

# 問題の自動生成に活用する知識ベースの自動構築手法の提案

古舘昌伸<sup>†1</sup> 福坂祥基<sup>†2</sup> 高木正則<sup>†2</sup>

我々は特定の地域に関する文化や歴史などについての知識を測るご当地検定を対象とした、ユーザ参加型ご当地検定用学習コミュニティシステムを開発・運用している。本研究では、作問による学習コンテンツの作成を通して地域住民や観光客などの一般ユーザの知識を収集し、地域に関する知識ベースの構築を目指している。しかし、先行研究から、地域に関する知識が体系化されておらず、収集すべき知識が明確化されていないという問題が明らかになった。そこで、我々は現在ある知識から地域に関する知識ベースを構築することを目的とし、研究を進めることとした。知識ベースの構築には莫大なコストがかかるため、本稿ではご当地検定の過去問題から知識ベースを自動構築する手法について検討する。知識ベースの構築により知識を体系化することで、ご当地検定の試験問題作成支援が可能となるほか、地域毎の知識ベースの比較や観光業等の他分野への応用など、ご当地検定以外の分野への展開も期待される。

## A Proposal of Method for Automatic Construction of Knowledge Base for Automatic Generation of Questions

MASANOBU FURUDATE<sup>†1</sup> SHOKI FUKUSAKA<sup>†2</sup>  
MASANORI TAKAGI<sup>†2</sup>

We have development and operation of the learning community system in local knowledge test that focused on Gotouchi Test. In this study, we collect knowledge from local inhabitants or tourist through the making of learning contents by the problem making. And we aim at the knowledge base construction about the area. However, we are known that we had a problem that it was not clarified that should collect knowledge because don't knowledge about the area by a precedent study. So, we aimed to construction of knowledge base on regions by already exists knowledge. The knowledge base construction entails a very big cost. So, this paper considered method for automatic construction of knowledge base on regions using past exams of local knowledge test. We can support examination questions making of the local knowledge test by construction a knowledge base. And the application to other fields such as the knowledge base comparison and tourism between areas is possible, too.

### 1. はじめに

近年、地域の魅力を再発見する試み(地元学や地場産業、フィールドワーク等)が行われるようになってきた。これらは自分たちが暮らしている地域について学ぶことにより地域の良さを再発見し、その良さを発信する人材の育成や地域特有の新たな観光資源を発掘することが目的である。

地域の魅力を再発見する試みの一つとしてご当地検定がある。ご当地検定とは、特定の地域に関する文化や歴史などについての知識を測る検定試験である。2010年時点で全国約250の実施団体が存在し、今後も実施団体数の増加が予想される[1]。しかし、ご当地検定では問題作成の負担が大きいことや、問題数の不足が大きな問題となっている[2]。この問題に対し、我々は地域住民や観光客などが地域に関する問題を投稿可能なユーザ参加型ご当地検定用学習コミュニティシステムを開発・運用している[3]。本システムにおいて、作問による学習コンテンツの作成を通してユーザから知識を収集することにより、地域に関する知識ベースの構築を目指してきた[4]。

しかし、本システムの運用を通し、現在地域に関する知識が体系化されておらず、どのような知識があるのかが不

明であること、それに伴い収集すべき知識が明確ではないことが判明した。そこで本稿では、既存の知識から地域に関する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を用いて地域の知識ベースを自動構築する手法を提案する。

### 2. 先行研究で開発したシステムの概要

#### 2.1 システムの概要とこれまでの取り組み

本システムは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)」をフィールドとしている。本システムの全体構成を図1に示す。本システムには、過去の「も

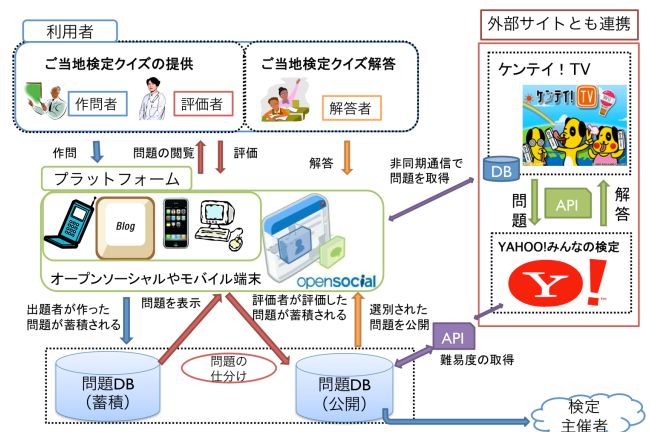


図1 システムの全体構成

<sup>†1</sup> 岩手県立大学大学院  
Iwate Prefectural University Graduate School

<sup>†2</sup> 岩手県立大学  
Iwate Prefectural University

りけん」に出題された全問題が監修されており、ポイントやバッジを取得しながら問題を解答できる機能が提供されている。また、従来少人数の専門家で作成された作問チーム（作問委員会）のみで行っていた問題案作成の部分を、観光客や地域住民等の一般市民に参加してもらうことで、作問委員会の負担が軽減され、試験問題（コンテンツ）の充実化も図れると考えた。更にコンテンツの投稿を行うユーザー数が増加することにより、より幅広い分野のコンテンツ収集が可能になると期待される。本システムにおける問題投稿時の画面例を図2に示す。

この他、検定試験の過去問題が出題される本システムで扱っている問題の難易度が高く、ユーザが気軽に参加できないという利用者からの意見に対し、「もりけん入門クイズ」を開発した。これはもりけんへの興味を持たせ、気軽に問題にチャレンジできるように難易度を下げた問題を手軽に解答出来るスマートフォンアプリケーションとなっている。Android版とiOS版、更にFacebookアプリケーション版を作成し、公開している[5][6]。本アプリケーションの画面例を図3に示す。マルチプラットフォームに対応することで、検定受験者以外のより多くの利用者層に利用してもらえることが期待される。更にシステムの利用者層を広

めるために、本アプリケーションを用いて、盛岡について知ってもらい、興味を持ってもらうことを目的としたセミナーを本学の大学生や地域住民の方を対象として開催した。

## 2.2 システムの運用状況

現在までに本システムの利用者数は320名となっている。現在までの利用状況を分析したところ、ほとんどのユーザが問題の解答機能のみの利用に留まっており、本システムの狙いであるユーザによるコンテンツ作成は行われなかった。また、検定試験が行われる直前はシステムへのアクセス数が増加していたのに対し、検定試験が終わるとアクセス数は減少していった。このことから、現状ではご当地検定の試験対策用としてのみ本システムが用いられており、ユーザが知識を発信する場として、またユーザ間で知識を深めていく場として機能していないことが分かった。現状の本システムの利用者のほとんどが検定試験受験者であるため当然の結果とも考えられるが、ユーザ自らが自主的に知識を発信させる仕組みがない点にも着目し、ゲーミフィケーションに着目した問題投稿を誘発する仕組み[7]や、質問行動を促し教える活動を誘発する仕組み[8]をこれまでに提案し、実装してきた。しかしユーザから知識を収集し、ご当地検定の問題に利用する仕組みを検討する上で、現在どのような知識があり、ユーザから収集したい知識はどのようなものなのかを明確化する必要があることが分かった。

そこで本研究では、ご当地検定の過去問題を利用し、今ある知識から地域に関する知識ベースを構築することとした。知識ベースの構築を手で行うのは莫大なコストがかかってしまい、かつ最新の情報を反映させることが困難なため、本稿では知識ベースの自動構築手法について検討していく。知識ベースについては、ご当地検定の作問支援を可能とするものを構築することを狙いとする。しかし、ご当地検定への利用のみならず、観光業などの他の分野への適用も考慮し、より有用性・汎用性の高い知識ベースを構築することを目標とする。

## 3. 既存の取り組みと関連研究

本章では、知識体系の自動構築・利用に関する主な先行研究や既存の取り組みについて紹介し、本研究で試験の過去問題を用いる有用性や、過去問題をリソースとした知識ベースの自動構築を行う意義を述べる。

柴木ら[9]は日本語の大規模オントロジーである日本語彙大系を用いて、日本語 Wikipedia から汎用オントロジーを半自動で構築することを試みている。一方、玉川ら[10]は日本語 Wikipedia のみをリソースとし、大規模かつ汎用的な日本語 Wikipedia オントロジーを構築している。どちらも Wikipedia を用いており、ユーザ参加型という特徴から“Wikipedia は語彙網羅性、即時更新制に優れており、半構造情報資源であることからフリーテキストと比べてオントロジーとのギャップが小さいため、オントロジー学習の

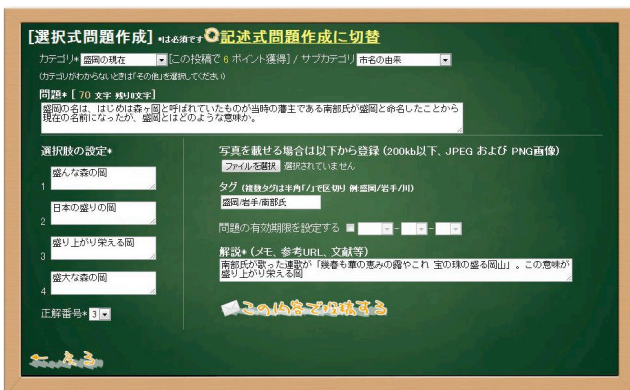


図2 システムの画面例（作問画面）



図3 もりけん入門クイズ画面例

情報資源として非常に有用”と述べられている。

DBpedia[11]はWikipediaの半構造情報をRDFに変換することによって、大規模なデータベースを構築している。リソースには英語版Wikipediaを利用しており、Linked Open Data (LOD) のハブとして様々な領域のLODとリンクされている。また、日本語版Wikipediaを対象とし、独自でマッピング作業を行っているDBpedia Japaneseも存在している[12]。

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に成功している。しかし、既存研究ではリソースがWikipediaのみであり、Web上に情報が記載されている知識のみにしか対応していない。本研究で対象とするご当地検定試験は地域によって扱うコンテンツが違い、更に時事問題など年を重ねる毎に対象とする知識が増え、情報更新の頻度が多いものもある。また、このような新規知識に加え、地域特有の限定的な知識などはWikipediaや他のWeb上に存在しない場合も多数ある。よって、ご当地検定試験に向けた知識ベースの構築を検討する際には、現状の仕組みだけでは対応することは困難であるといえる。

更に作問委員会が編集を行い、もりけんの参考書となっている「もりけん本」及び「もりけん本スーパー」がある[13]。本書には「盛岡の歴史」や「盛岡の文化」、「盛岡の現在」などのカテゴリ別に情報が記載されている。しかし、現状のカテゴリ分けは大雑把なものが多く、妥当性に難点がある。また、情報が乱立しており、体系化されているとは言い難い。

また、菅原ら[14]は試験問題の作成現場における作問負担の問題に着目し、一問一答形式の問題を解答した際の誤回答情報を用いて誤答選択肢を自動生成するための知識ベースを構築している。菅原らの研究で用いる知識ベースの例を図4に示す。しかし、現状では知識ベースの構築を人手で行っており、かつ誤回答情報のみをリソースとしているため、網羅性に難点がある。同じように津森ら[15]は、専門用語などの辞書的な構造を持つ語彙の習得を支援すべく、語彙を木構造で表現した語彙情報と問題テンプレート

を用いて、学習者の理解状況に応じた選択問題を自動生成する方法を検討している。この語彙情報についても予め用意しておく必要がある他、ご当地検定の知識ベースにおいては必ずしも語彙同士が辞書的な構造(親子関係)を表せるわけではない。更に、我々は問題作成者にとって語彙情報の構築における負担が大きいと考える。

よって、本稿では問題作成者に負担をかけずに、Web上の情報や既存の本に依存しない知識ベースの構築を目的とし、ご当地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。本研究のように検定試験の過去問題から知識ベースを構築する研究はまだ行われていない。また、本研究により構築した地域に関する知識ベースに図4のような誤解答の知識ベースを組み合わせることによって、問題文と解答、誤答選択肢の自動生成が可能になる。

#### 4. 期待される効果

検定試験の過去問題から知識ベースを自動構築する手法を確立することにより、期待される効果を以下に述べる。

- 他検定試験への応用
- 特定の地域に関する知識ベースという新しい知識体系の構築
- 様々な地域毎の知識ベースを構築することにより、新しい視点での地域間の比較
- 乱立していた地域に関する知識が体系化され、一つにまとめられる
- 不足している知識が可視化されることで、その知識を収集するのに最適な手段の検討が可能となる
- ご当地検定用試験問題の自動生成支援による作問の負担軽減
- 観光業などの他分野への展開

以上のように、地域に関する知識ベースを構築すること、及び自動構築手法を確立することは、ご当地検定のみならず様々な分野にとって有益なものとなることが考えられる。

#### 5. 過去問題からの知識ベース構築

##### 5.1 もりけんの問題形式

本研究で対象としているもりけんで出題される問題形式について紹介する。もりけんは年に1回開催され、1級・2級・3級の試験があり、1級は50問、2・3級は100問出題される。もりけんの級別の出題形式の内訳をまとめたものを表1に示す[16]。2, 3級は多肢選択形式の問題のみが出題されるが、1級はそれに合わせ一問一答形式の問題も出題される。どの級でもほとんどの問題が答えは名詞となるものだが、答えが数字や説明文になるものも存在する。本研究では答えが説明文となる問題は対象外とする。また、各級で画像を用いた問題(図5)や地図を用いた問題(図6)も出題される。これらの問題は文字情報が少なく、問題文

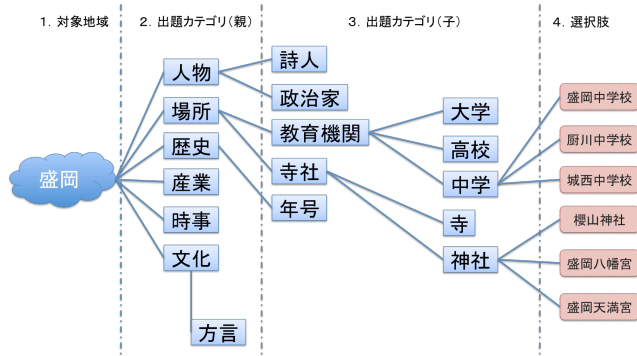


図4 菅原らの研究で用いる知識ベースの例



表 1 もりけんの出題形式内訳

級	出題数	出題内容	合格基準
1 級	50 問 (記述式 20 問, 4 者択一の選択問題 30 問)	■盛岡についての高度な事項 1. 盛岡の現在                      2. 盛岡の気候と地理 3. 盛岡の産業                      4. 盛岡の文化 5. 盛岡の先人・著名人          6. 盛岡の歴史	80%以上の正解をもって合格とします。
2 級	100 問 (4 者択一の選択問題)	■盛岡についてのやや高度な事項 1. 盛岡の現在                      2. 盛岡の気候と地理 3. 盛岡の産業                      4. 盛岡の文化 5. 盛岡の先人・著名人          6. 盛岡の歴史	70%以上の正解をもって合格とします。
3 級	100 問 (4 者択一の選択問題)	■盛岡についての基礎的な事項 1. 盛岡の現在                      2. 盛岡の気候と地理 3. 盛岡の産業                      4. 盛岡の文化 5. 盛岡の先人・著名人          6. 盛岡の歴史	70%以上の正解をもって合格とします。

や正解選択肢から分類を行うことが困難なため、本研究では対象外とする。更に、「～について、正しくないものは次のうちどれですか」というような誤りを発見する問題についても、正解選択肢が間違っている情報となるため知識ベースに格納することは出来ない。よってこのような問題も本研究では対象外とする。また、問題文に知識ベースに格納したい情報が二つ入っている問題もあった。このような問題文は二つに分割し、二つの問題として扱うこととする。

## 5.2 人手での知識ベース構築

我々はまず、もりけんの過去問題から人手で知識ベース構築を試みた。なお、知識ベースはツリー構造で表せると仮定し、かつほとんどの問題において答えが名詞になっていることから、知識の分類を行うべき対象(分類対象)を正解選択肢に定めた。現在、もりけんの過去問題は全部で1750問存在する。その中から2013年度のもりけん3級問題96問(2013年度の3級問題は100問存在するが、5.1節で述べた本研究で対象外とする問題7問を省き、更に問題を分割した計96問)について、以下の手順で知識ベースの構築を行った。

- (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クラスとして追加(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インスタンスとして追加

上記の知識ベース構築手法の概要を図7に示す。なお、図中において四角の要素が分類対象を表し、丸の要素が分類対象の親カテゴリを表す。図7中の設問1では、「毘沙門橋」が答えになっており、この問題は「橋についての知識」が問われている。よって、この問題のカテゴリは「橋」

設問33 写真の建物は次のどれですか。

1. 米内浄水場
2. 旧陸軍覆馬場
3. 旧第九十銀行本店
4. 旧盛岡高等農林学校旧本館



図 5 画像を用いた問題例

設問47 岩手県の県名の由来となった「鬼の手形伝説」が残る「三ツ石神社」のある場所は、地図中のどれですか。

1. A
2. B
3. C
4. D



図 6 地図を用いた問題例

となる。橋は「構築物」と親子関係にあり、構築物は「構築物」と親子関係にあるので建築物、構築物の親カテゴリがカテゴリツリーに追加される。そして、橋の下にこの問題の分類対象である「毘沙門橋」が追加される。図7中の設問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」となり、カテゴリツリーには人物が追加され、その子要素に「金田一京助」が追加される。そして、金田一京助の下に、この問題の分類対象である「アイヌ語」を追加する。

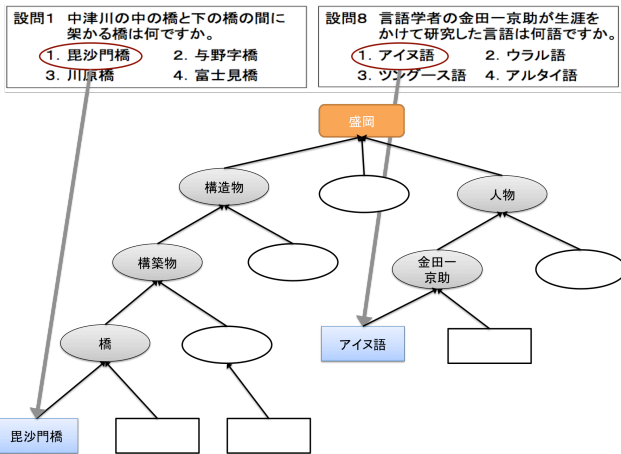


図 7 人手による知識ベース構築手法の概要

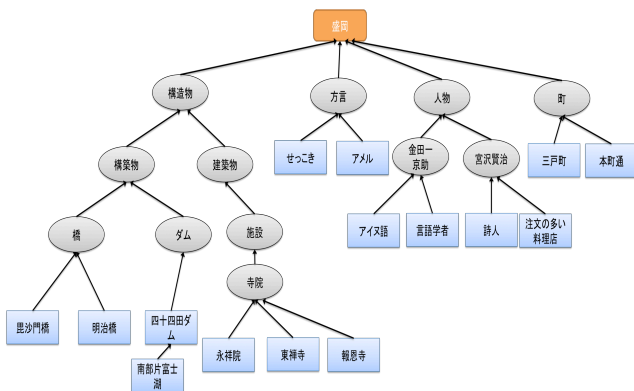


図 8 人手で構築した知識ベースの一部

以上のような手順により、盛岡をルートとしたカテゴリツリーによる盛岡の知識ベースを構築した。構築した知識ベースの一部を図 8 に示す。構築した知識ベースの特徴として、カテゴリ階層の深さにばらつきがあることが挙げられる。また、カテゴリ数にもばらつきがあり、特に建造物と人物の子カテゴリ、およびインスタンス数が他と比べ多かった。更に図 7 の設問 8 のような問題の場合、分類対象である答えの意味のみを参照したカテゴリ分類は行えず、ルートの「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、分類を行う必要があることが分かる。つまり、辞書的な構造を持つ日本語語彙大系などの既存の汎用オントロジーのみを用いたカテゴリ分類は行えない。また、分類対象とその親カテゴリとの親子関係の繋がりがどのような意味合いを持つのかを付与する必要がある。例えば図 8 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。更に問題によっては、問題文中に答えを導くためのヒントとなる情報が存在していることが分

表 2 人手で構築した知識ベースの詳細

問題数	96 問
カテゴリ数 (クラス数)	61
インスタンス数	137
ツリーの深さ (最大)	6
ツリーの深さ (最小)	2
ツリーの深さ (平均)	3.54

かった。この情報についても知識ベースに格納し、作問の際に過去問題と同じようなヒントを示せるようにする必要はある。また、分類対象についてはほとんどの問題において答えが分類対象該当したが (96 問中 81 問)、必ずしも該当するとは限らない結果となった。例えば「～は誰ですか」というような問題の場合、問題文は答えである人物名の説明文となっている。この場合、答えである人物についてが「その問題で問われている知識」となり、その人物の説明文である問題文を分類対象とし、人物名の子要素として追加する必要がある。カテゴリ分類については、問題文中の単語を参照することである程度分類可能なことが分かった。例えば「～の町名 (旧町名) は何ですか」という問題の場合は、問われている知識は必ず「町についての知識」となるため、「町」カテゴリとなる。他にも、「盛岡弁」という単語が問題文中に出現した場合は「方言」カテゴリとすることが可能である。分類対象の選定や問題文中の単語参照によるカテゴリ分類については、独自ルールが必要となることが判明した。

人手で構築した知識ベースについて、カテゴリ数 (クラス数) とインスタンス数、及びツリーの深さを示したものを表 2 に示す。

### 5.3 知識ベースの要件定義

5.2 節より、本研究で構築する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
- 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 (意味的な繋がり) を付与する
- 答えを導くためのヒント (付加情報) を付与する
- 分類対象を答えのみに固定しない

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。

### 5.4 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 9 に示す。5.2 節、5.3 節を元に、我々は以下の 3 つのモジュールを作成し、知識ベースの自動構築を行うこととした。

表 3 知識ベースの構成要素

構成要素	説明
ルート	<ul style="list-style-type: none"> <li>知識ベースのルートとなる要素</li> <li>今回は盛岡についての知識ベースなので「盛岡」となる</li> </ul>
クラス	<ul style="list-style-type: none"> <li>「その問題で問われている知識」とルートとの関連を表す問題のカテゴリ</li> </ul>
インスタンス	<ul style="list-style-type: none"> <li>知識の分類を行うべき対象である分類対象を表したもの</li> <li>人名などの固有名詞もインスタンスとして扱う</li> </ul>
主語 (リソース)	<ul style="list-style-type: none"> <li>「問題で問われている知識」を表す。基本的にクラスに該当するが、主語が固有名詞の場合はインスタンスとして扱う</li> <li>主に「(盛岡の) 橋」や「石川啄木」のような人名などが主語に値する</li> <li>主語を元にカテゴリ分類を行う</li> </ul>
述語 (プロパティ)	<ul style="list-style-type: none"> <li>クラス (カテゴリ) とインスタンス (分類対象) の関係性を表したものの</li> </ul>
目的語 (プロパティの値)	<ul style="list-style-type: none"> <li>分類対象であるインスタンスとなる要素</li> </ul>
付加情報	<ul style="list-style-type: none"> <li>各問題で答えを導くヒントとなるような情報</li> <li>主語・述語・目的語のいずれかを補足する情報となっている</li> </ul>
付加情報フラグ	<ul style="list-style-type: none"> <li>付加情報がトリプルの中のどの要素の情報かを表すフラグ</li> </ul>

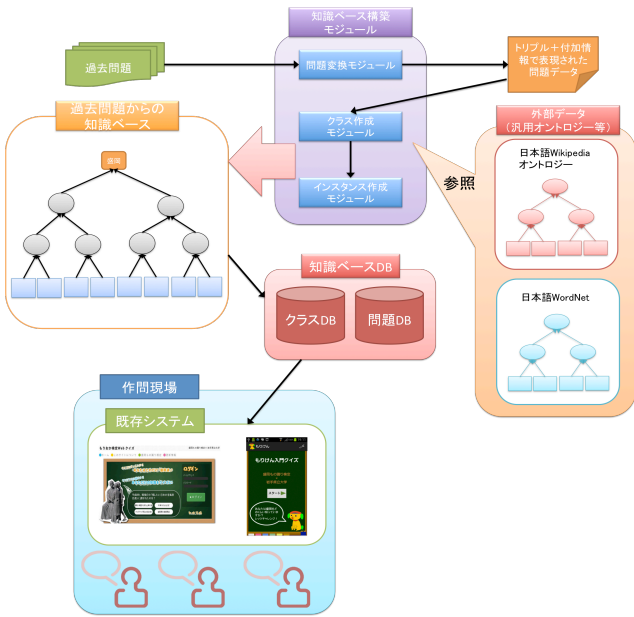


図 9 知識ベース自動構築手法の概要

- (1) 過去問題を要素別に変換する問題変換モジュール
- (2) 問題データにカテゴリを表したクラスを追加するクラス作成モジュール
- (3) クラスの実データとなるインスタンス作成モジュール

以上のモジュールを用いて過去問題から知識ベースを構築し、知識ベースのデータベースを作成する。そのデータベースで、実際のもりけん作問現場において問題の自動生成による作問支援を行えるようにすることを狙う。以下で各モジュールについての詳細な機能を述べる。

#### 5.4.1 問題変換モジュール

まず、各問題から問題文と正解選択肢を抽出する。抽出した情報を用いて「主語 (リソース) - 述語 (プロパティ) - 目的語 (プロパティの値)」のトリプル形式[17]に変換する。それぞれ主語が各問題で問われている知識、述語がインスタンスとクラスの関係性、目的語がその問題の分類対象となるインスタンスに該当する。更にこのトリプルに問題文から抽出した付加情報データ（問題文中にある答えを導くヒント情報）を付与する。付加情報は主語、述語、目的語のいずれかについての付加情報となるので、どの要素についての付加情報かを表す付加情報フラグも付与することとする。

以上の機能により、知識ベースを構成する「トリプル+付加情報」形式の問題データが生成される。ここで、クラス、インスタンスを含めた知識ベースの構成要素についてまとめたものを表 3 に示す。なお、「トリプル+付加情報」形式の問題を自動生成するに当たり、もりけんの問題は各問題で問題文の表現が統一されていない問題があったため、今回はもりけんの問題を「トリプル+付加情報」形式に変換しやすいように改変した。例えば人の名前を問う問題で

は、「～は誰ですか」という聞き方に統一し、建築物や名産品などの名前を問う問題は「～の名前は何ですか」という聞き方に統一した。

問題文と正解選択肢を「トリプル+付加情報」形式に変換する具体的な方法について以下で述べる。

#### (1) 主語の抽出

主語については問題文に対し、特定のルールを用いた自然言語処理を行うことで抽出することとした。このルールに対し優先度を設けることで、問題で問われている知識である主語を抽出することが可能だと考えられる。このルールについては、5.2 節での過去問題の分析により著者が独自に導き出した。ここで用いるルールの一覧と、5.2 節で分析した 96 問の問題における、各ルールを適用可能な問題数を表 4 に示す。なお、ルールの優先度については上から高い順となっている。優先度の高いルールから順番に問

表 4 主語抽出ルールと適用問題数及び優先度

番号	主語抽出ルール	適用可能問題数
1	基本情報カテゴリの辞書にマッチするワードが出現した場合、基本情報カテゴリとし、主語を「盛岡市」とする	5
2	問題文に「過去あった場所」「現在」というワードが出現した場合、答えを主語とし、述語を「過去あった」とする	3
3	問題文に建築物の名詞と「過去あった」というワードが出現した場合、文中に出現する建築物の名詞を主語にする	2
4	問題文で「～は誰ですか」という聞き方をしている場合は、答え（人名）を主語とする	12
5	問題文で「～の町名（旧町名）は何ですか」という聞き方をしている場合、「（盛岡の）町」を主語とする	3
6	問題文に「（名詞）の名前（別名）は何ですか」という聞き方をしている場合、名詞の部分の主語とする	36
7	問題文に「盛岡弁」という言葉が出現した場合は、方言カテゴリとし、主語を「」で囲まれている方言とする	3
8	問題文に「歌碑」という言葉が出現した場合は「歌碑」を主語にする	2
9	問題文に人名が出現する場合は、人名を主語とする	13
10	問題文にスポーツ名が出現した場合はスポーツ名を主語とする	2
11	問題文に行事名が出現した場合はその行事名を主語とする	1
12	問題文に氏族の名前が出現した場合はその氏族名を主語とする	2
13	問題文に「盛岡市の～（名詞）」という言葉が出現した場合、名詞を主語にする	3
14	問題文で最初に出てくる名詞から「～の」「～は」「～が」の助詞までを主語にする	9

題文に当てはめ、ルールとマッチした時点でそのルールに基づき、主語を確定する。なお、表 4 中の基本情報カテゴリとは、地域（盛岡）の基本情報を表すカテゴリとなっている。ここには面積や人口、隣接市町村についての情報（問題）などが当てはまる。どのような情報（問題）が基本情報となり得るのかについては、Wikipedia にある盛岡市の記

事の Infobox の項目を参考に基本情報カテゴリ辞書を作成した。Wikipedia の各記事には Infobox と呼ばれる各記事の要約情報を提供する表が存在し、記事のタイプによって Infobox のテンプレートがある[18]。盛岡市の記事には「日本の市」テンプレートの Infobox が記載されている。

## (2) 述語の抽出

述語については、問題のカテゴリや問い方によって抽出方法が変わる。現時点では主語の前のワードを抽出する方法と、主語のワードから「～は何ですか」の間のワードを抽出する方法を考えている。例えば、「国の重要文化財に指定されている建築物の名前は何か」という問題の場合は、「建築物」が主語となり、述語は「国の重要文化財に指定されている」となる。この場合は主語である「建築物」の前のワードを抽出すれば良い。また、「盛岡文士劇の平成 25 年の時代物の演目は何ですか」という問題の場合は、「盛岡文士劇」が主語となり、述語は「平成 25 年の時代物の演目」となる。この場合は主語である「盛岡文士劇」から「～は何ですか」の間のワードを抽出すれば良い。このようにして述語を抽出出来ないかを検討中である。なお、ここでの問題文解析には、既存の形態素解析サービスやキーフレーズ抽出サービス等[19]を利用することを考えている。

## (3) 目的語の抽出

目的語は基本的には正解選択肢とする。ただし、表 4 中のルール 2 とルール 4 については、正解選択肢が主語になるので問題文中から分類対象である目的語を抽出する必要がある。その場合は、基本的に問題文中に出現する固有名詞を目的語とする。

## (4) 付加情報と付加情報フラグの抽出

付加情報の抽出については主語、述語、目的語の抽出で抽出されなかったワードとする。また、付加情報フラグについては、係り受け解析ツール[19]などを用いて、付加情報となるワードが何に対し係り受け関係を持っているのかで判断する。

### 5.4.2 クラス作成モジュール

クラス（カテゴリ）の作成については、主語の単語が属するカテゴリとする。例えば主語が人名の場合は人物、施設名の場合は施設のクラスを作成する。主語の属するカテゴリの判断には、保科ら[20]の先行研究を参考にし、日本語 Wikipedia オントロジーや日本語 WordNet でインスタンス参照を行い、クラス取得をすることとした。なお、日本語 Wikipedia オントロジーはインスタンス数が豊富にあるが、インスタンスの上位概念が不足している特徴がある[10]。よって、上位概念（クラス）の抽出には日本語 WordNet と組み合わせる必要があると考えられる。

### 5.4.3 インスタンス作成モジュール

インスタンス作成モジュールでは、インスタンスである目的語、クラスとの関係性を示す述語、そして付加情報と付加情報フラグをクラス作成モジュールで作成したクラス



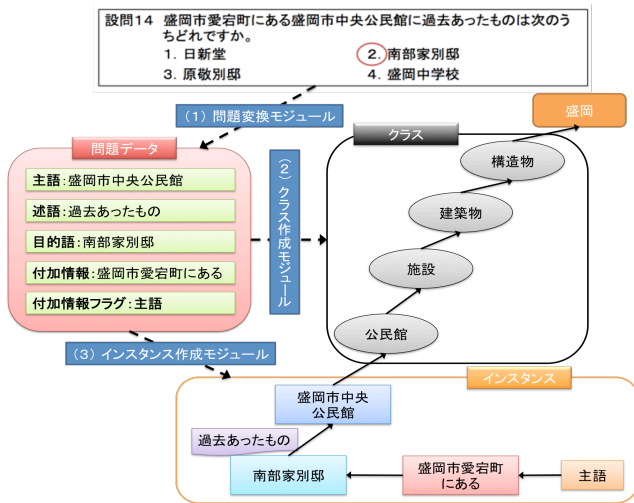


図 10 各モジュールを利用した知識ベース構築例

(主語)の直下に追加する。各モジュールを利用した知識ベース構築手法の具体例を図 10 に示す。

### 5.5 評価方法について

本研究の評価手法について検討したい。まず、作問委員会の協力を得て、過去問数問を利用し、人手で構築した知識ベースと本手法で自動的に構築した知識ベースを比較する。また、本研究で構築した知識ベースは、主に作問支援に用いることを考えているため、構築した知識ベースから実際に作問を行ってもらい、作成された問題ともりけんの過去問題との比較を行う。作問を行ってもらう人は、専門家である作問委員会のほか、本学の学生にも行ってもらう予定である。本実験により、過去問題と同じような問題を生成できれば、質の高い知識ベースの構築に成功したと言える。また、学生にも実験に参加してもらうことで、知識ベースを用いることにより専門家以外の人たちにも問題作成が可能であることも検証する。

## 6. まとめと今後の課題

本稿では、ユーザから地域に関する知識を収集するためには知識の体系化が必要であると考え、地域に関する知識ベースの構築を行うことを目的とした。そこで、ご当地検定の過去問題から知識ベースを自動生成する手法について提案した。自動生成においては、問題文と正解選択肢から「トリプル+付加情報」形式の問題データを自動生成し、外部オントロジー等を用いてクラス作成をしていくことにより、知識ベースとしていく方法を適用した。今後は評価実験を行っていく予定である。また、本研究においては現在検討段階のものも多く、以下の課題が残されている。

- 今回対象外とした画像や地図を用いた問題への対応
- クラス作成モジュールにおいて、主語のカテゴリが既存オントロジーでは抽出不可能な場合の対応
- 既存の言語解析ツールでの対応可能範囲の把握と対応不可能な場合の代替案の検討

- 付加情報の抽出方法の再検討
- 他地域・他分野（観光業など）への展開

上記の課題について検討していき、問題の自動生成に活用可能な知識ベースの自動構築手法を確立していきたい。

## 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ご感謝を申し上げます。

## 参考文献

- 1) (財)地域活性化センター: 地域の魅力を発信するご当地検査報告書(2009).
- 2) 伊藤重男: 「ご当地検定」に関する実証研究, 名古屋経済短期大学紀要, Vol.51, pp.1-13(2010).
- 3) もりおか検定 Web クイズ, <http://sakumon.jp>
- 4) 菅原遼介, 奥津翔太, 古舘昌伸, 高木正則: ユーザ参加型ご当地検定システムの開発と運用, マルチメディア, 分散, 協調とモバイル(DICOMO2012)シンポジウム(2012).
- 5) もりけん入門クイズ(Android版), <https://play.google.com/store/apps/details?id=jp.sakumon.moridai&hl=ja>
- 6) もりけん入門クイズ(iOS版), <https://itunes.apple.com/jp/app/moriken-ru-menkuizu/id665713390>
- 7) 古舘昌伸, 菅原遼介, 奥津翔太, 高木正則, 山田敬三, 佐々木淳: ユーザ参加型学習コミュニティシステムにおける知識発信支援機能の提案, 情報処理学会 情報教育シンポジウム 2012(2012).
- 8) 古舘昌伸, 菅原遼介, 奥津翔太, 井上裕之, 高木正則, 山田敬三, 佐々木淳: 地域に関する知識ベース構築に向けた知識収集方法の提案, 日本教育工学会第 29 回全国大会(2013).
- 9) 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).
- 10) 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).
- 11) Soren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary Ives: DBpedia: A Nucleus for a Web of Open Data, 6th International Semantic Web Conference, Vol.4825, pp.722-735 (2007).
- 12) DBpedia Japanese, <http://ja.dbpedia.org/>
- 13) もりけん本スーパerver.2, 盛岡商工会議所(2012).
- 14) 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).
- 15) 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).
- 16) 地元学検定『もりけん』, <http://www.ccimorioka.or.jp/jinzai/moriken.html>
- 17) RDF Primer, <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>
- 18) Help:Infobox - Wikipedia, <http://ja.wikipedia.org/wiki/Help:Infobox>
- 19) テキスト解析 - Yahoo!デベロッパーネットワーク, <http://developer.yahoo.co.jp/webapi/jlp/>
- 20) 保科宗淳, 大河原渉, 山口高平: 領域オントロジーと Linked Data を利用した観光情報推薦, 第 26 回人工知能学会全国大会, 3C2-OS-13b-8 (2012.6)