

ナレッジグラフ推論チャレンジ 2023 応募シート

1. 応募者に関する情報

- 氏名またはチーム名：KG_Honda
- 所属：本田技研工業株式会社
- メールアドレス（代表）：ko_01_uchida@jp.honda
- 応募者に学生が含まれる： いいえ
- 応募者の代表が学生である： いいえ

2. 応募部門

推理小説部門

3. 構築したナレッジグラフについて

- 構築対象としたナレッジグラフ
The Resident Patient
- 構築したナレッジグラフの基本情報
 - データサイズ
6092 トリプル、514KB
 - データ形式
RDF 形式 (.ttl ファイル)
- 構築したナレッジグラフのデータ入手先
メール添付

4. ナレッジグラフ構築に用いた「言語モデル」および「構築手法」について

- ナレッジグラフ構築に用いた「言語モデル」
GPT-4 (2023-03-15-preview)
- ナレッジグラフ構築に用いた「データ」
The Resident Patient TTL ファイル (ResidentPatient.ttl)
- ナレッジグラフの構築手法の説明
概要：ResidentPatient.ttl に含まれる動詞を抽出し、GPT-4 を用いて各動詞の同義語を生成。生成した同義語を参照元の動詞と新規のプロパティで接続し、ナレッジグラフを拡張。

構築手順：

- ① kgc:hasPredicate を述語に持つ知識トリプルの目的語（＝動詞）を抽出（動詞計 142 件）
- ② GPT-4 を用いて各動詞の同義語を生成（同義語計 1466 件）

プロンプト：

“下記の動詞の同義語をカンマ区切りで教えてください。熟語はローワーキャメルケースの英語で出力してください。¥n {動詞}”

出力例：

“find”: [discover, detect, …], “go”: [move, proceed, …], …

一部の単語(suicide, kill など)については各 LLM のフィルタリングで出力できなかったため、[INTIMIDATE Synonyms | Collins English Thesaurus \(collinsdictionary.com\)](https://www.collinsdictionary.com/ja/synonyms/INTIMIDATE) の同義語を参考に few-shot を作成し、web 版の ChatGPT で同義語を生成。

- ③ ②で生成した同義語と参照元の動詞を新規プロパティ kgc: synonym で接続した知識トリプルを、ResidentPatient.ttl の末尾に追加（知識トリプル計 1466 件）

追加した知識トリプルの例：

<http://kgc.knowledge-graph.jp/data/predicate/discover>

kgc:synonym <http://kgc.knowledge-graph.jp/data/predicate/find>.

<http://kgc.knowledge-graph.jp/data/predicate/detect>

kgc:synonym <http://kgc.knowledge-graph.jp/data/predicate/find>.

● パフォーマンス情報

・ 計算機のスペック

OS: Windows 10

プロセッサ: 12th Gen Intel(R) Core(TM) i5-1235U

動作周波数: PBF1.30GHz, EBF0.90GHz

コア数: 10

スレッド数: 12

メインメモリ: 16GB

・ 構築に掛かった時間

構築手順①（kgc:hasPredicate を述語に持つ知識トリプルの目的語抽出）：5 秒

構築手順②（各動詞の同義語生成）：426 秒

構築手順③（知識トリプルの追加）：129 秒

計：9 分 20 秒

- 参考情報

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, Xindong Wu. Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering. 2024; 1-20.

5 章“LLM-Augmented KGs”を参考に、先述した手法でナレッジグラフの構築を行った。

5. 構築したナレッジグラフの評価

- 評価方法

2 種類の大規模言語モデル(以降 LLM と表記)を利用し、「{人物名}が{動詞}したもの」を検索するためのテストデータを 2 セット作成した。その後、拡張前のナレッジグラフ (ResidentPatient.ttl)と拡張後のナレッジグラフ (ResidentPatient_rev.ttl)それぞれを用いてテストデータを含む SPARQL クエリを実行し、正答率を比較した。

■テストデータの作成手順

- ① 下記 3 つのプロパティを述語に持つシーンを抽出し、各シーンの 3 つのプロパティの目的語を格納するテストデータを作成する。

- kgc: subject
- kgc: hasPredicate
- kgc: what

抽出したシーンの例：

```
<http://kgc.knowledge-graph.jp/data/ResidentPatient/290>
  rdf:type    kgc:Situation ;
  kgc:source  "若い男は灰を叩き落とした"@ja ;
  kgc:source  "The young man dropped beating ash"@en ;
  kgc:subject <http://kgc.knowledge-graph.jp/data/ResidentPatient/Young_man> ;
  kgc:hasPredicate <http://kgc.knowledge-graph.jp/data/predicate/drop> ;
  kgc:what <http://kgc.knowledge-graph.jp/data/ResidentPatient/Ash> ;
  kgc:at_the_same_time <http://kgc.knowledge-graph.jp/data/ResidentPatient/286> .
```

作成したテストデータの例：

```
[http://kgc.knowledge-graph.jp/data/ResidentPatient/Young\_man,
http://kgc.knowledge-graph.jp/data/predicate/drop,
http://kgc.knowledge-graph.jp/data/ResidentPatient/Ash]
```

- ② 各テストデータについて 5 つずつ、2 要素目(kgc:hasPredicate の目的語)を同義語に置換したテストデータを新規に作成する。

新規に作成したテストデータの例：

[http://kgc.knowledge-graph.jp/data/ResidentPatient/Young_man,
<http://kgc.knowledge-graph.jp/data/predicate/fall>,
<http://kgc.knowledge-graph.jp/data/ResidentPatient/Ash>]

なお、②における同義語の出力には、以下 2 つの LLM を使用した。

- GPT-3.5 (gpt-35-turbo)
- Bard (gemini-pro)

プロンプト：

“下記の動詞の同義語を 5 つ、カンマ区切りで教えてください。熟語はローワーキャメルケースの英語で出力してください。¥n {動詞}

以上の手順で、利用した LLM ごとに計 852 件のテストデータを生成した。

■実行した SPARQL クエリ

```
PREFIX kgc: <http://kgc.knowledge-graph.jp/ontology/kgc.owl#>
PREFIX rp: <http://kgc.knowledge-graph.jp/data/ResidentPatient/>
SELECT DISTINCT ?what
WHERE {
  {
    ?situation kgc:what ?what;
      kgc:hasPredicate {predicate};
      kgc:subject {subject}.
  }
  UNION
  {
    {predicate} kgc:synonym ?synonym.
    ?situation kgc:what ?what;
      kgc:hasPredicate ?synonym;
      kgc:subject {subject}.
  }
}
```

クエリ内の赤太字は変数である。{subject}には各テストデータの 1 要素目 (kgc:subject の目的語)、{predicate}には 2 要素目(kgc:hasPredicate の目的語)が挿入される。クエリ実行

後、検索結果に対応するテストデータの 3 要素目(kgc:what の目的語)が含まれていたら正解とし、LLM ごと、KG ごとにテストデータの正答率を出力した。

● 評価結果

言語モデルごと、KG ごとの正答率は以下の通りである。

		kg_source	
		ResidentPatient.ttl	ResidentPatient_rev.ttl
testdata_source	GPT3.5	17.1%	71.9%
	Bard	18.1%	66.0%

GPT3.5 で同義語を出力したテストデータ、Bard で同義語を出力したテストデータともに、拡張後の KG が拡張前の KG の正答率を上回る結果となった。拡張後の KG には参照元の動詞と繋がった同義語が多く登録されているため、より多くのテストデータで正しい検索結果を出力することができたと考えられる。

何らかの動詞を用いて小説の内容を検索する際、拡張前の KG では小説内で利用されている単語を使って検索しなければ期待する出力を得ることができなかった。しかし、検索時に必ずしも同じ単語を使うとは限らない。

この問題を解決する方法として、検索時に入力されうる動詞を KG に追加で登録しておくことが挙げられるが、この作業を全て手動で行うと膨大な工数が掛かることが予想される。今回、言語モデルを活用することで検索の表記ゆれに対応した KG の構築を 10 分程度で完了させ、大幅な工数短縮を実現した。

一方、今回追加で登録した動詞は各言語モデルの知識由来のものであり、WordNet 等をソースとする同義語に比べると不確実である。より良い正答率を得るための手段として、辞書データなど外部のリソースと紐づけた言語モデルによる同義語の出力および KG 拡張が挙げられる。

6. ナレッジグラフの構築に利用したプログラム（オプション）

7. 資料の共有について

● 応募フォーム

- 公開の可否：
（ ）公開してよい
（○）非公開とする

● 応募したナレッジグラフ

- 公開の可否：
 - () 公開してよい
 - (○) 非公開とする
- 公開形式：
 - () ナレッジグラフ推論チャレンジのサイトで公開
 - () 独自のサイトで公開してリンクを希望

→公開先 URL (※)：

- **応募したプログラム等**

- 公開の可否：
 - () 公開してよい
 - (○) 非公開とする
- 公開形式：
 - () ナレッジグラフ推論チャレンジのサイトで公開
 - () 独自のサイトで公開してリンクを希望

→公開先 URL (※)：