根拠を説明可能な質問応答システムのための日本語マルチホップ QA データセット構築

石井愛[1] 井之上直也[1,2] 関根聡[1]
[1]理化学研究所 AIP [2]北陸先端科学技術大学院大学
{ai.ishii,satoshi.sekine}@riken.jp naoya-i@jaist.ac.jp

概要

推論の根拠を説明可能な質問応答システムの実現のためには，知識データを適切に活用するスキルを開発するための根拠情報を含むデータセットが必要だと考える．本稿では，クラウドソーシングにより新たに生成する 3 つ組を用いた日本語のマルチホップ QA データセット構築の枠組みを提案し，初期版のデータセットを公開する．調査分析結果から，多様な表現の質問に応じて，知識データの連鎖や数値比較等の様々なスキルが要求されることが示された．

1 はじめに

答えを導き出すために複数の情報や知識を用いて推論し，その推論の根拠を説明できる質問応答システムの実現には，構造化された知識を連鎖させながら適切に活用するスキルが重要となると考える．そのようなスキルに関連する既存のデータセットとしては，HotPotQA[1]，R4C[2]，2WikiMultiHopQA[3]等の推論の根拠となる情報を含むマルチホップ QA タスクがある．R4C は HotPotQA をベースとして，推論過程を 3 つ組の組み合わせで付与したデータセットである．R4C の 3 つ組は HotPotQA の文単位の根拠から単純に 3 つ組を抽出することでは得られず，推論に関連した箇所を適切に出力することで高い評価を得られることが示されている[2].

構造化された知識としては，Wikidata[4]，森羅プロジェクト[5]の Wikipedia 構造化データ[6]等，3 つ組形式で利用可能な知識データがある．これらの知識を活用するスキルを開発するためのデータセットとしては，2WikiMultiHopQA のように知識データの 3 つ組を用いて質問を生成することが考えられる．

質問:
油彩画『モナ・リザ』が所蔵されている美術館の最寄り駅は？
答え:
パレ・ロワイヤル＝ミュゼ・デュ・ルーヴル駅
根拠:
(モナ・リザ, 所蔵されている美術館, ルーヴル美術館)
(モナ・リザ, 絵画の種類, 油彩画)
(ルーヴル美術館, 最寄り駅, パレ・ロワイヤル＝ミュゼ・デュ・ルーヴル駅)

図 1　構成質問の例

ただし，その質問を解く際に用いる知識データが質問作成時に用いたものと同一の場合，その知識データを活用するスキルはうまく開発されない可能性がある．また，既存の知識データでは計算機が利用しやすいよう，3 つ組の関係の記述は統一されている．そのため，作成された質問にバリエーションが生まれにくいと考える．

そこで本稿では，クラウドソーシングにより新たな 3 つ組を生成し，それをベースに質問，回答，および根拠の 3 つ組のセットを作成するデータセット構築の枠組みを提案する．クラウドソーシングによる新たな 3 つ組を用いることで，用いる知識データに依らず知識データの活用スキルを開発でき，かつ，多様で自然な表現が含まれるデータセットを生成することを狙う．本稿の貢献は以下のとおりである．

- クラウドソーシングに基づく 3 つ組を用いたマルチホップ QA データセット構築の枠組みを提案する．
- データセットの初期版および，データセット構築の枠組みを公開する[i].
- 生成したデータセットの定量的・定性的な分析結果を示す．

[i] https://github.com/aiishii/jpn_explainable_qa_dataset

2 データセットの概要

2.1 データセットの目的

本データセットでは，マルチホップ推論が必要な質問 Q と回答 A および，その根拠を 3 つ組のセットとして提供する．質問 Q から回答 A までの推論経路を導出するタスクにより，知識データを適切に活用するスキルが開発されることを目的とする．

2.2 問題の種類

2 つの 3 つ組を前提として，根拠の説明を求める質問は，3 つ組のどこか 1 つが共通しているか，並行しているかの 2 パターンと考えられる．本データセットでは，共通するエンティティをブリッジエンティティとする(1)構成問題，2 つの 3 つ組を比較する(2)比較問題の 2 種類の問題を作成する．

(1) 構成問題　2 つのページから生成された 2 つの 3 つ組 $(e,r1,e1)$と　$(e1,r2,e2)$を中心に，$e1$をブリッジエンティティとして，e，$r1$，および$r2$ を用いて質問，$e2$を用いて回答を作成する．ブリッジエンティティは，図 1 の“ルーブル美術館”に相当する．

(2) 比較問題　2 つのページに内在する 3 つ組 $(e1,r1,e2)$と $(e3,r2,e4)$ の$e2$ と$e4$を比較する質問を作成する．生年月日（例：$e1$と$e3$どちらが先に生まれたか？）や設立年，所在地や出身国（例：$e1$と$e3$の出身国は同じか？）等の情報から質問を作成する．

3 データセットの構築

本データセットは，(1)構成問題のための 3 つ組の生成，(2)3 つ組を用いた構成問題の生成，(3)比較問題の生成の 3 つのクラウドソーシングタスクにより生成する．対象とするページが人物や企業に関するページに偏らないよう，拡張固有表現[7](ENE)が Wikipedia ページに付与された分類データ[8]を使用し分布を調整する．

3.1 クラウドソーシングタスク

(1)構成問題のための 3 つ組の生成　まず，人気のページの上位[ii]から，ENE の分布が Wikipedia 全体とほぼ同等となるように対象とするページ群を選定する．その際，固有表現のハイパーリンクが含まれにくい概念の説明のページや，データセットにふさわしくないと考えられるページが含まれるカテゴリを手作業で確認し除外した．クラウドソーシングでは，ページ右側の表（Infobox）および冒頭部分（Abstract）のハイパーリンクを対象に，ページタイトルを Subject，リンク先を Object とする 3 つ組の関係を記述するタスクを実施する．　関係記述用画面（図 2）では，クラウドワーカーが関係を記述するボックスを選択すると，右側の Web ページ部分の対象箇所がハイライトされる．

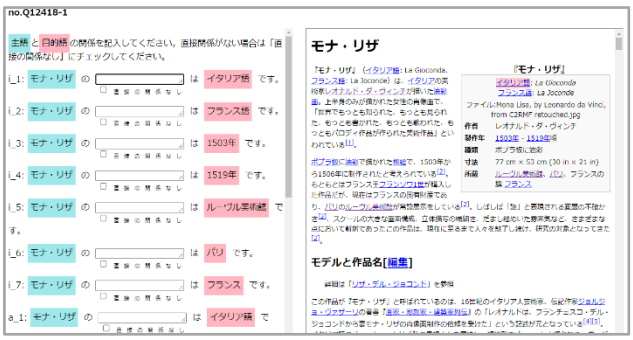

図 2　3 つ組の関係記述画面

(2) 3 つ組を用いた構成問題の生成　起点とするページの 3 つ組に含まれるブリッジエンティティの候補について，ENE の分布が調整されるよう重みづけしてランダムに選定し，ブリッジエンティティでつながる 2 つの 3 つ組のペアを生成する．その際，Infobox に含まれる情報のほうが問題生成に使われやすいという事前の調査結果から，Abstract に含まれブリッジエンティティが 6 割以上選択されるよう

図 3　構成質問生成時の根拠選択画面

[ii] 複数の Cirrussearch ダンプファイルに含まれる Popularity を平均してを用いた．

調整する．クラウドソーシングタスクでは，ブリッジエンティティでつながる 2 つのページの 3 つ組をぞれぞれ表で表示し（図 3），クラウドワーカーは 2 つの表の情報を用いて問題を作成し，使用した 3 つ組を各表から選択して根拠のセットとする．

(3)比較問題の生成　まず，対象のページ群を Wikipedia ページに付与されているカテゴリ情報でグルーピングし，ランダムにグループを選定する．そのグループ内からランダムに 2 ページのペアを作成する．その際，ENE の分布が調整されるよう重みづけをする．クラウドソーシングタスクでは，クラウドワーカーは左右に表示された 2 つのページ（図 4）を見て，質問と答えを作成し，その際使用した情報を 3 つ組の形となるよう入力する．なお，左右に 2 つのページを表示する画面では，Infobox を表示するかどうかを制御し，ランダムに 4 割程度 Infobox が表示されるようにする．

図 4　2 つのページの比較画面

3.2 データセットの統計

以下に作成したデータセットの統計を示す．クラウドソーシングは Lancers[iii]のタスク形式で実施し，問題として成立しないセットを簡易的な自動チェックおよび手作業にて構成問題 127 件，比較問題 378 件除外した．

表 1　データセットの統計

構成問題	比較問題			総計
	YES/NO	単語	計	
300	326	444	770	1070

3.3 データセットのサンプル

データセットのサンプルとして，付録表 4 に構成問題の例，付録表 5 に比較問題の例を示す．

4 分析

作成したデータセットについての調査，分析結果を示す．

4.1 問題の分野の多様性

問題作成時に対象とする Wikipedia ページの ENE カテゴリが Wikipedia 全体の分布に近くなるよう調整した結果，表 4 の分布となった．分布は多少前後するものの，人気の高いページに非常に多く含まれる人名カテゴリへの偏りはある程度避けられている．構成問題の起点ページ，ブリッジエンティティのページ，比較問題のページの ENE カテゴリの種類数はそれぞれ，46，56，58 と幅広い分野の質問が作成されたといえる．表 4 の CONCEPT 等，事前調査によりブリッジエンティティとなるような固有名詞が含まれにくいカテゴリや，「○○の一覧」等のまとめページが多く含まれる IGNORED，R18 の制限に相当するようなページが含まれるカテゴリはあらかじめ除外したため 0%となっている．

表 2　ENE カテゴリの分布（上位 10 カテゴリ）

ENE カテゴリ	Wikipedia 全体	構成問題		比較問題
		起点	ブリッジ	
人名	31.2%	39.7%	25.7%	26.3%
CONCEPT	6.1%	0.0%	0.0%	0.0%
市区町村名	5.1%	7.3%	8.7%	11.5%
音楽名	4.6%	4.0%	3.3%	8.8%
企業名	3.3%	5.0%	12.3%	6.5%
番組名	3.1%	4.7%	2.0%	6.2%
学校名	2.9%	3.7%	3.3%	1.2%
IGNORED	2.1%	0.0%	0.0%	0.0%
映画名	1.9%	4.7%	3.3%	5.8%
鉄道駅名	1.7%	1.7%	0.0%	3.6%

[iii] https://www.lancers.jp/

4.2 問題を解くために要求されるスキル

付録表 4 に示した構成問題は，クラウドソーシングで作成した 3 つ組を用いるため，その 3 つ組の関係がどのように表現されているかに質問の自然さが左右されている．

付録表 5 に示した比較問題は 1〜4 が数値を比較する問題である．数値の意味をとらえた上で比較や計算するスキルが要求される．3 は生年月日を用いて初土俵時の年齢を計算した上で比較する必要がある．4 は個数を数える問題であり，「豊富なのはどちら」という質問に対し，個数が多いほうを答える必要がある．5，6 は数値以外を比較する問題であり，5 のように片方が持つ属性値について質問し，それを持つのはどちらかを問う問題，または 6 のように両方の属性値が合致しているかどうかを問う問題がある．6 のように属性によらず「石川県」が含まれているかを問う，「ゆかりのある」という表現の意味をとらえるスキルを要求する問題も作成された．数値を比較する問題の割合は 34%であった．これは作成開始当初生年月日を比較する問題が多く作成されたため，数値の比較以外の問題を作成するインストラクションをしたことによる．

4.3 根拠の 3 つ組の特徴

Wikidata と比較した特徴　クラウドソーシングで作成された 3 つ組の一部を Wikidata と比較した結果，本研究の 3 つ組のほうが関係の数は多い傾向があった．Wikipedia は関係の記述が統一されているため，汎用的である．対して本研究の 3 つ組は，表記ゆれが多く質問のバリエーションを高める用途には適していると考える．ただし，低品質な記述も散見されるため，改善方法の検討が必要である．

3 つ組抽出元のページ内位置　表 3 に問題の根拠として使用された 3 つ組のページ内での出現位置を示す．構成質問のブリッジエンティティはほぼ設定どおりの割合であり，比較問題ではクラウドワーカーに見せる画面で比較問題に使用しやすい Infobox を表示するかどうかをランダムにしている効果から，Abstract とそれ以外の本文をあわせた割合と比較して Infobox が 5%程度少ない結果となった．Both は Infobox と Abstract 両方に含まれる場合の割合であり，other は 3 つ組の Object がページ内の表現と完全一致しない割合である．比較問題の 3 つ組はクラウドワーカーがページ内の該当箇所をコピーして Object エンティティとし，関係を自由記述するようインストラクションに記載しているが，ページ内の不要な情報を削除したケースや表 5 の 5 例のように，ページ内で存在しない情報を 3 つ組にしたケースが含まれる．

表 3 使用された 3 つ組の出現箇所

	infobox	abstract	both	本文	other
構成	36.6%	63.4%	—	—	—
比較	24.78%	17.52%	40.24%	12.63%	4.72%

5 おわりに

知識データを適切に活用するスキルの開発を目的とした，推論の根拠を 3 つ組のセットで含む日本語のマルチホップ QA データセット構築の枠組みを提案した．調査分析結果から，幅広いカテゴリの多様な表現の質問が含まれること，その多様な表現の意味を捉えた上での知識データの連鎖や数値比較等，様々なスキルが要求されることを示した．構築は継続中であるがこれまで構築した初期版となるデータセット，および構築のための枠組みとなるスクリプト等を https://github.com/aiishii/jpn_explainable_qa_dataset にて公開予定である．

今後の課題として，2WikiMultiHopQA にて提案されているような構成問題と比較問題を組み合わせたバリエーションや，2 つの 3 つ組から新たな関係を導いて用いる問題の追加があげられる．また，データセットの規模を大きくしていく上で品質を確保する手段についても今後検討する予定である．

謝辞

本研究は JSPS 科研費 JP20269633 および 19K20332 の助成を受けたものです．

参考文献

[1] Z. Yang *et al.*, “HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering,” *EMNLP 2018*, pp. 2369-2380, 2018, doi: 10.18653/V1/D18-1259.

[2] N. Inoue, P. Stenetorp, and K. Inui, “R4C: A Benchmark for Evaluating RC Systems to Get the Right Answer for the Right Reason,” *ACL 2020*, pp. 6740-

6750, Jul. 2020, doi: 10.18653/V1/2020.ACL-MAIN.602.

[3] X. Ho, A.-K. Duong Nguyen, S. Sugawara, and A. Aizawa, “Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps,” in *COLING 2020*, 2020, pp. 6609-6625, doi: 10.18653/v1/2020.coling-main.580.

[4] D. Vrandečić and M. Krötzsch, “Wikidata: A free collaborative knowledgebase,” *Commun. ACM*, vol. 57, no. 10, pp. 78-85, Sep. 2014, doi: 10.1145/2629489.

[5] “森羅 SHINRA - Wikipedia 構造化プロジェクト.” http://shinra-project.info/ (accessed Jan. 13, 2023).

[6] S. Sekine, A. Kobayashi, and K. Nakayama, “SHINRA: Structuring Wikipedia by Collaborative Contribution,” *AKBC-2019*, 2019.

[7] S. Sekine, “Extended Named Entity Ontology with Attribute Information,” *LREC08*, 2008.

[8] 関根聡, 安藤まや, 小林暁雄, and 隅田飛鳥, “拡張固有表現定義の更新と日本語 Wikipedia 分類データ 2019,” 言語処理学会 第26回年次大会, 2020.

A 付録

表 4　構成問題の例

no	質問	回答	根拠
1	妙高市に隣接し，市の花をユキツバキに制定している市の属する都道府県はどこ？	長野県	(妙高市，隣接自治体，飯山市), (飯山市，属する都道府県／都道府県，長野県), (飯山市，市の花，ユキツバキ)
2	イギリスのロックバンド，イニュエンドウが 1991 年 10 月 14 日にリリースしたシングルの B 面曲は？	炎のロックン・ロール	(イニュエンドウ，収録曲／シングルカット，ショウ・マスト・ゴー・オン),(イニュエンドウ,国，イギリス),(イニュエンドウ,ジャンル，ロックバンド)，(ショウ・マスト・ゴー・オン，販売単位，シングル), (ショウ・マスト・ゴー・オン, B 面，炎のロックン・ロール), (ショウ・マスト・ゴー・オン，イギリスのリリース／リリース, 1991 年 10 月 14 日)
3	金曜ナイトドラマ漂着者に出演者していた，1994 年 12 月 5 日生まれの俳優の所属事務所は？	エイベックス・グループ	(漂着者，出演者，太田奈緒), (漂着者，放送枠，金曜ナイトドラマ), (太田奈緒，生年月日, 1994 年 12 月 5 日), (太田奈緒，職業，俳優), (太田奈緒，所属事務所，エイベックス・グループ)
4	自動車ジープのメーカーの本社所在地は？	アムステルダム	(ジープ，製品種類，自動車), (ジープ，メーカー，ステランティス), (ステランティス，本社所在地，アムステルダム)
5	ガンダムシリーズでハサウェイ・ノアが主人公である作品が掲載されていた雑誌は？	月刊ガンダムエース	(ハサウェイ・ノア，登場するアニメ作品，機動戦士ガンダム 閃光のハサウェイ)，(ハサウェイ・ノア，登場する作品，ガンダムシリーズ一覧), (機動戦士ガンダム 閃光のハサウェイ，掲載誌，月刊ガンダムエース), (機動戦士ガンダム 閃光のハサウェイ，主人公，ハサウェイ・ノア)
6	林芙美子の長編小説である「めし」が映画化された際の映画監督は？	成瀬巳喜男	(めし，作者，林芙美子), (めし，分類，長編小説),(めし，監督，成瀬巳喜男), (成瀬巳喜男，職業，映画監督)

表 5　比較問題の例

no	質問	回答	根拠
1	奈良市とドバイはどちらも人口が 100 万人以上の都市ですか？	NO	(奈良市，人口，約 35 万 2000 人),(ドバイ，人口，約 331 万人)
2	小林多喜二とチャールズ・ディケンズはどちらも幼い頃に出身地から他の場所へ家族とともに移住していますが、どちらがより幼い頃にそれを体験していますか？	チャールズ・ディケンズ	(小林多喜二，初めて移住を体験した年齢, 5 歳), (チャールズ・ディケンズ，初めて移住を体験した年齢, 2 歳)
3	若乃花幹士 (2 代)と貴ノ浪貞博では、どちらがより若くして初土俵を踏んだでしょうか？	若乃花幹士 (2 代)	(若乃花幹士 (2 代)，生年月日, 1953 年 4 月 3 日),(若乃花幹士 (2 代)，初土俵, 1968 年 7 月場所)，(貴ノ浪貞博，生年月日, 1971 年 10 月 27 日), (貴ノ浪貞博，初土俵, 1987 年 3 月場所)
4	レインボーシックス シージとファイナルファンタジーXV では、対応プラットフォームが豊富なのはどちらでしょうか？	レインボーシックス シージ	(レインボーシックス シージ，対応プラットフォーム, Windows、PlayStation 4、Xbox One、PlayStation 5、Xbox Series X/S),(ファイナルファンタジーXV，対応プラットフォーム, PlayStation 4（PS4）・Xbox One)
5	佐川宣寿と新原浩朗は、国税庁長官を経験しているのは佐川宣寿ですか？	YES	(佐川宣寿，国税庁長官, 2017 年 7 月 5 日), (新原浩朗，国税庁長官，経験なし)
6	馳浩と谷本正憲はどちらも石川県にゆかりのある政治家ですか？	YES	(馳浩，職業，政治家)，(馳浩，出身地，石川県金沢市), (谷本正憲，職業，政治家), (谷本正憲，職歴，石川県知事(公選第 13・14・15・16・17・18・19 代)