早押しクイズのパラレル問題の自動生成

橋元佐知　佐藤理史　宮田玲　小川浩平
名古屋大学大学院工学研究科
hashimoto.sachi@d.mbox.nagoya-u.ac.jp

概要

本稿では、日本語の早押しクイズの「パラレル問題」を自動生成する方法を提案する。この方法では、アノテーション済みの既存のパラレル問題を利用し、その一部の要素を置換することによって、新たな問題を作成する。置換する要素には、対比ペア、答の上位概念、問題の対象ドメイン、の 3 種類がある。これらの置換先の候補として、特定のカテゴリに対するワードリストや対比ペアリストを準備しておく。本システムを用いて、これまでに、500 問以上の新たなパラレル問題を作成した。

1　はじめに

答が一意に定まるクイズ問題は、これまで質問応答 (QA) システムの開発において、しばしば用いられてきた [1][2]。日本でも、クイズを題材にした日本語 QA データセット JAQKET[3] が開発され、これを用いた質問応答システムのコンペティションも行われた [4]。

このようなデータセットの構築のためには、大量のクイズ問題が必要となるが、良質のクイズ問題を作成することは、それほど容易ではない。問題文とその答を作ることが最初の関門であるが、それ以外にも「裏取り」と呼ばれる答の正当性・一意性を確認する作業や、問題の難易度の適切さを確認する作業が必要であり、かなりの知力と労力、経験を必要する [5]。QA システムが目指しているのは、クイズ問題を解く能力の実現であるが、クイズ問題を作るためには、それとは異なる知的能力が必要である。実際、クイズ問題作成は、ある種の創作という側面を持ち、日本ではクイズ作家とよばれる人達やクイズ制作専門会社が存在する。

コンピュータによるクイズ問題自動生成の目的は、単に QA システム開発用のデータセット構築のために留まらない。良質のクイズ問題の自動生成が可能となれば、クイズ大会やクイズ番組への問題提供など商用利用の道が開ける。さらに、教育現場での活用を想定したクイズ問題の自動生成の研究もいくつか存在する [6][7]。

本稿では、日本語の早押しクイズの「パラレル問題」を対象に、クイズ問題を自動生成する方法を提案する。

2　パラレル問題の構造と特徴

本研究が対象とする「パラレル問題」とは、

アイルランドの首都はダブリンですが、アイスランドの首都はどこでしょう？

のような、「〜は〜ですが、〜は何でしょう？」という形式の問題である。その基本構造を図 1 に示す [8]。パラレル問題は 4 つのパート (A, B, C, D) と答 (W) から構成され、各パートはそれぞれ中核要素 (X, Y, Z, Q) を持つ。これら以外にも、中核要素の上位概念や問題の対象ドメインが含まれる場合もある。本研究で扱う問題の構成要素の一覧を表 1 に示す。

パラレル問題の最大の特徴は、X と Z (あるいは、まれに Y と Q($= W$)) が、何らかの対比軸になっていることである [5]。本稿では、この対比軸の分類を対比関係と呼ぶ。この対比関係の種類は限られている (表 2)。例えば、「アイルランド (X) の首都はダブリンですが、アイスランド (Z) の首都はどこでしょう？」という問題では、「アイルランド」と「アイスランド」が音・表記の類似という観点で対比関係をなす。「日本で一番 (X) 高い山は富士山ですが、二番目 (Z) に高い山は何でしょう？」という問題では、「一番」と「二番目」が連続という観点で対比関係をなす。「アイルランド (X) の首都はダブリンですが、中国 (Z) の首都はどこでしょう？」や「日本で一番 (X) 高い山は富士山ですが、三番目 (Z) に高い山は何でしょう？」という問題は、対比関係が不自然であり、パラレル問題として不適切である。

パラレル問題のもうひとつの特徴は、前半部 (AB パート) と後半部 (CD パート) が並行的となる点である。つまり、X–Y の関係と Z–Q($= W$) の関係は、

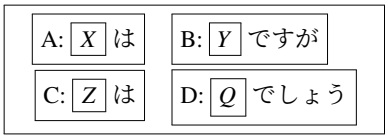

図 1　パラレル問題の基本構造

表 1　パラレル問題の構成要素

タグ	内容
X	A パートの中核要素 (対比ペアの一方)
Y	B パートの中核要素 (前半部の答要素)
Z	C パートの中核要素 (対比ペアのもう一方)
Q	D パートの中核要素 (疑問表現)
D	(問題の対象) ドメイン
x	X の (広義の) 上位概念
y	Y の (広義の) 上位概念
z	Z の (広義の) 上位概念 (通常は x と同一)
W	答
w	答 W の (広義の) 上位概念 (通常は y と同一)

同一である。例えば、「アイルランド (X) の首都はダブリン (Y) ですが、アイスランド (Z) の公用語は何 (Q) でしょう？」という問題は、「アイルランド–首都–ダブリン」と「アイスランド–公用語–アイスランド語」という関係が一致しないため、パラレル問題として不適切である。

3　問題文生成の基本戦略

前節で述べた 2 つの特徴は、パラレル問題を作る際に満たすべき制約となる。この 2 つの制約を満たす問題を作る方法として、既存のパラレル問題を利用し、その一部の要素を置換する方法を採用する。置換する要素には、いくつかの選択肢がある。

3.1　対比ペアを置換する

問題文に含まれる対比ペア (X と Z) を別の対比ペアに置換することにより、新たな問題を作ることができる。以下にその概略を示す。

1. 元となる問題を選択する。
   例：「アイルランドの首都はダブリンですが、アイスランドの首都はどこでしょう？ (答: レイキャビク)」
2. 置換元の対比ペアを抽出し、問題をフレームに一般化する。
   フレーム = X の首都は Y ですが、Z の首都はどこでしょう？ (答: W)
   X, Z = アイルランド, アイスランド
3. 置換先の対比ペアを決定する。
   X, Z = オーストリア, オーストラリア
4. 問題作成に必要な答要素 (Y と W) を決定する。
   Y = ウィーン, W = キャンベラ
5. フレームに要素を代入し、問題を完成させる。
   「オーストリアの首都はウィーンですが、オーストラリアの首都はどこでしょう？ (答: キャンベラ)」

表 2　対比関係の種類

種類	対比軸の具体例
1. 対義	大きい ↔ 小さい, 最初 ↔ 最後
2. 連続	一番目 ↔ 二番目, 春 ↔ 夏
3. 一対一対応	衆議院 ↔ 参議院, 芥川賞 ↔ 直木賞
4. 音・表記の類似	アイルランド ↔ アイスランド, 戦争と平和 ↔ 戦争と平和の法

ここで、ステップ 3 と 4 の実現がポイントとなる。

ステップ 3 では、フレームに代入できる置換先対比ペアを選択する必要がある。置換元の対比ペアと同種の対比ペア、つまり、語のカテゴリ (上記の例では「国名」) が同一であり、対比関係の種類 (「音・表記の類似」) が同一であるような対比ペアを選択すれば、フレームに代入できる可能性が高い。加えて、対比関係の種類が「対義 (『最初に』と『最後に』)」の場合は、これを「連続 (『最初に』と『次に』)」に置換することにより、問題を作れる場合もある。

ステップ 4 で答要素 (Y と W) を決定することは、「オーストリア (オーストラリア) の首都はどこでしょう」という問題を解くことに等しい。つまり、ステップ 4 を実現するには、クイズ問題を解く方法を適用すればよい。

3.2　対比ペア以外の要素を置換する

対比ペア以外の要素 (y, D) の置換によっても、新たな問題を作ることができる。以下にその概略を示す。

1. 元となる問題を選択する。
   例：「日本で最も大きいトンボはオニヤンマですが、最も小さいトンボは何でしょう？ (答: ハッチョウトンボ)」
2. 置換元要素を選択し、問題をフレームに一般化する。
   フレーム = 日本で最も大きい y は Y ですが、最も小さい w(= y) は何でしょう？ (答: W)
   y = トンボ
3. 置換先要素を決定する。
   y = カマキリ
4. 問題作成に必要な答要素 (Y と W) を決定する。

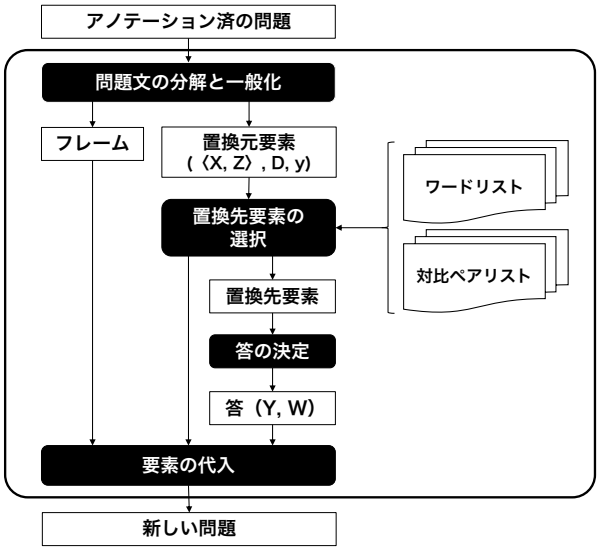

図 2　パラレル問題自動生成システムの構成

Y = オオカマキリ, W = ヒナカマキリ

5. フレームに要素を代入し、問題を完成させる。
「日本で最も大きいカマキリはオオカマキリですが、最も小さいカマキリは何でしょう (答: ヒナカマキリ)

基本的な流れは、対比ペアの置換の場合と同じであり、置換元要素の選択 (ステップ 2) と置換先要素の決定 (ステップ 3) のみが異なる。ステップ 3 では、置換元の語と同じカテゴリの語を選択すると、フレームに代入できる可能性が高い。

なお、上記の例では、答要素 Y の上位概念である $y(= w)$ を置換したが、ドメイン D の置換 (「日本 → 世界」) によっても、新しい問題を作ることができる。

3.3　自動化への準備

上記の方法を自動化するために、次の準備を行う。

1. 追加アノテーション
既存のパラレル問題には、すでに構成要素 (表 1) がアノテーションされているが [8]、それに加えて、各要素に国名などのカテゴリ情報を付与する。さらに、対比ペアには、対比関係の種類の情報を付与する。
2. ワードリスト・対比ペアリストの作成
付与したカテゴリに対して、そのカテゴリに属する語を集めたリストや対比ペアリストを作成する。

4　パラレル問題自動生成システム

作成したパラレル問題自動生成システムの全体像を図 2 に示す。本システムの入力はアノテーション済みのパラレル問題であり、前節に示した流れに沿って、新たなパラレル問題を作成・出力する。以下では、置換先の候補となる、ワードリストと対比ペアリストの作成法、および、答要素 (Y, W) の決定法を説明する。

4.1　ワードリストの作成

問題の構成要素 (X, Z, D, y) のカテゴリに対するワードリストの作成には、次の 2 種類の方法を使用した。ワードリストに収録する語は、原則として、日本語版 Wikipedia の記事タイトルに限定した。

1. **記事冒頭文を利用した同一カテゴリ語の収集**
Wikipedia 記事の最初の 1 文 (記事冒頭文) が類似していたり、その文の最後の名詞句が同一であれば、それらの記事のタイトルは同一カテゴリに属する可能性が高い。この性質を利用して特定のカテゴリに属する語を収集する。具体的には、既存のパラレル問題から特定のカテゴリに属する語の具体例を収集し、それを出発点として、そのカテゴリに属する語を収集する。記事冒頭文の類似性判定には、TF-IDF の重み付き Bag-of-Words のコサイン類似度を用い、値が 0.65 以上である場合に、同一カテゴリに属すると判定する。
2. **「○○の一覧」ページからの抽出**
Wikipedia の「○○の一覧」ページでまとめられている項目を収集する。

作成したワードリストの語に対して、ローマ字読み（訓令式）を追加した。また、「略称」「人物の愛称」などのカテゴリでは、「機関・機構の略語」「歴史上の人物の幼名」のような下位のカテゴリを設け、各語に正式名称等の情報を追加した。

対比ペアを構成する要素 (X, Z) では 78 種類のカテゴリが存在するが、これまでに 11 種類のカテゴリのワードリストを作成した。それ以外の要素 (D, y) では、13 種類のカテゴリ中、2 種類のカテゴリのワードリストを作成した。

4.2　対比ペアリストの作成

対比ペアリストは、次の 2 種類の方法で作成した。

1. **ワードリストから作成する (音・表記の類似)**
   特定のカテゴリに対するワードリストから、音や表記が類似するペアを次の方法で収集し、そのカテゴリの対比ペアリストを作る。
   (a) 表記および音 (ローマ字読み) の編集距離が、それぞれの文字列の平均長の 1/2 を超えないペアを抽出する。
   (b) 得られたペアのリストを、音の編集距離が小さい順 (同一の場合は、表記の文字列長の差が小さい順) にソートする。
   (c) ソートされたリストの先頭から、ペアを順に取得する。ただし、既に取得したペアに含まれる語を含むペアは取得しない。
2. **既存のパラレル問題から収集する (対義・連続)**
   既存のパラレル問題から、カテゴリが「属性・程度」か「属性・順序」で、かつ、対比関係が「対義」である対比ペアを「対義ペアリスト」に、対比関係が「連続」である対比ペアを「連続ペアリスト」に収録する。この 2 つの対比ペアリストは、対比関係の種類が「対義」の問題から、「連続」の問題を生成する際に使用する。

4.3 答要素の決定

作成する問題の答要素 (Y, W) の決定には、次の 2 つの方法を使用する。以下では、答要素 Y を決定する方法を示すが、W の決定法も同様である。

1. **ワードリストに情報が存在する場合**
   ワードリストの情報をそのまま利用する。例えば、「GDP は Y の略ですが」という問題文が得られ、ワードリストに「GDP」の正式名称の情報がある場合は、その正式名称を Y として採用する。
2. **検索エンジンを利用する**
   Yahoo! 検索エンジンで、得られた問題文の A パートを完全一致で検索する。検索結果の上位 10 件のスニペットを取得し、(1) 入力した問題文の直後から句読点まで、(2) 句読点から入力した問題文の直前まで、(3) 入力問題文以外の太字部分、を抽出して、答候補とする。次に、得られた答候補を、問題文の A パートの末尾に追加して、完全一致で再び検索を行い、最もヒット数の多かった答候補を Y とする。

表 3 生成したパラレル問題数

対比ペアの置換による生成				
対比関係の種類	カテゴリ		生成数	
(a) 対義 → 連続	-	(26)	88	(4)
(b) 音・表記の類似	地名・日本	(46)	92	(1)
	書物	(26)	58	(3)
	国名	(16)	55	(2)
	略称	(57)	51	(3)
	人・スポーツ	(52)	46	(1)
	地名・外国	(47)	32	(1)
	スポーツ	(26)	21	(1)
	愛称	(19)	19	(1)
	その他	(60)	53	(3)
合計			515	(20)
対比ペア以外の置換による生成				
対比関係の種類	カテゴリ		生成数	
(a) 要素 D の置換	地域名	(25)	32	(2)
(b) 要素 y の置換	生物名	(18)	4	(1)
合計			36	(3)

カテゴリの括弧内の数字は、そのカテゴリの対比ペアリストまたはワードリストに要素数を表す。生成数の括弧内の数字は、生成で利用した既存の問題の数を表す。

5 生成例と考察

本システムで実際に作成できたパラレル問題の数を表 3 に、生成例を付録 A に示す。

表 3 からわかるように、今回取り組んだカテゴリの多くでは、少数の既存の問題から、数十問の新しい問題を生成することができた。より多くのカテゴリに対して、ワードリストや対比ペアリストを準備すれば、さらに多くの問題を生成することができる。

ただし、表 3 は正しく Y や W を求められた問題の生成数であり、システムが出力した問題の 4 割程度が不適切な問題であった。不適切な問題には、問題の答が存在しないか、複数存在するために、誤った Y や W が出力されたものが多かった。例えば、フレーム「X(国名) の国の鳥は Y だ」を利用して生成された問題文

> スロバキアの国の鳥は Y ですが、スロベニアの国の鳥は何でしょう？ (答: W)

では、スロバキアとスロベニアがどちらも国鳥を定めていないため、答を求めることができず、誤った Y や W が出力された。

今後は、より多くのカテゴリやフレームを利用した問題の生成や、適切な問題の生成精度の向上、生成された問題の評価に取り組みたい。

参考文献

[1] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building Watson: An overview of the DeepQA project. **AI Magazine**, Vol. 31, No. 3, pp. 59–79, Jul. 2010.

[2] Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan L. Boyd-Graber. Quizbowl: The case for incremental question answering. **CoRR**, 2019.

[3] 鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也. JAQKET: クイズを題材にした日本語 QA データセットの構築. 言語処理学会第 26 回年次大会発表論文集, pp. 237–240, 2020.

[4] AI 王 ～クイズ AI 日本一決定戦～, （2022-01 閲覧）. https://sites.google.com/view/nlp2021-aio/.

[5] 伊沢拓司. クイズ思考の解体. 朝日新聞出版, 2021.

[6] Naoshi Sakamoto. Automated generation of fill-in-the-blanks-type quizzes using wikipedia. **International Journal of Computer Theory and Engineering**, Vol. 9, No. 5, pp. 367–373, 2017.

[7] 史佳奥原, 雄一清, 康之田原, 昭彦大須賀. Linked data を用いた俯瞰的な多肢選択式問題自動生成手法の提案. 情報処理学会論文誌, Vol. 60, No. 10, pp. 1738–1756, oct 2019.

[8] 橋元佐知, 佐藤理史, 宮田玲, 小川浩平. 競技クイズ・パラレル問題の基本構造と文型. 言語処理学会第 27 回年次大会発表論文集, pp. 1420–1424, 2021.

A パラレル問題の生成例

A.1 対比ペアの置換による生成例

A.1.1 対比関係「対義」を「連続」に置換する生成例

元の問題	1 年で**最初に**来る二十四節気は「小寒」ですが、**最後に**来る二十四節気は何でしょう？ (答: 冬至)
生成問題	1 年で**最初に**来る二十四節気は「小寒」ですが、**次に**来る二十四節気は何でしょう？ (答: 大寒)

A.1.2 音・表記の類似ペアの置換による生成例

カテゴリ	地名・日本
元の問題	**白石市**があるのは宮城県ですが、**黒石市**があるのはどこでしょう？ (答: 青森県)
生成問題	**横浜市**があるのは神奈川県ですが、**小浜市**があるのはどこでしょう？ (答: 福井県)
カテゴリ	国名
元の問題	**アイルランド**の首都はダブリンですが、**アイスランド**の首都はどこでしょう？ (答: レイキャビク)
生成問題	**スロバキア**の首都はブラチスラバですが、**スロベニア**の首都はどこでしょう？ (答: リュブリャナ)
カテゴリ	書物
元の問題	『**東海道中膝栗毛**』を書いたのは十返舎一九ですが、『**西洋道中膝栗毛**』を書いたのは誰でしょう？ (答: 仮名垣魯文)
生成問題	『**銀河鉄道の夜**』を書いたのは宮沢賢治ですが、『**銀河鉄道 999**』を書いたのは誰でしょう？ (答: 松本零士)
カテゴリ	スポーツ
元の問題	**サッカー**のプレイヤーは 1 チーム 11 人ですが、**フットサル**のプレイヤーは 1 チーム何人でしょう？ (答: 5 人)
生成問題	**サッカー**のプレイヤーは 1 チーム 11 人ですが、**ブラインドサッカー**のプレイヤーは 1 チーム何人でしょう？ (答: 5 人)
カテゴリ	略称
元の問題	楽譜に書かれる記号で、D.C. といえばダカーポですが、D.S. といえば何でしょう？ (答: ダルセーニョ)
生成問題	楽譜に書かれる記号で、p といえばピアノですが、mp といえば何でしょう？ (答: メゾピアノ)
カテゴリ	人物の愛称
元の問題	**豊臣秀吉**の幼名は「日吉丸」ですが、**伊達政宗**の幼名は何でしょう？ (答: 梵天丸)
生成問題	**源義経**の幼名は「牛若丸」ですが、**源義家**の幼名は何でしょう？ (答: 不動丸)
カテゴリ	人・社会
元の問題	日本の歴代総理大臣で、**義一**、**角栄**といえば名字は田中ですが、**赳夫**、**康夫**といえば名字は何でしょう？ (答: 福田)
生成問題	日本の歴代総理大臣で、**騏一郎**といえば名字は平沼ですが、**純一郎**といえば名字は何でしょう？ (答: 小泉)
カテゴリ	漢字表記
元の問題	漢字で**水球**と書く球技はもちろん水球ですが、**氷球**と書く球技は何でしょう？ (答: アイスホッケー)
生成問題	漢字で**板球**と書く球技はクリケットですが、**門球**と書く球技は何でしょう？ (答: ゲートボール)

A.2 対比ペア以外の要素の置換による生成例

カテゴリ	地域名
元の問題	**米国**で最も人口が多い都市はニューヨークですが、二番目に多い都市はどこでしょう？ (答: ロサンゼルス)
生成問題	**イギリス**で最も人口が多い都市はロンドンですが、二番目に多い都市はどこでしょう？ (答: バーミンガム)
カテゴリ	生物名
元の問題	日本で最も大きい**トンボ**はオニヤンマですが、最も小さい**トンボ**は何でしょう？ (答: ハッチョウトンボ)
生成問題	日本で最も大きい**バッタ**はショウリョウバッタですが、最も小さい**バッタ**は何でしょう？ (答: ノミバッタ)