その 2 年後の宝亀 11 年（780 年）、鹿を 捕えようとして音羽山に入り込んだ坂上田村麻呂（758 年 - 811 年）は、修行中の賢心に出会った。

質問文　係り受けが適切ではない質問文
2 年後の宝亀 11 年（780 年）、鹿を捕えようとして音羽山に入り込んだ修行中の賢心に 出会ったのは何か？
解答：坂上田村麻呂（758 年 - 811 年）

図 8．係り受けが不適切な質問文

4.3　質問文の内容が不適切な質問文

説明文　Wikipedia 中の仁和寺を対象としたページ中の説明文

御室桜は日本さくら名所 100 選に選定されている。

質問文　一意に定まらない質問文
日本さくら名所 100 選に選定されているのは何か？
解答：御室桜

図 9．質問文の内容が不適切だと判断した例

生成した質問文の中には，文法的には誤りではないが主題を特徴付ける固有名詞が存在しないために不適切であると考えられる質問文が存在した． 例として Wikipedia 中の仁和寺を対象としたページ中の説明文と一意に定まらない質問文の出力結果例を図 9 に示す．説明文中の下線部分は，主題である．質問文中の下線部分は，質問文に質問表現を成形した結果である．

5. おわりに

本稿では，Wikipedia 中の説明文を質問文へ変換する手法を提案した．Wikipedia 中の様々な説明文を質問文に変換するためにはそれぞれ説明文のパターンに適した処理を行う必要がある．

今後，本稿で文法的に不適切であると考察した問題を解決する．そして，質問文の内容が不適切であると判断した質問文の分析を行い，一意に定まらない原因を追求する．また，ユーザの知識量に応じた質問文を提示するしくみについても考えていく．例えば，ユーザの未知である事柄に対しては有名な解答が導かれる質問文を提示し，既知である事柄に対しては無名な解答が導かれる質問文を提示する．これにより，ユーザの知識量に合わせた適切な質問提示ができると考える．

参考文献

[1] 高野　澄：第 1 回京都検定　問題と解説，2005 年，京都新聞出版センター，pp207

[2] 工藤拓，山本薫，松本裕治：Conditional Random Fields を用いた日本語形態素解析,情報処理学会自然言語処理研究会,SIGNL-161, 2004.

[3]山内亮子, 福本淳一： 質問応答システムにおける類似回答の統合, 自然言語処理研究会, NL193-1, pp.1-6, 2009.9.