JEMHopQA: 日本語マルチホップ QA データセットの改良

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

概要

説明可能な QA システム開発のための日本語マルチホップ QA データセットを改良した JEMHopQA について報告する．クラウドソースを用いた既存のプロセスに LLM を用いた作成プロセスを追加することでデータセットを拡張し，品質改善のため後処理におけるチェックの強化を行ったものである．後処理において除外・修正したエラー内容をあわせて報告する．

1 はじめに

知識と推論スキルの両方を必要とする複雑なタスクにおいて，高性能な大規模言語モデル（LLM）の高い精度が報告されている [1]．しかし，そのような LLM が QA 問題の解決に必要な知識をどの程度持っているのか，またその知識を活用する推論をどの程度正確に実行しているのかは，正確には明らかになっていない．LLM は推論の際に，どれくらいの頻度で「幻覚」のような知識に頼っているのだろうか．このような幻覚は，人間が注意深く作成した構造化知識ベース（KB）によって改善できるのだろうか？

そのような調査のためのベンチマークは，英語では，マルチホップ QA データセット [2, 3] から人間による注釈付きマルチホップ QA データセット [4, 5, 6, 7] に至るまで，ベンチマークが既に多数利用可能であるが，他の言語では稀である．

本論文では，JEMHopQA(Japanese Explainable Multi-Hop Question-Answering) を紹介する．JEMHopQA は石井ら [8] が報告した Wikipedia ベースの日本語では初となる説明可能なマルチホップ QA データセットを改良したものであり，1,179 組の質問と答え，および答えを導出するためのトリプルの形の知識を含んでいる（図 1，図 2）．このデータセットを用いることで日本語において，LLM に内在する知識の網羅性や，問題を解決するために知識を適切に運用するスキルなど，知識の観点からの評価が可能である．

質問: ルーヴル美術館が所在する都市の市長の名前は？
導出: (ルーヴル美術館, 所在地, パリ)， (パリ, 市長, アンヌ・イダルゴ)
回答: アンヌ・イダルゴ

図 1: 構成質問の例

質問: 『天空の城ラピュタ』と『となりのトトロ』の公開日が早いのは、『となりのトトロ』ですか？
導出: (天空の城ラピュタ, 公開年, 1986 年)， (となりのトトロ, 公開年, 1988 年)
回答: NO

図 2: 比較質問の例

問題作成では，自然で多様な質問を作成しつつスケーラビリティを確保するために，クラウドソースおよび LLM を用いて作成した上で，後処理による品質改善を実施した．石井ら [8] が報告した 1,070 件をもとに，2,203 件の追加，後処理にて 2,094 件の削除および 576 件の修正を行い，最終的に約 100 件増加となる 1,179 件に拡張および品質改善を行ったものである．後処理では自動チェック項目の追加および，手作業によるチェックを強化した．

本稿では，まず JEMHopQA のデータセットの仕様（2 章），次にデータセット構築（3 章）について説明し，最後に作成されたデータセットの統計と分析結果および，想定するタスクについて述べる（4 章）．想定するタスクで用いることができる評価スクリプトは JEMHopQA データセットとともに公開されている[1)]．

1） https://github.com/aiishii/JEMHopQA

2 データセットの仕様

2.1 データ設計の方針

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，根拠となる導出 D の組を提供する．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想定する．

導出の設計にはいくつかの選択肢があるが，井之上ら [4] と同様に予測された根拠の評価を高い解像度で可能とし，かつ，既存の KB（LLM 内の潜在的な KB を含む）の知識の網羅性および運用能力を評価するために適した形式として，2 つのエンティティ間の半構造化された関係（トリプル）を表す形式を採用する．

形式的には，導出 D は導出ステップ $d_i \in D$ から成り，$d_i \equiv \langle d_i^s, d_i^r, d_i^o \rangle$ のトリプル形式とする．d_i^s と d_i^o はそれぞれ主語と目的語のエンティティ（Wikipedia の記事タイトルに対応）を表し，d_i^r は主語と目的語のエンティティ間の関係を表わす名詞句（「場所」，「リリース日」など）である．本データセットの質問はマルチホップ推論が必要な質問であるため，各質問と回答のペアは常に 2 つ以上の導出ステップを伴う．

2.2 質問のタイプ

JEMHopQA には構成と比較の 2 つの質問のタイプが含まれる：

構成質問 では，2 つの導出トリプル (d_1^s, d_1^r, d_1^o), (d_2^s, d_2^r, d_2^o) がブリッジエンティティ（$d_1^o = d_2^s$ となるエンティティ）によって接続されており，質問に暗黙的に含まれているブリッジエンティティを見つけて推論する能力が求められる (図 1).

比較質問では，同じ関係 $d_1^r = d_2^r$ を持つ 2 つのトリプル (d_1^s, d_1^r, d_1^o), (d_2^s, d_2^r, d_2^o) を取得し，その関係に沿った 2 つの目的語エンティティに対する推論が求められる (図 2).

3 データセット構築

本章では，JEMHopQA の問題作成方法の詳細を述べる．データセットを作成するための Wikipedia ページ選定と，クラウドソーシングの方法は [8] と同様であり，GPT-3.5 を用いた構成問題の生成プロセス，後処理の自動チェック項目および手動チェックの観点が異なる点である．クラウドソーシングインターフェースの詳細は [8] を参照のこと．

3.1 Wikipedia ページ選定

まず，前処理として質問の作成に用いる Wikipedia ページを選定する．選定の際には，クラウドワーカーに馴染みのあるエンティティを対象とするためページビュースコアを用い，さらに質問を多様にするため拡張固有表現 (ENE)[9] カテゴリが Wikipedia 記事に付与された日本語 Wikipedia 分類データ 2019[10] を利用し，分布が近くなるように調整する．詳細は [8] を参照のこと．

3.2 構成問題の作成

構成問題は 2 つのステップで作成する：(1) 導出トリプルの作成，(2) (1) を用いた構成問題の作成．問題作成にはクラウドソーシングと GPT-3.5 の両方を使用する．

Step (1)：導出トリプルの作成　主語エンティティ（記事タイトル）と目的語エンティティ（Infobox や Abstract にあるハイパーリンク部分）の関係をアノテーションするタスクである．クラウドワーカーは，ハイライトされた目的語エンティティ部分の前後のテキストを読み，主語と目的語のエンティティのペアの間の関係をテキストボックスに関係を入力する．報酬は 1 タスクあたり 16 円である．GPT-3.5 では，クラウドワーカーと同様の情報を入力として関係を出力するプロンプトを用いた (付録図 3 上).

Step (2): 構成問題の作成　Step(1) で作成したブリッジエンティティを共有する 2 つのページ間のトリプルから質問を作成するタスクである．クラウドソーシングでは，クラウドワーカーがトリプル選択画面でトリプルを選択し，選択したトリプルを用いて作成した質問を入力画面のテキストボックスに入力する．この作業に対して 11 円の報酬が支払われた．GPT-3.5 では，あらかじめ決定された導出トリプルのペアを入力し，質問を出力するプロンプトを用いた（付録図 3 下）.

3.3 比較問題の作成

比較質問を作成するために，まず，3.1 節で得られた 1 万件の記事の中から，同じページカテゴリを持つページのペアをランダムに抽出し，候補ページのペアを作成する．その際，Wikidata[2] で同

2) https://www.wikidata.org/

じ “instance_of” カテゴリを持たないものをフィルタリングする．抽出されたペアに対して，YES，NO（図 2），OTHER（「ノートルダム大聖堂とアンヴァリッド宮殿のどちらが世界遺産か」のように回答がエンティティである場合）の 3 つの回答タイプのいずれかをランダムに割り当て，これらの質問タイプの最終的な比率が 1:1:2 になるようにする．

このタスクでは，クラウドワーカーは 2 つのページの比較画面を参照し，入力画面のテキストボックスに質問と関連するトリプルを入力する．比較問題の作成については，クラウドソーシングのみで十分なサンプルを得ることができたため，GPT を用いた作成は実施しなかった．クラウドワーカーにはこのタスクインスタンスごとに 11 円の報酬が支払われた．

3.4 データセットの後処理

上述した問題作成プロセスの結果得られたデータセットには多数のエラーが含まれるため，品質を担保するための後処理を実施する．まず必須となるエンティティの欠落などのタスクの前提のチェックを自動的に実施し，その後，著者らが手動で削除もしくは修正を実施した．石井ら [8] の後処理と異なる点として，ブリッジエンティティおよび回答となるエンティティが一意に定まるかどうかのチェックを自動チェックおよび手動チェックにて強化した点があげられる．手動チェックでは，Wikipedia では判別できなくても一意に定まらない可能性がある質問について，他の情報源も確認しながらチェックを実施した．

	構成問題		比較問題
	クラウド	GPT	クラウド
石井ら [8]	300	0	770
追加作成	140	1,895	168
自動削除	64	1,431	116
手動削除	130	271	82
手動修正	207	102	267
最終セット	246 (55.9%)	193 (10.2%)	740 (78.9%)

表 1: 後処理の統計

問題作成プロセス実施により，石井ら [8] のデータセットに加え，クラウドソーシングでは，それぞれ 140 問の構成問題と 168 問の比較問題，GPT-3.5 では 1,895 問の構成問題が作成された．このデータセットに対し，上記の後処理を実施した結果，表 1 に示した通り，比較問題は 78.9%，構成問題ではクラウドソーシング，GPT-3.5 でそれぞれ 55.9%と 10.2%のみが残った．これは構成問題の作成タスクのほうが複雑であることに起因する．また，構成問題の質問作成プロセスにおいて，クラウドソーシングではインターフェースにて質問の作成に適したトリプルを選択することができたが，GPT による方法はトリプルを選択するプロセスを含まないものだったため，自動削除数が多くなった．

手動削除でのエラー修正の例を付録表 5 に示す．手動削除でよくある GPT のエラー（10%）は，質問をマルチホップにするためにブリッジエンティティが利用されていないケースであった．例えば，(檀れい, 元配偶者, 及川光博) と (及川光博, 卒業大学・学部, 成城大学法学部) のトリプルが与えられた場合に，「檀れいの卒業校はどこですか？」という質問が作成されたケースである．クラウドソーシングと GPT の両方で見られるもう 1 つのタイプのエラー（両方の手動削除の 11%）は，2 つのエンティティ間の関係が，主語と目的語のエンティティペアを一意に決定しないケースである．例えば，「関口宏の高校の同級生である俳優が亡くなった年月日は？」という質問について，Wikipedia ページ上には同級生は一人しか書かれていない場合であっても，他にも俳優の同級生が存在する可能性があり，ブリッジエンティティを一意に特定できないケースである．

また，チェックの過程で「ルーヴル美術館が所在する都市の市長の名前は？」といった市長や代表取締役の等の人物に関する，将来的に正解が変わる可能性がある問題が一定数含まれることがわかった．これらの問題については，`time_dependent` フラグを付与し区別できるようにした．

4 データセットの統計と分析

最後に，作成されたデータセットの詳細および，このデータセットを用いた想定するタスクとそのタスクで用いることができる評価指標について述べる．

4.1 データセットの統計

表 2 に最終的なデータセットの統計を示す．比較質問は，YES/NO/OTHER (=entity) の回答タイプの比率を 1:1:2 とし，構成問題と同様にバランス

の取れた回答タイプとした．また，3.4 節で述べた time_dependent フラグがついた問題は，Train，Dev でそれぞれ 39 件，3 件であった．

	構成問題	比較問題 (YES, NO, OTHER)	ALL
Train	392	667 (174+173+320)	1,059
Dev	47	73 (22+23+28)	120
ALL	439	740 (196+196+348)	1,179

表 2: データセットの統計

4.2 回答タイプの種類

作成された構成問題は，図 1 のようにブリッジエンティティを介してエンティティを回答するか，日付のような数値を回答するかのどちらかであった．比較問題は，次の 3 つの方法のいずれかで解くことができる：(i) トリプルの数字を比較する（図 2），(ii) 2 つの主語エンティティが目的語を共有しているかどうかを調べる（例えば，”奥州市と酒田市はどちらも東北地方の都市ですか？”），または (iii) 主語エンティティのどちらが該当する目的語を持つかを調べる（例えば，川崎重工業と任天堂，本社が京都にあるのは？）．表 3 に回答のタイプの分布を示す．

	Train	Dev
構成問題 (total)	392 (37.0%)	47 (39.2%)
エンティティ回答	255 (24.1%)	32 (26.7%)
数値回答	137 (12.9%)	15 (12.5%)
比較問題 (total)	667 (63.0%)	73 (60.8%)
数値比較	297 (28.1%)	33 (27.5%)
目的語共有	208 (19.6%)	30 (25.0%)
エンティティ選択	162 (15.3%)	10 (8.3%)

表 3: 回答タイプの分布

4.3 トピックカテゴリの多様性

Wikipedia ページ選定時の ENE カテゴリ分布調整の結果，表 4 に示す通り，最終的なデータセットにおいて 11 カテゴリ[3]中で，ブリッジエンティティや比較になりにくい 3 つのカテゴリ（Disease，Deity，Color）カテゴリを除いた 8 カテゴリをカバーし，分布においても，Wikipedia のページビュー上位記事の分布である Wikipedia TopView[4]では 5 割を超える Person カテゴリへの偏りを回避し，Wikipedia 全体の分布に近づけることができた．

ENE	Wikipedia 全体	Wikipedia TopView	JEMHop QA
Person	31.16%	54.58%	39.36%
Product	24.13%	29.75%	31.17%
Facility	12.38%	0.63%	9.84%
Location	8.97%	1.94%	10.98%
Organization	8.10%	9.57%	12.21%
Event	3.15%	3.72%	0.76%
Natural Object	2.58%	0.88%	0.13%
Individual Living Thing	0.34%	0.45%	0.34%
Disease	0.23%	0.68%	0.00%
Deity	0.14%	0.05%	0.00%
Color	0.02%	0.00%	0.00%

表 4: エンティティカテゴリの分布

4.4 想定するタスク

JEMHopQA を用いたタスクとして，QA システムに，質問に対する答えと，根拠となる導出の出力を要求するタスクを想定する．具体的には，質問 Q が与えられたとき，(i) 答え A を予測し，(ii)A の根拠となる導出 D を生成するタスクである．このタスクの評価指標には，導出 D のトリプルの関係が制約のある集合に限定されていないため，類似性の許容が求められる．類似性を考慮した評価指標および，JEMHopQA を用いた GPT-4 の評価と分析結果は [11] にて報告する．

5 おわりに

本稿では，日本語マルチホップ QA 用のデータセットを改良した JEMHopQA について報告した．クラウドソーシングと LLM を用いた問題作成プロセスにより，自然で多様な質問が作成できることが示された．ただし，専門家の手作業による後処理は品質改善のために不可欠であり，この部分のスケーラビリティの確保については今後の課題である．

3) 第 2 レベルカテゴリの中で，時間カテゴリと数カテゴリを除いた 11 カテゴリを対象とした

4) https://pageviews.wmcloud.org/topviews/?project=ja.wikipedia.org の 2017 年から 2022 年の月ごとのページビュートップ 1000 ページのスコアを平均して用いた．

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] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. **arXiv preprint arXiv:2302.04023**, 2023.

[2] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[3] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop Questions via Single-hop Question Composition. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 539–554, may 2022.

[4] Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6740–6750, Online, July 2020. Association for Computational Linguistics.

[5] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[6] Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7358–7370, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[7] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 346–361, 2021.

[8] 石井愛, 井之上直也, 関根聡. 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. 言語処理学会第29回年次大会発表論文集 (NLP2023), pp. 2088–2093, 2023.

[9] Satoshi Sekine. Extended named entity ontology with attribute information. In **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)**, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[10] 関根聡, 安藤まや, 小林暁雄, 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回年次大会発表論文集 (NLP2020), pp. 1221–1224, 2020.

[11] 石井愛, 井之上直也, 鈴木久美, 関根聡. マルチホップQAの根拠情報を用いたLLMの“偽”正解の分析. 言語処理学会第30回年次大会発表論文集 (NLP2024), 2024.

付録

トリプル生成用プロンプト:TITLE を説明する文である TEXT を読んで，TEXT 内に含まれる各 WORD は TITLE の何にあたるかの関係 RELATION をそれぞれ答えてください。
- 各 WORD は TARGET を改行で区切ったものです。
- 各 RELATION は，端的な表現で答えてください。
- TITLE と WORD が直接関係を持たない場合は”なし”と答えてください。
- WORD に対応する RELATION を EXAMPLE の OUTPUT に示す JSON 形式で出力してください。
—
EXAMPLE:
TITLE：トレイシー・ウィルソン
TEXT: トレイシー・ウィルソンは，カナダのケベック州ラシーン出身の女性フィギュアスケート選手。…(略)
TARGET： カナダ 出身 フィギュアスケート選手 …(略)
OUTPUT:{”カナダ”:”出身”, ”出身”:”なし”, ”フィギュアスケート選手”:”職業”, …(略)}

構成問題生成プロンプト: 以下の指示に従って質問を作成してください。
- TITLE，ATTR1，ATTR2，ANSWER の情報を使って，ANSWER が回答になる質問 QUESTION を作成してください。
- QUESTION はできる限り自然な文で作成してください。
- TITLE が書籍・雑誌・新聞や作品，映画のタイトルの場合は TITLE を『』でかこってください。
- TITLE に () が含まれる場合，例えば”ナヨン (TWICE)”であれば、”TWICE のナヨン”等と () 内の自然な表現に書き換えてください。
- 出力は QUESTION のみを答えてください。
—EXAMPLE:
TITLE：花田虎上　ATTR1：弟　ATTR2：得意技 ANSWER: 上手投げ QUESTION：花田虎上の弟の得意技は何ですか？
TITLE：魔女の宅急便 (1989 年の映画)　ATTR1：主題歌　ATTR2：活動期間の開始年 ANSWER:1971 年 QUESTION：1989 年の映画『魔女の宅急便』の主題歌の歌手が音楽活動を始めた年はいつですか？” …(略)

図 3: トリプル・構成問題生成プロンプトの例

対応	エラー内容	質問 (⇒修正した質問)	導出	回答
修正	ブリッジエンティティが利用されない	檀れいの卒業校はどこですか？ ⇒　檀れいの元夫が卒業したのは何大学何学部ですか？	(檀れい, 元配偶者, 及川光博), (及川光博, 卒業大学・学部, 成城大学法学部)	成城大学法学部
削除	ブリッジエンティティが複数	関口宏の高校の同級生である俳優が亡くなった年月日は？	(関口宏, 高校の同級生, 林隆三), (林隆三, 没年月日,2014 年 6 月 4 日)	2014 年 6 月 4 日
修正	回答の異表記が存在する	カップヌードル販売会社の東京本社所在地は？ ⇒　…本社所在地の区は？	(カップヌードル, 販売会社, 日清食品), (日清食品, 東京本社所在地, 東京都新宿区新宿)	新宿区
修正	比較問題形式でない	楽曲『おいでシャンプー』と『君の名は希望』の作詞者は？ ⇒　…はどちらも秋元康ですか？	(おいでシャンプー, 作詞者, 秋元康), (君の名は希望, 作詞者, 秋元康)	YES
削除	GPT の捏造	『ダイヤモンドは砕けない』の作者が直木賞を受賞したのは何年ですか？	(ダイヤモンドは砕けない, 作者, 荒木飛呂彦), (荒木飛呂彦, 受賞年,2013 年)	2013 年
フラグ	将来的に正解が変わる	ルーヴル美術館が所在する都市の市長の名前は？	(ルーヴル美術館, 所在地, パリ), (パリ, 市長, アンヌ・イダルゴ)	アンヌ・イダルゴ

表 5: 手動チェックにおける修正・削除例