ナレッジグラフ推論チャレンジ 2024 応募シート

１．応募者に関する情報

-氏名またはチーム名：山田俊幸、大向一輝
-所属：東京大学
-メールアドレス（代表）：to-yamada@g.ecc.u-tokyo.ac.jp
- 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：一般部門

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**

  インターネットやウェブの歴史を扱うために重要な出来事のナレッジグラフ。既存のウェブ周辺の主題を扱った年表等を共通形式に整形・統合・名寄せし、固有表現を抽出して Wikidata と紐づけたものを RDF の形式にしました。

- **構築したナレッジグラフの基本情報**
  - データサイズ
    出来事 551 件、11708 トリプル、536KB
  - データ形式
    RDF（Turtle）
  -
- **構築したナレッジグラフのデータの入手先**

  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」**

  Open AI ChatGPT 4o。「マイ GPT」の機能を利用して使用

- **ナレッジグラフ構築に用いた「データ」**

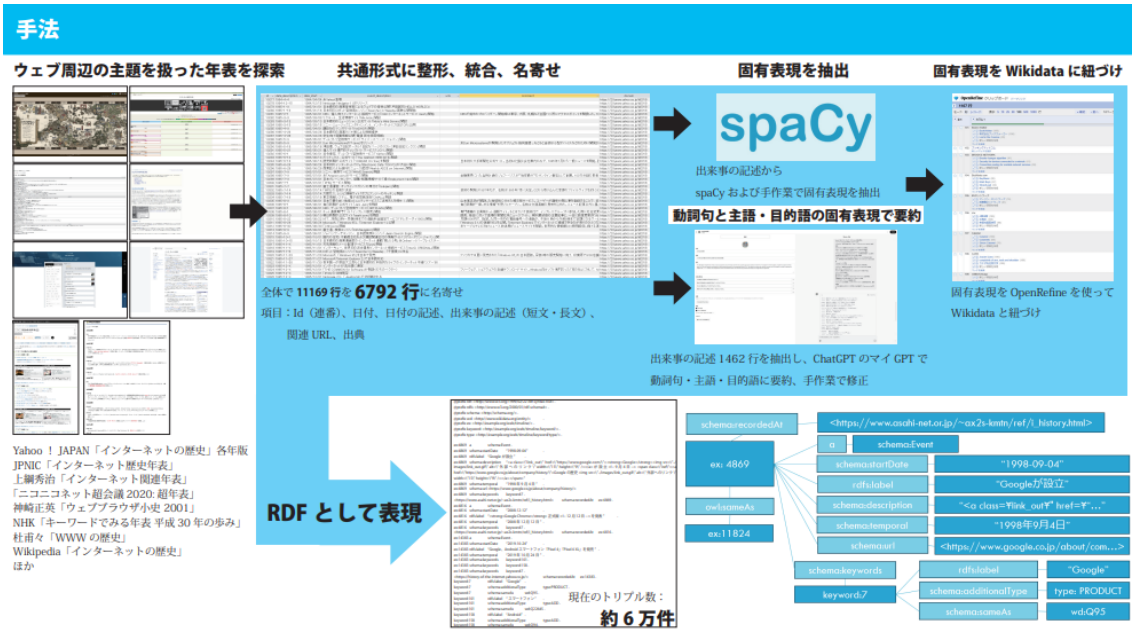
  LLM に以下のようなプロンプトを与えてマイ GPT を作成しました。

  「以下の内容を一行ずつ「S が O を V した」のような短文に要約してください。受動態は能動態に変換してください。 S は人名または団体名などの人間の集団を指し、不明な場合は「？？？」で補ってください。
  以上を S・V・O の 3 つに区切ってタブ区切りの TSV で出力してください。」

  これにすでに既存のウェブを主題とする年表等から収集していた出来事の説明テキスト 1462 行のテキストを処理させました。出来事は、件数の少ない年については全件、件数の多い年についてはランダムに選びました。

- **ナレッジグラフの構築手法の説明**

  今回のナレッジグラフは、LLM を使用しない手作業ですでに構築していたナレッジグラフの一部を拡張したものです。下記のような既存のウェブ周辺の主題を扱った年表等を収集し、共通形式に整形・統合・名寄せし、手作業や spaCy 等のライブラリによって固有表現を抽出し、固有表現を OpenRefine の照合機能を利用して Wikidata と紐づけたものを、RDF の形式に出力しています（次ページの図参照）。



図（Joint Symposium of Linked Pasts 10 and Linked Pasts Japan 1 ポスターより抜粋）

情報源

Yahoo ！ JAPAN「インターネットの歴史」
https://history-of-the-internet.yahoo.co.jp/
JPNIC「インターネット歴史年表」
https://www.nic.ad.jp/timeline/
上網秀治「インターネット関連年表」
https://www.asahi-net.or.jp/~ax2s-kmtn/ref/i_history.html
「ニコニコネット超会議 2020: 超年表」
https://dic.nicovideo.jp/a/%E3%83%8B%E3%82%B3%E3%83%8B%E3%82%B3%E3%83%8D%E3%83%83%E3%83%88%E8%B6%85%E4%BC%9A%E8%AD%B02020%3A%E8%B6%85%E5%B9%B4%E8%A1%A8
神崎正英「ウェブブラウザ小史 2001」
https://www.kanzaki.com/works/2001/pub/ua-history.html
NHK「キーワードでみる年表 平成 30 年の歩み」
https://www3.nhk.or.jp/news/special/heisei/chronology/
杜甫々「WWW の歴史」
http://www.tohoho-web.com/wwwxx018.htm
Wikipedia「インターネットの歴史」
https://ja.wikipedia.org/wiki/%E3%82%A4%E3%83%B3%E3%82%BF%E3%83%BC%E3%83%8D%E3%83%83%E3%83%88%E3%81%AE%E6%AD%B4%E5%8F%B2

（各年表の情報は 2020-2021 年に収集）

ただし、既存の複数の年表から情報を収集して構築した関係上、出来事の説明テキストについは複数の年表の記述を使用しており、権利上の懸念や表記の統一の点で課題を抱えていました。
今回上記のデータを用いて LLM を利用し、出来事（本ナレッジグラフでは schema:Action として設定。以下同じ）の主体（schema:agent）と対象（schema:object）の情報を付与し、そこから出来事の説明テキスト（rdfs:label）を改めて生成しています。このことで表記の統一が取りやすく、また権利的な懸念がなくなり公開可能となりました。

今回公開するナレッジグラフについては、主体・対象の語彙が wikidata と紐づけできたものに限っています。wikidata との紐づけにより語彙の表記の統一も効率的に行うことができました。

全体のデータは MS Access のデータベースで管理しており、クエリにより Turtle 形式に出力しています。

- **パフォーマンス情報**

  ChatGPT の制約の範囲ですが、50 行の処理に 47 秒程度かかります。

- **参考情報**

  https://kaken.nii.ac.jp/ja/grant/KAKENHI-PROJECT-22K18448/

5．構築したナレッジグラフの評価

- 今回のナレッジグラフは現在構築過程のナレッジグラフのうち公開可能な一部を切り出したものです。ナレッジグラフ全体は、日本におけるウェブの通時的変遷を捉えるべく、普及過程を調査分析し、科学技術史の一部としての研究分野の確立に向けた基礎資料群の形成と公開、ならび研究コミュニティの立ち上げを目的としており、その中で共有・修正されることを期待しています。
- ナレッジグラフの全体も現在構築過程のため知識の全体像は不詳ですが、全体の 21.3%をサンプルとして今回のアプローチで作業した結果、そのうちの 38.5%（全体の 8.2%）を今回公開できました。
- ナレッジグラフは原則として Schema.org に沿うかたちで構築しており、また出来事に関わる主要な語彙は Wikidata とリンクしています。
- LLM の利用によっても全ての行で手作業での修正を行っており、手動で構築するコストは発生しています。ただし LLM の処理によって語彙の切り出しの省略化ができ、修正点についても定形的なものが多くほとんどがまとめて処理できたことから、想定されていた作業量を半減（たとえば、1 人月を 0.5 人月程度に）することができたと考えています。
- 今回のような作業は以前から検討していましたが作業負担から後回しになっていました。構築過程で近年の LLM の普及があり、これを受けて行った今回の試行により全体の 2 割について現実的な作業時間で実施することができ、残りについても作業の道筋ができたと思われます。作業経験から最適化も期待でき、全件について同様の処理を進めていく予定です。

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）

- 独自に作成したもので公開可能なプログラムはありません。
- 年表テキストの収集や整形は随時スクリプトなどを作成。データは MS Access で複数のテーブルにより管理しています。
- ほか、語彙と wikidata との紐づけに OpenRefine の照合機能を使用しています。

７．資料の共有について

応募フォーム

- 公開の可否：
  （〇）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：
  （〇）公開してよい
  （　）非公開とする
- 公開形式：
  （〇）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
  （〇）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（※）：