# グラフベース RAG の知識構造化に関するモジュール別検討

西田典起 ♠ Rumana Ferdous Munne♠ Shanshan Liu♠ 徳永なるみ ♠
山縣友紀 ♠ Fei Cheng♢ 古崎 晃司 ♣ 松本裕治 ♠
♠ 理化学研究所 ♢ 京都大学 ♣ 大阪電気通信大学
noriki.nishida@riken.jp

## 概要

本研究では，グラフベース RAG（GraphRAG）におけるモジュール設計の重要性を体系的に分析する．文書集合から知識グラフを構築するトリプル抽出，知識グラフを知識単位に分割するコミュニティ分割，および知識を言語化するレポート生成に着目し，それぞれの設計選択が質問応答性能に与える影響を，二つの多段質問応答ベンチマークで評価した．その結果，トリプル抽出の品質が下流性能のボトルネックとなり得ること，およびコミュニティ分割では，事実の詳細さとトピック集約性のバランスが，高精度な検索と効果的な多段推論に重要であることが分かった．さらに，テンプレートベースのレポート生成は，LLM ベース手法と比べて，質問応答性能と計算効率の両面で優れていた．これらの知見を組み合わせることで，従来の GraphRAG 設定を大きく上回る性能を達成した．

## 1 はじめに

変化の速い分野や専門性の高い分野では，LLM の内部知識を常に最新に保つことは難しい [1, 2]．Retrieval-Augmented Generation（RAG）は，推論時に外部知識を参照することで，この制約を補う枠組みである．近年，RAG [3, 4, 5] の拡張としてグラフベース RAG（GraphRAG）が提案されている [6, 7, 8, 9, 10]．GraphRAG は，文書集合から `(subject, relation, object)` 形式の三つ組 (トリプル) を抽出し，それらを知識グラフとして構造化する．推論時には，質問に関連する知識をグラフから検索し，取得した構造化知識を条件として回答を生成する．これにより，文書横断的に分散していた事実が整理され，網羅的な検索や多段推論の改善が期待される．

一方で，事実性や多段推論が求められる質問応答において，GraphRAG の内部設計に関する体系的な指針は十分に確立されていない．特に，どの構成要素が性能を左右し，設計や品質が最終的な質問応答性能にどのように影響するかは明らかでない．

本研究では，GraphRAG をモジュール単位で分析し，設計選択とその品質が質問応答性能に与える影響を明らかにする．単に知識グラフの有効性を検証するのではなく，どの条件で機能し，どの構成要素が本質的かに焦点を当てる．トリプル抽出，コミュニティ分割，レポート生成に着目し，(1) トリプル抽出手法と品質，(2) 知識単位の粒度，(3) レポート生成における LLM の必要性を検討する．これらのモジュールを体系的に変更し，CDR-QA および DocRED-QA の二つの多段質問応答ベンチマークで評価した．

実験の結果，GraphRAG の有効性は，知識グラフの有無そのものではなく，抽出・分割・言語化の設計に強く依存することが分かった．まず，トリプル抽出の品質は性能の主要なボトルネックであり，ノイズの少ない高精度な抽出が安定した質問応答性能をもたらした．次に，コミュニティ分割では，適度に細かくトピック集約的な粒度が，根拠事実の検索精度を高め，多段推論を改善した．さらに，レポート生成では，テンプレートベース手法が，LLM による要約よりも高い性能と計算効率を示し，事実欠落やハルシネーションも抑制した．これらの知見を組み合わせることで，従来設定に比べて大幅な回答精度向上が得られた．本研究で用いた実装および二つの質問応答ベンチマークを公開する[1)]．

## 2 手法

**GraphRAG フレームワーク** 本研究では，図 1 に示す 7 段階からなる GraphRAG パイプラインを実装する．まず，文書から `(subject, relation,`

1) https://github.com/norikinishida/kapipe. 本論文は，TACL 採録論文 “Dissecting GraphRAG: A Modular Analysis of Knowledge Structuring for Factoid Question Answering” (Nishida et al.; to appear) の要約版である．

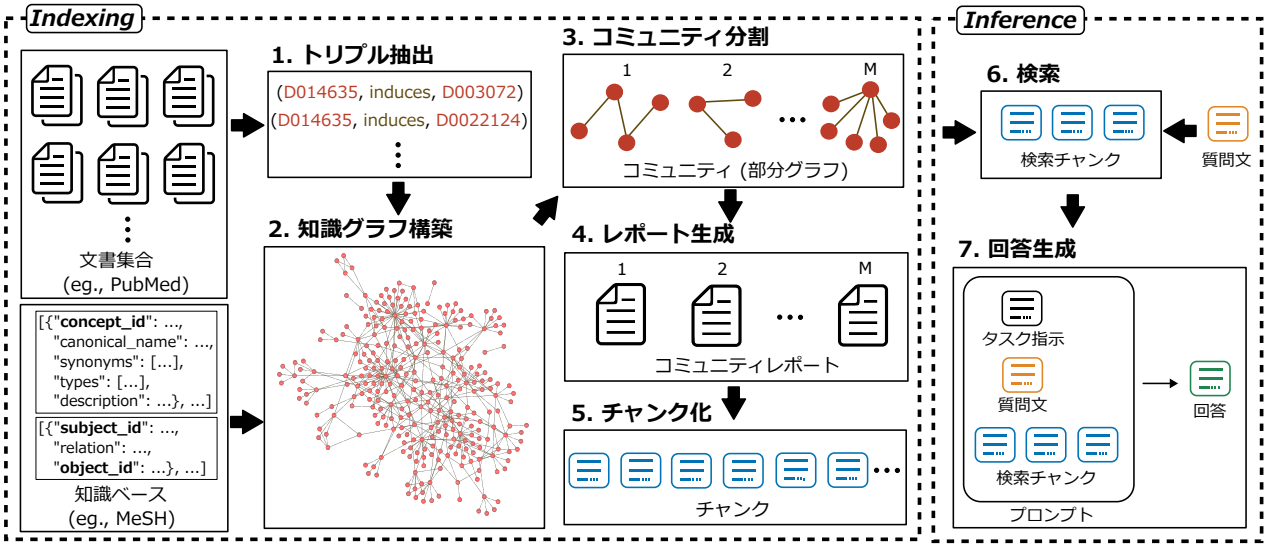


図 1: 本研究で実装する GraphRAG パイプラインの概要図.

object) 形式のトリプルを抽出し，知識ベース概念 ID (MeSH 用語や DBpedia ラベルなど) に基づいて統合することで知識グラフを構築する．次に，知識グラフを基本的な知識単位となるコミュニティに分割し，各コミュニティを自然言語レポートとして言語化する．生成されたレポートは固定長チャンクに分割・索引化され，推論時には質問に関連するチャンクを検索し，それらを条件として LLM が回答を生成する．本研究では，全工程のうち，知識の構造化・分割・言語化を直接規定する三つの主要モジュール，すなわちトリプル抽出，コミュニティ分割，レポート生成に着目する．他の構成要素は固定し，設計選択および品質が質問応答性能に与える影響を切り分けて分析する．

**トリプル抽出**　トリプル抽出は，固有表現抽出 (NER)，エンティティ曖昧性解消 (ED)，文書レベル関係抽出 (DocRE) の三段階からなり，テキスト中のエンティティ検出・正規化・関係同定を行う．本研究では，トリプル抽出において，スキーマ制約と教師あり学習への依存度が異なる二つの戦略を比較する．(1) **SLM-SFT** では，Biaffine-NER [11], BLINK [12], ATLOP [13] といったタスク特化の教師あり小規模言語モデルを用いる．これらは高精度 (precision) なトリプルを生成できる一方で，エンティティ型や関係型は学習データ [14, 15] に依存する．(2) **LLM-ICL** では，NER と DocRE を LLM (GPT-4o-mini [16]) で直接行い，ED では BLINK の Bi-Encoder で得た候補概念を LLM でリランキングする．few-shot 事例を用いた in-context learning により多様なエンティティ型や関係型に対応できる一方で，制約が弱く，ノイズや幻覚的関係を含みやすい [17, 18, 19].

**コミュニティ分割**　本研究では，粒度の異なる三つの分割手法を比較し，質問応答に適した知識単位の粒度を検討する．これらは，話題単位の粗粒度分割からトリプル単位の細粒度分割までを網羅する．(1) **Hierarchical Leiden** は，Leiden アルゴリズム [20] を再帰的に適用し，知識グラフを階層的に分割する手法である．粒度は比較的粗いが，グラフ全体を考慮した分解を行えるため，構造的ベースラインとして用いる [6]．(2) **Neighborhood Aggregation** は，各ノードとその 1 ホップ近傍からなる，細粒度かつ局所的なコミュニティを形成する手法である．エンティティ中心の関係文脈を保持し，ハイパーパラメータを持たず決定論的に適用される．(3) **Triple-Level Factorization** は，各トリプルを独立したコミュニティとして扱う最も細粒度な分割手法である．単一事実に対応するため関係文脈は乏しく，多段推論には不向きであることが予想される．

**レポート生成**　各コミュニティは自然言語レポートに変換され，構造化知識と言語モデルを接続する役割を担う．本研究では，LLM による要約生成と，テンプレートによる固定形式生成の二つの手法を比較する．(1) **LLM-based Report Generation** では，コミュニティ内のエンティティと関係を入力として，LLM (GPT-4o-mini) がタイトル，要約，知見からなる構造化要約 (JSON) を生成し，テンプレートにより平文レポートへ変換する．(2) **Template-Based**

| データセット | トリプル抽出 | Ans. Recall | Ev. Recall@10 | NER | | | ED | DocRE | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | P | R | F1 | Acc. | P | R | F1 |
| CDR-QA | Manual | **11.7** | **26.3** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| | SLM-SFT | 7.7 | 14.5 | 87.7 | 89.1 | 88.4 | 90.7 | 66.0 | 70.2 | 68.0 |
| | LLM-ICL | 4.8 | 8.4 | 60.6 | 80.8 | 69.3 | 89.5 | 38.2 | 90.2 | 53.7 |
| DocRED-QA | Manual | **18.3** | **45.1** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| | SLM-SFT | 18.2 | 33.0 | 88.2 | 88.0 | 88.1 | 85.5 | 66.5 | 58.9 | 62.5 |
| | LLM-ICL | 14.5 | 5.6 | 70.7 | 76.1 | 73.3 | 87.5 | 10.8 | 18.7 | 13.7 |

表 1: CDR-QA および DocRED-QA におけるトリプル抽出手法の比較と，NER，ED，DocRE の性能．

**Report Generation** では，知識グラフ構造に基づき，次数にもとづく中心エンティティを用いたタイトルと，全エンティティおよび関係の列挙からなるレポートを決定論的に生成する．流暢さは低いが，情報の完全性と計算効率に優れる．

**チャンク化・検索・回答生成**　各コミュニティレポートは固定長チャンクに分割して索引化され，推論時には質問に関連する上位 $k$ 件のチャンクを取得し，質問文とともに LLM (GPT-4o) へ入力して回答を生成する．

## 3　実験

**実験設定**　本研究では，GraphRAG を体系的に分析するため，生物医学分野の CDR-QA と一般分野の DocRED-QA を新たに構築した．両データセットは，複数文書に対して人手でアノテーションされたトリプルに基づく知識グラフ上で回答可能な，多段推論かつ複数正解を持つ質問を備え，質問は Neighborhood，Intersection，Multi-Hop の三種類に分類される（詳細は付録 A）．質問応答性能は，各質問について生成回答が正解エンティティ（その同義語を含む）をいくつ含んでいるかを数え，全質問で合算した割合として定義される **Answer Recall** で評価する．加えて，**Evidence Recall@k** として，各質問の回答根拠トリプルが上位 $k$ 件の検索結果に含まれる割合を測定する．特に断りのない限り，実験は GraphRAG のデフォルト設定で行う [6]：人手によるトリプル集合を用いてグラフを構築し，Hierarchical Leiden で分割し，LLM によってレポート生成したチャンクを Contriever [21] で検索し，上位 10 件を条件として GPT-4o で回答を生成する．

**結果と考察**　GraphRAG の設計構成が質問応答性能に与える影響を調べるため，文脈を用いない Direct LLM，文書チャンクを検索する Document-RAG，GraphRAG（default 設定），およびモジュール

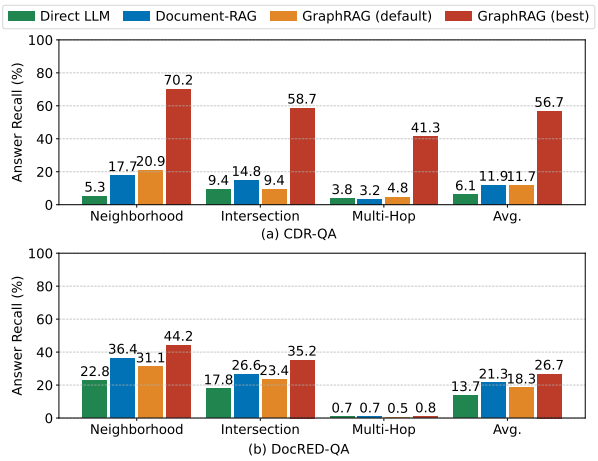


図 2: CDR-QA（上）および DocRED-QA（下）における 4 つのシステム構成の Answer Recall.

分析に基づく最良構成の GraphRAG（best）を比較した．図 2 に示すように，GraphRAG（default）は Document-RAG をわずかに下回り，知識グラフを用いるだけでは性能向上が保証されないことが分かる．一方で，最適な設計選択を組み合わせた GraphRAG（best）は大きな性能向上を示し，CDR-QA で 56.7%，DocRED-QA で 26.7%の Answer Recall を達成した．これらの結果は，GraphRAG の有効性が，知識グラフの有無ではなく，抽出・分割・言語化に関する設計選択に強く依存することを示している．

次に，トリプル抽出手法の違いが GraphRAG の性能に与える影響を調べるため，手動トリプル抽出（Manual），タスク特化 SLM による抽出（SLM-SFT），および LLM による抽出（LLM-ICL）を比較した．表 1 に示すように，他のモジュールを固定して評価した結果，手動トリプル抽出は上界を与え，自動手法の中では SLM-SFT が一貫して LLM-ICL を上回った．これは DocRE における抽出精度の差（CDR：66.0 vs. 38.2，DocRED：66.5 vs. 10.8）によるものであると考えられる．これらの結果は，GraphRAG の

| コミュニティ分割 | Answer Recall | | | Evidence Recall@10 | | | コミュニティ統計量 | |
|---|---|---|---|---|---|---|---|---|
| | Manual | SLM-SFT | LLM-ICL | Manual | SLM-SFT | LLM-ICL | コミュニティ数 | 平均ノード数 |
| **CDR-QA** | | | | | | | | |
| Hierarchical Leiden | 11.7 | 7.7 | 4.8 | 26.3 | 14.5 | 8.4 | 330 | 9.6 |
| Neighborhood Aggregation | **33.5** | **23.6** | **16.9** | **58.8** | **42.9** | **37.2** | 1,263 | 4.9 |
| Triple-Level Factorization | 24.7 | 17.5 | 11.0 | 27.2 | 17.8 | 9.7 | 2,435 | 2.0 |
| **DocRED-QA** | | | | | | | | |
| Hierarchical Leiden | 18.3 | 18.2 | 14.5 | 45.1 | 33.0 | 5.6 | 3,099 | 13.7 |
| Neighborhood Aggregation | **26.5** | **25.1** | **19.7** | **74.3** | **56.4** | **10.4** | 17,859 | 3.7 |
| Triple-Level Factorization | 22.4 | 20.4 | 19.7 | 54.8 | 39.9 | 7.7 | 29,205 | 2.0 |

表 2: CDR-QA および DocRED-QA におけるコミュニティ分割手法の比較．“Answer Recall” および “Evidence Recall@10” 下の各列は，入力グラフの構築に用いたトリプル抽出手法に対応する．また，手動構築グラフを入力としたときのコミュニティ数と平均コミュニティサイズ (ノード数) も併せて示す．

| データセット | レポート生成 | Answer Recall | | | Evidence Recall@10 | | | 総時間 [分] |
|---|---|---|---|---|---|---|---|---|
| | | HL | NA | TF | HL | NA | TF | |
| CDR-QA | LLM-based | 11.7 | 33.5 | 24.7 | 26.3 | 58.8 | 27.2 | 46.129 |
| | Template-based | **24.4** | **56.7** | **28.6** | **41.1** | **70.4** | **32.0** | 0.005 |
| DocRED-QA | LLM-based | 18.3 | 26.5 | 22.4 | 45.1 | **74.3** | 54.8 | 590.905 |
| | Template-based | **22.9** | **26.7** | **25.3** | **62.4** | 74.1 | **59.3** | 0.070 |

表 3: CDR-QA および DocRED-QA におけるレポート生成手法の比較．“Answer Recall” および “Evidence Recall@10” 下の各列は，入力コミュニティの構築に用いた分割手法（Hierarchical Leiden：HL，Neighborhood Aggregation：NA，Triple-Level Factorization：TF）に対応する．また，手動構築グラフに Hierarchical Leiden を適用して得られた全コミュニティに対する，レポート生成にかかった総処理時間（分）も併せて示す．

性能がトリプルの再現性だけでなく精度にも強く依存し，ノイズや幻覚的関係を含む抽出が下流の質問応答性能を大きく低下させることを示している．

知識グラフ分割の粒度が質問応答性能に与える影響を調べるため，粗粒度な Hierarchical Leiden，中間粒度な Neighborhood Aggregation，および極細粒度の Triple-Level Factorization を比較した．表 2 に示すように，Neighborhood Aggregation はすべてのデータセットおよび抽出設定において一貫して最良の性能を示し，Answer Recall と Evidence Recall@10 の双方で他手法を上回った．Hierarchical Leiden が少数かつ粗粒度のコミュニティを生成し，Triple-Level Factorization が多数かつ極細粒度のコミュニティを生成するのに対し，Neighborhood Aggregation は中程度の粒度を保ち，性能とのバランスが最も良好であった．この結果は，コミュニティが粗すぎると検索精度が低下し，細かすぎると推論に必要な文脈が分断される (網羅的な検索が難しくなる) ことを示しており，Neighborhood Aggregation がその両者を適切に両立することで，検索と多段推論を効果的に支援できることを示唆している．

コミュニティレポートの生成手法として，LLM による要約生成とテンプレートによる決定論的生成を比較した．表 3 に示すように，テンプレートベース手法は，ほぼすべての分割設定およびデータセットで LLM ベース手法を上回り，Answer Recall と Evidence Recall@10 の双方で高い性能を示した．さらに，処理時間も桁違いに短く，計算効率でも優れていた．この結果は，事実中心の質問応答では，コミュニティレポートの流暢さよりも，エンティティや関係を漏れなく表現することが重要であり，テンプレートベース生成が検索品質と回答性能の両立に有効であることを示している．

各モジュールの寄与については，付録 B のアブレーション実験を参照されたい．

## 4 おわりに

本研究では，GraphRAG の主要モジュール設計が質問応答性能に与える影響を体系的に分析した．GraphRAG の有効性は知識グラフの有無そのものではなく，トリプル抽出の品質，知識単位の粒度設計，および言語化方法に強く依存することを示した．

# 参考文献


[1] Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-aware language models as temporal knowledge bases. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 257–273, 2022.

[2] Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. Time waits for no one! analysis and challenges of temporal misalignment. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 5944–5958, Seattle, United States, July 2022. Association for Computational Linguistics.

[3] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[4] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: few-shot learning with retrieval augmented language models. **J. Mach. Learn. Res.**, Vol. 24, No. 1, January 2023.

[5] Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, and Lili Qiu. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. 2024.

[6] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2024.

[7] Rong-Ching Chang and Jiawei Zhang. Communitykg-rag: Leveraging community structures in knowledge graphs for advanced retrieval-augmented generation in fact-checking, 2024.

[8] Mufei Li, Siqi Miao, and Pan Li. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. 2025.

[9] Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A. Rossi, Subhabrata Mukherjee, Xianfeng Tang, Qi He, Zhigang Hua, Bo Long, Tong Zhao, Neil Shah, Amin Javari, Yinglong Xia, and Jiliang Tang. Retrieval-augmented generation with graphs (graphrag). 2025.

[10] Haoyu Han, Harry Shomer, Yu Wang, Yongjia Lei, Kai Guo, Zhigang Hua, Bo Long, Hui Liu, and Jiliang Tang. Rag vs. graphrag: A systematic evaluation and key insights. 2025.

[11] Juntao Yu, Bernd Bohnet, and Massimo Poesio. Named entity recognition as dependency parsing. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6470–6476, Online, July 2020. Association for Computational Linguistics.

[12] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6397–6407, Online, November 2020. Association for Computational Linguistics.

[13] Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. **CoRR**, Vol. abs/2010.11304, , 2020.

[14] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. **Database J. Biol. Databases Curation**, Vol. 2016, , 2016.

[15] Pierre-Yves Genest, Pierre-Edouard Portier, Elöd Egyed-Zsigmond, and Martino Lovisetto. Linked-docred - enhancing docred with entity-linking to evaluate end-to-end document-level information extraction pipelines. In **Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '23, p. 3064–3074, New York, NY, USA, 2023. Association for Computing Machinery.

[16] OpenAI. Gpt-4 technical report, 2024.

[17] Somin Wadhwa, Silvio Amir, and Byron Wallace. Revisiting relation extraction in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15566–15589, Toronto, Canada, July 2023. Association for Computational Linguistics.

[18] Yilmazcan Ozyurt, Stefan Feuerriegel, and Ce Zhang. Document-level in-context few-shot relation extraction via pre-trained language models. 2024.

[19] Kriti Bhattarai, Inez Y. Oh, Zachary B. Abrams, and Albert M. Lai. Document-level clinical entity and relation extraction via knowledge base-guided generation. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii, editors, **Proceedings of the 23rd Workshop on Biomedical Natural Language Processing**, pp. 318–327, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[20] V. Traag, L. Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. **Scientific Reports**, Vol. 9, p. 5233, 03 2019.

[21] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. 2021.

| Question Type | Question-Answer Pair |
|---|---|
| | **CDR-QA** |
| Neighborhood | **Question**: *What chemical compounds induce Defects, Intraventricular Septal?*<br>**Answer**: *Sulfasalazine*; *Aspirin*; *flumioxazin* |
| Intersection | **Question**: *What diseases are commonly induced by both Oral Contraceptives and Unfractionated Heparin?*<br>**Answer**: *Venous Thromboembolism*; *Venous Thrombosis* |
| Multi-hop | **Question**: *What chemical compounds induce the diseases that are also induced by Clonazepam?*<br>**Answer**: *Amiodarone*; *Alprazolan*; *Ammonia*; . . . |
| | **DocRED-QA** |
| Neighborhood | **Question**: *On which platform was Shadowrun Returns released?*<br>**Answer**: *Microsoft Windows*; *Linux* |
| Intersection | **Question**: *Which sports team do both Davy Jones (baseball) and Bo McLaughlin play for?*<br>**Answer**: *Chicago Cubs*; *Baltimore Orioles* |
| Multi-hop | **Question**: *What did the person who has Emily Brontë as a sibling write?*<br>**Answer**: *Villete (novel)*; *Jane Eyre* |

表 4: CDR-QA および DocRED-QA における 3 種類の質問タイプの例．下線は質問テンプレートに埋め込まれたエンティティを示す．

## A　データセット

CDR-QA および DocRED-QA 作成のための知識グラフの構築手順を述べる．各データセットについて，トリプル抽出用コーパスに付与された手動アノテーション済みトリプルを用いて，有向知識グラフを構築した．CDR-QA では，1,500 件の PubMed アブストラクトからなる CDR コーパス [14] を使用し，MeSH 用語にリンクされた化学物質・疾患の言及 (メンション) と，`Chemical-Induce-Disease` 関係を利用した．DocRED-QA では，4,051 件の Wikipedia 記事からなる Linked-DocRED コーパス [15] を使用し，DBpedia にリンクされた固有表現と，`located_in` などの関係を用いた．

次に，各知識グラフから，異なる推論パターンを評価するための三種類の質問を生成する．これらは，解に到達するために必要なグラフ走査の回数と構造が異なる．(1) **Neighborhood 質問**は，特定の関係（エッジ）を介して，あるノードに直接接続されているエンティティを問う単一ホップの質問である．例えば，ある化学物質によって誘発される疾患を尋ねる．(2) **Intersection 質問**は，同一の関係を介して二つの異なるノードの双方に接続されているエンティティを問う．例えば，二つの薬剤の両方によって誘発される疾患を尋ねる．(3) **Multi-Hop 質問**は，複数の関係からなる連鎖を通じて接続されるエンティティを問う．例えば，化学物質から疾患を介して接続される別の化学物質を尋ねる．各質問タイプについて，あらかじめ定義したテンプレートにもとづき，各データセットにつき 128 件の質問正解ペアを生成した．質問例を表 4 に示す．

| TE | CC | RG | CDR-QA | DocRED-QA |
|---|---|---|---|---|
| Manual | NA | Temp. | **56.7** | **26.7** |
| LLM | NA | Temp. | 28.0 (-28.7) | 21.1 (-5.6) |
| Manual | HL | Temp. | 24.4 (-32.3) | 22.9 (-3.8) |
| Manual | NA | LLM | 33.5 (-23.2) | 26.5 (-0.2) |
| LLM | HL | LLM | 4.8 (-51.9) | 14.5 (-12.2) |

表 5: CDR-QA および DocRED-QA におけるアブレーション結果．最良構成を基準とし，1 つのモジュールのみを他の手法に置き換えた場合の Answer Recall（%）を示す．括弧内は最良構成からの再現率低下を表す．略語：TE ＝トリプル抽出，CC ＝コミュニティ分割，RG ＝レポート生成，HL ＝ Hierarchical Leiden，NA ＝ Neighborhood Aggregation，Temp. ＝ Template-based.

## B　アブレーション実験

本節では，GraphRAG のどのモジュールが質問応答性能全体に最も大きく寄与しているのかを検討する．対象は，トリプル抽出（TE），コミュニティ分割（CC），レポート生成（RG）の三つである．このため，最も高性能であった構成を基準とし，一度に一つのモジュールのみを最も性能の低かった代替手法に置き換えるアブレーション実験を行う．表 5 に示すように，トリプル抽出またはコミュニティ分割を最も性能の低い手法に置き換えた場合，CDR-QA では Answer Recall が大きく低下し（それぞれ–28.7，–32.3），DocRED-QA でも中程度の低下が見られた（それぞれ–5.6，–3.8）．一方，レポート生成手法の変更による低下は比較的小さく，CDR-QA では–23.2，DocRED-QA では–0.2 にとどまった．これらの結果は，すべてのモジュールが GraphRAG の性能に寄与している一方で，トリプル抽出の品質とコミュニティ分割の粒度が特に重要であることを示している．