

試験問題をリソースとした知識ベース自動構築手法の検討

Study on a Method for Automatic Construction of Knowledge Base Using Exam Questions as Resources

古館 昌伸^{*1}, 福坂 祥基^{*2}, 高木 正則^{*2}
 Masanobu FURUDATE^{*1}, Syoki FUKUSAKA^{*2}, Masanori TAKAGI^{*2}

^{*1}岩手県立大学大学院ソフトウェア情報学研究科
^{*1}Graduate School of Software and Information Science, Iwate Prefectural University
^{*2}岩手県立大学ソフトウェア情報学部
^{*2}Faculty of Software and Information Science, Iwate Prefectural University
 Email: g2311032@s.iwate-pu.ac.jp

あらまし：近年，Wikipedia などの情報を利用して汎用的なオントロジーを構築する研究が実施されている．本研究では問題の自動生成に活用できる知識ベースの構築を目的とし，多肢選択式問題の情報から知識ベースを自動構築する手法を提案する．本手法を実現できれば検定試験の過去問題から知識を体系化でき，各検定試験の出題分野の傾向や年度ごとの比較を容易にできるようになる．また，検定試験の作問を支援することもできる．

キーワード：知識ベース，オントロジー，検定試験，多肢選択式問題，ご当地検定

1. はじめに

近年，Wikipedia などの Web 上の知識をリソースとして，汎用的な大規模オントロジーを構築する研究が行われている⁽¹⁾⁽²⁾．これらの研究では，これまでより低コストで非常に大規模なオントロジーを半自動的に構築することが可能である．しかし，Web 上で扱われている知識のみにしか対応していない．

一方，試験問題の作成現場では，作問負担が問題となっており，問題の自動生成に関する研究が行われている⁽³⁾⁽⁴⁾．これらの研究では，Web 上の情報や問題作成者が入力した情報をリソースとして問題を生成することが多い．Web 上の情報を活用した場合，Web 上に記載されていない情報に関する問題が生成できず，問題作成者が入力した情報を活用する場合は，問題作成者に負担が生じる．本研究では，問題作成者に負担をかけずに Web 上の情報に依存しない問題自動生成用の知識ベースの構築を目的とし，検定試験の過去問題をリソースとした知識ベースの自動構築手法を検討する．

2. 構築する知識ベースの概要

本研究では，検定試験で最も多く使われている多肢選択式や一問一答式の問題で，答えが名詞の問題を活用して知識ベースの構築を行う．実際に活用する問題には，共同研究を行っている盛岡商工会議所が主催する岩手県盛岡市のご当地検定「盛岡もの識り検定（以下もりけん）」の過去問題を用いる．

我々はまず以下の手順により，もりけんの過去問題から盛岡をルートとした知識ベースの構築を人手で行った．知識ベース構築手順の概要を図 1 に示す．

- (1) 問題文と答え（分類対象）を抽出
- (2) 問題文から「その問題で問われている知識」を読み取り，この知識の上位概念をカテゴリとし

- て各問題に付与
 - (3) 盛岡をルートとしたカテゴリツリーに(2)で付与したカテゴリを追加（親子関係の構築）
 - (4) 追加したカテゴリの末端に分類対象である答えを追加
- 構築した知識ベースの一部を図 2 に示す．なお，図 1，図 2 において，四角の要素が分類の対象を表

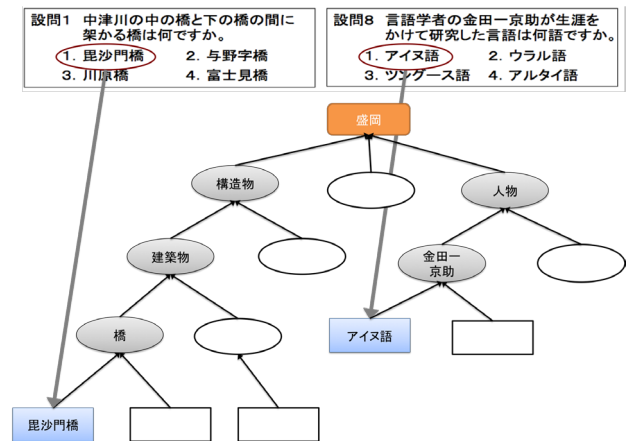


図 1 人手による知識ベース構築手順の概要

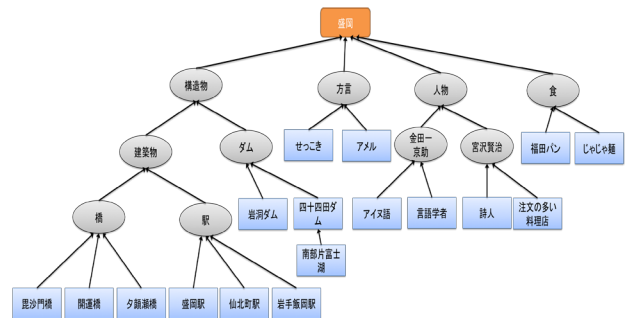


図 2 人手で構築した知識ベースの一部

し、丸の要素が分類対象の親カテゴリを表す。図 1 の設問 8 では、答えは「アイヌ語」だが、問題では金田一京助についての知識が問われているため、この問題のカテゴリは「人物」となる。このように、答えの名詞のみを参照したカテゴリ分類は行わず、ルート「盛岡」とその問題で問われている知識（アイヌ語）を結ぶ知識（金田一京助が研究した言語）を考慮し、カテゴリの分類を行った。また、分類対象とその親との親子関係の意味合いを付与する必要がある。例えば図 2 では「宮沢賢治」の子要素として「詩人」と「注文の多い料理店」があるが、これらは「宮沢賢治（の職業は）詩人」、「宮沢賢治（の作品は）注文の多い料理店」という繋がりになっており、それぞれ意味合いが全く異なる。作問支援という観点においては、この親子間の関係性が非常に重要な役割を果たすと考えられる。

手動での分類作業の結果を踏まえ、問題自動生成のための知識ベースの要件を以下のように定義した。

- 多様な階層の深さに対応
- 答えの名詞（分類対象）の上位概念ではなく、その問題で問われている知識とルート（盛岡）との関連を考慮した分類を行う
- 分類対象とその親との関係性（意味的なつながり）を付与する

以上の要件を満たす知識ベースを構築することにより、リソースとした過去問題とほぼ同じ問題が生成できるだけでなく、類似問題の生成も期待できる。

3. 知識ベースの自動構築手法

本稿で提案する知識ベースの自動構築手法の概要を図 3 に示す。2 章を元に、我々は以下の手順で知識ベースの自動構築を行うこととした。

- (1) 各問題について問題文と正解選択肢を抽出し、トリプル⁽⁵⁾に変換
- (2) カテゴリ階層の作成
- (3) インスタンスの作成
- (4) 作成したカテゴリとインスタンスの関係性を付与

3.1 各問題をトリプルに変換

まず、各問題から抽出した情報を用いて「主語-述語-目的語」のトリプル形式に変換する。もりけんの場合、主語は問題文の先頭から助詞が出現する前までの名詞が該当する。述語はそれ以降から「～はどこですか」などの文末より前の部分となる。また、目的語は正解選択肢が該当する。

3.2 カテゴリ階層の作成とインスタンスの作成

トリプルに変換後、分類対象（インスタンス）を選出する。本稿ではインスタンスを目的語（正解選択肢）として扱う。また、主語の名詞が属するカテゴリを「その問題で問われている知識」のカテゴリと判断し、主語の名詞が属するカテゴリ階層を作成する。その後、作成したカテゴリ階層の末端に先ほど選出したインスタンスを追加する。

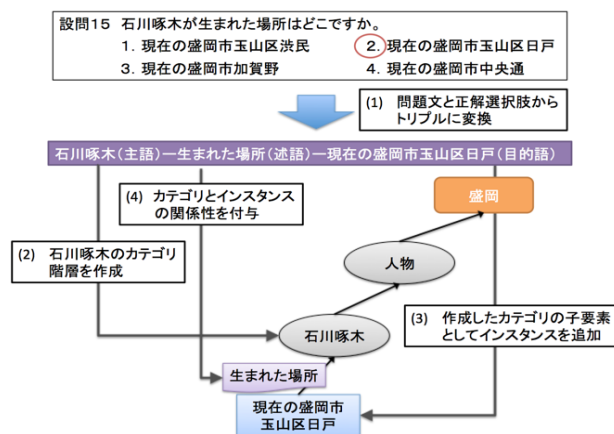


図 3 知識ベース自動構築手法の概要

3.3 カテゴリとインスタンスの関係性を付与

3.2 節で追加したカテゴリ、及びインスタンスの関係性を付与する。トリプルの場合、この関係性の部分は述語が該当するため、述語をカテゴリとインスタンスの関係性とし、付与することとする。

4. まとめと今後の課題

本稿では、検定試験の過去問題から知識ベースを構築することを目的とし、ご当地検定の過去問題から地域に関する知識ベースを自動生成する手法について提案した。本研究により問題の自動生成支援が可能となるほか、各検定試験の出題分野の傾向や年度ごとの比較を行うことが可能となる。今後は評価実験を行っていく予定である。また、今回は画像や地図を用いた問題を対象外としたが、それらへの対応も行っていく。更に今回構築した知識ベースを観光業などの他分野へ展開したり、他の検定試験の知識ベース構築への適用についても検討していきたい。

謝辞

本研究の一部は科学研究費（若手研究（B）、No.24700904）の研究助成を受けたものである。また、本研究に協力して頂いた盛岡商工会議所と文化地層研究会の皆様には感謝を申し上げます。

参考文献

- (1) 柴木優美, 永田昌明, 山本 和英: 日本語語彙大系を用いた Wikipedia からの汎用オントロジー構築, 情報処理学会 研究報告, NL194-4(2009).
- (2) 玉川奨, 桜井慎弥, 手島拓也, 森田武史, 和泉憲明, 山口高平: 日本語 Wikipedia からの大規模オントロジー学習, 人工知能学会論文誌 25(5), pp.623-636(2010).
- (3) 菅原遼介, 高木正則: 記述式問題の誤回答を用いた誤答選択肢自動生成システムの開発, 情報処理学会 情報教育シンポジウム 2013(2013).
- (4) 津森伸一, 海尻賢二: 理解状況に適応した選択問題生成方法の検討, 教育システム情報学会誌 26(3), pp.240-251(2009).
- (5) RDF Primer, <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>