地域資料における災害データの Knowledge Graph 生成手法の提案

1 はじめに

近年, 人工知能技術は急速な発展を遂げている. その中で大規模言語モデル (Large Language Model: LLM) および人間の知識をグラフ構造で表現する Knowledge Graph [1] が注目を集めている. 特に, OpenAI 社が発表した LLM をベースとした生成 AI である GPT-4o [2] はさまざまな自然言語処理タスクで優れた成果を残している. また, Knowledge Graph はさまざまな知識とそのつながりをグラフ構造として表現するデータ構造であり, 多種多様な情報とそのつながりを体系的に表現できるという利点や自然言語文のような非構造データから構造的な情報に変換して扱えるという利点がある. これらの利点は異なるデータソースからの情報の統合やそれらの間の関係性を明確にするために有効である. また, Knowledge Graph を用いることで複数のデータに含まれる知識や情報を可視化および比較して集約することが容易となる. 災害情報のような時系列や地域で変化が生じることが多いデータについても, それらからの知識抽出や集約において有効だと考えられる.

本研究では, GPT-4o を利用して各地域の資料に含まれる災害データから Knowledge Graph を生成する手法とその利用方法について提案し, 評価実験によりその有効性を検証した.

2 要素技術

2.1 Knowledge Graph

Knowledge Graph [1] とは, さまざまな知識を体系的に連結し, その関係をグラフ構造で表した知識ネットワークのことである. Knowledge Graph は, head, tail を要素に持つ entity 集合と, その entity 間の関係を表現する relation を要素に持つ relation 集合によって構成されている. 図 1 に entity をノード, relation をエッジとするグラフとして表現した Knowledge Graph の例を示す. また, Knowledge Graph の表現方法として (head, relation, tail) という 3 つ組構造（triple）の集合で表すこともできる.

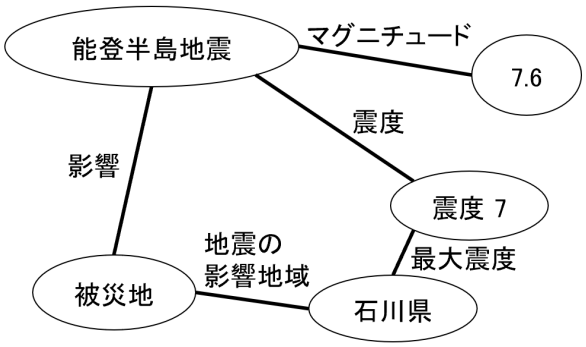


図 1: Knowledge Graph の例

Knowledge Graph は異なる種類のデータを統合的にグラフとして扱うため, 多元的な情報を同時に取得することが可能である. 例えば, 検索エンジンにおける意味検索や医療分野におけるデータ管理, レコメンデーションシステムなど, 多様に応用されている. Knowledge Graph を活用することでデータの背景にある関係性を可視化し, より効率的に情報を管理・活用することが可能である.

2.2 内閣府による災害データ

内閣府による災害データ [3] は日本国内で発生した自然災害に関する情報を網羅的に提供しているデータである. このデータには, 災害の発生日時, 場所, 規模, 被害状況, 対応措置などが含まれ, 各種報告書や統計資料が PDF 形式でまとめられている. また, このデータは時系列データとして構成されており, 災害発生日から数日もしくは数か月ごとに被害状況や復旧の進展に関する情報が随時更新される. そのため, 災害の全体像や推移を追うことが可能である.

3 提案手法

Knowledge Graph はさまざまな知識とそのつながりをグラフ構造として表現するデータ構造であり, 異なるデータセットの相互関係を明確にすることができる. 本研究では過去の種々の災害データセットを GPT-4o へ入力してその結果として Knowledge

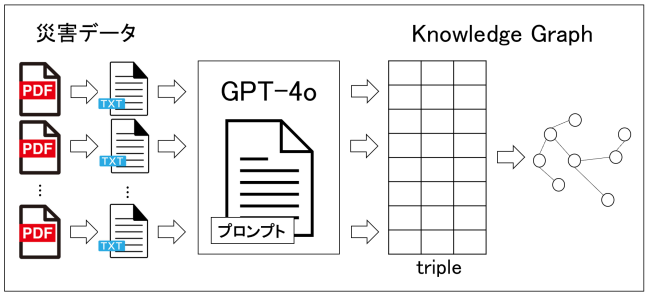


図 2: 提案手法のモデル概略図

Graph の各 entity と relation を triple 形式で抽出し, それらを統合することにより Knowledge Graph を構築する手法を提案する. この手法により, 複数の災害データを時系列的かつ相互に関連付けた知識の統合と可視化が可能となり, 災害対応や分析のための新たな知見の取得が期待される.

図 1 に提案手法のモデル概略図を示す. まず, PDF 形式で提供される災害データをテキストデータに変換し, GPT-4o へ入力する. この変換によってデータの扱いやすさが向上し, モデルにおける entity と relation の抽出が円滑になる. さらに, 各災害の発生日時や場所, 被害状況などの情報が triple として表現され, 相互関係を明示する Knowledge Graph として可視化される.

4 実験

提案手法の有効性を確認するために, 実験 1 として災害データセットからの triple 抽出, 実験 2 として抽出された triple からの Knowledge Graph 構築という 2 種類の実験をした. 以下では実験データおよび 2 種類の実験について説明する.

4.1 実験データ

実験データとして内閣府による災害データのうち, 2024 年 1 月 1 日に発生した能登半島地震の報告書を扱う. そのうち, 地震発生日の翌日である 1 月 2 日および 1 週間後の 1 月 8 日, 1 ヶ月後の 2 月 2 日, 3 ヶ月後の 4 月 2 日, 5 ヶ月後の 6 月 4 日, 7 ヶ月後の 8 月 21 日, 9 ヶ月後の 10 月 1 日時点の 7 つの報告書を使用する. このように作成日時の異なる報告書を用いることで地震の発生から復旧までの過程や被害状況の変化を時系列的に把握し, 時間の経過に伴う状況の変化を分析する.

4.2 実験 1

実験 1 では, 内閣府による災害データにおける 2024 年 1 月 1 日に発生した能登半島地震の報告書のうち, 1 月 2 日から 10 月 1 日までの報告書を対象に, GPT-4o を用いて triple を抽出した. GPT-4o への入力は入力トークン数の上限の関係上, 報告書の最初の 10 ページに限定している. triple の抽出には以下のようなプロンプト設定を使用した.

- 災害報告書からの Knowledge Graph の作成
- 重要な entity の抽出
- entity 間の関係性の推測
- triple の重要度 (1 ~ 5 の整数) の設定
- 30 個まで triple を生成

重要度について説明する. 重要度は各 triple が持つ情報の重要性を示し, GPT-4o がそれぞれの triple に応じて 1 から 5 のスコアを設定する. 1 は情報がそれほど重要でないことを示し, 5 は特に重要度が高いことを示す. この重要度設定により, Knowledge Graph 生成の際に特に重視すべき情報を優先的に把握することを可能とする.

4.3 実験 2

実験 2 では, 内閣府による災害データから実験 1 で抽出した triple を統合し, Knowledge Graph を可視化した. 実験 1 で設定した重要度について, k 番目の entity を E_k, E_k を含む triple のうち j 番目の triple を $\mathrm{T}_{\mathrm{E}_k j}$, その重要度を $I(\mathrm{T}_{\mathrm{E}_k j})$, $\mathrm{T}_{\mathrm{E}_k j}$ の個数を n とすると, E_k の重要度 $I(\mathrm{E}_k)$ は (1) 式のように表される.

$$I(\mathrm{E}_k) = \frac{\sum_{j=1}^{n} I(\mathrm{T}_{\mathrm{E}_k j})}{n} \tag{1}$$

この entity の重要度をグラフ上でノードの大きさとして表現することで重要な entity を視覚的に強調する.

5 実験結果

5.1 実験 1 の結果

表 1 に実験 1 でそれぞれの災害データから抽出した triple の総数およびその合計を示す. 表 1 より,

表 1: 抽出された triple の総数

2024 年	1 月 2 日	1 月 8 日	2 月 2 日	4 月 2 日
triple 数	20	20	20	30
2024 年	6 月 4 日	8 月 21 日	10 月 1 日	合計
triple 数	20	30	24	164

表 2: 1 月 2 日の結果

head	relation	tail	重要度
能登半島地震	地震の発生日時	発生日時	5
能登半島地震	地震の震源地	震源地	5
能登半島地震	震度 7 を観測	石川県	5
能登半島地震	津波の発生可能性	津波注意報	4
能登半島地震	ライフラインの被害	停電	4
⋮	⋮	⋮	⋮

1 月 2 日および 1 月 8 日, 2 月 2 日の triple 数は 20 で一定である一方, 4 月 2 日以降の triple 数は 20 より多くなる傾向がある. これは時間の経過に伴い災害に関する情報が追加され, 重要な情報が増えたためと考えられる.

表 2, 3, 4 に実験 1 で 1 月 2 日および 1 月 8 日, 4 月 2 日時点の報告書から抽出した triple の一例をそれぞれ示す. なお, 表中の “能登半島地震” は “令和 6 年能登半島地震” の略記である. 表 2, 3, 4 より, “能登半島地震 - 地震の発生日時 - 発生日時” のように複数の時点で共通して抽出される triple が存在するとわかった. さらに, 表 2 の “能登半島地震 - 震度 7 を観測 - 石川県” と表 3 の “石川県 - 最大震度 - 震度 7”, 表 4 の “石川県 - 地震の震度 - 震度 7” はいずれも意味的には類似しており, 能登半島地震において重要な情報である. このような意味的に類似する情報の抽出により, 同一の災害に関する要素を多角的に確認できることがわかった. また, 津波注意報の解除およびライフラインの被害, 電力や水道の復旧活動といった要素も triple として抽出されており, 災害発生から復旧までの過程を包括的に捉えられていることがわかった. 一方, 表 2, 3, 4 には “能登半島地震 - 地震の発生日時 - 発生日時” や “能登半島地震 - 地震の震源地 - 震源地” のように entity と relation の意味が重複する triple が抽出されており, これらは情報の冗長性から不必要な triple であると考えられる. また, 地震の発生日時や津波注意報の解除日時において具体的な日時を表す triple は得られなかった.

表 3: 1 月 8 日 の結果

head	relation	tail	重要度
能登半島地震	地震の発生日時	発生日時	5
石川県	最大震度	震度 7	5
津波注意報	津波注意報の解除	解除日時	3
停電	停電地域	石川県	4
ライフライン	ライフラインの復旧	復旧活動	4
⋮	⋮	⋮	⋮

表 4: 4 月 2 日 の結果

head	relation	tail	重要度
能登半島地震	地震の発生日時	発生日時	5
石川県	地震の震度	震度 7	5
石川県	地震後の復旧活動	復旧活動	4
復旧活動	電力の復旧活動	電力	4
復旧活動	水道の復旧活動	水道	4
⋮	⋮	⋮	⋮

5.2 実験 2 の結果

図 3 に実験 2 で可視化した Knowledge Graph の一部を示す. 図 3 より, 地震の発生源である “石川県” と地震の関係性が豊富に存在するように Knowledge Graph が作成されていることがわかった.

図 4 に実験 2 で可視化した Knowledge Graph の

5.3 考察

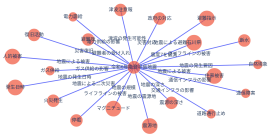
GPT-4o はテキストデータから比較的適切な triple を抽出できることがわかった. 一方, うまく関係性をとれない triple も抽出されており, このような triple が抽出されないようにプロンプトを修正する必要があると考えられる.

本研究では報告書作成時刻の異なるデータを扱ったが，時系列による違いは可視化した Knowledge Graph には反映されなかった．今後の課題として時系列を考慮できる Knowledge Graph の可視化手法を検討する必要がある.

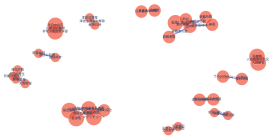
6 まとめと今後の課題

本研究では作成時点の異なる災害報告書から triple を抽出し, Knowledge Graph を可視化する手法を提案して，実験によりその有効性を確認した.

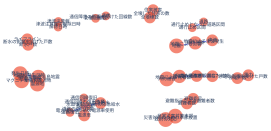
今後の課題として, より適切な triple 取得のためのプロンプト調整および時系列による Knowledge Graph の変化の可視化が挙げられる.



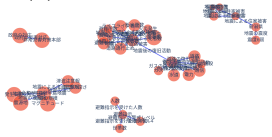
(a) 2024 年 1 月 2 日



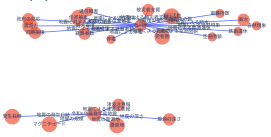
(b) 2024 年 1 月 8 日



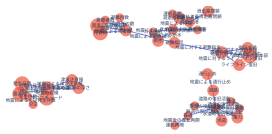
(c) 2024 年 2 月 2 日



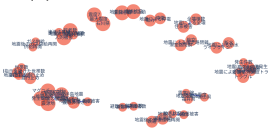
(d) 2024 年 4 月 2 日



(e) 2024 年 6 月 4 日



(f) 2024 年 8 月 21 日



(g) 2024 年 10 月 1 日

図 3: それぞれの Knowledge Graph

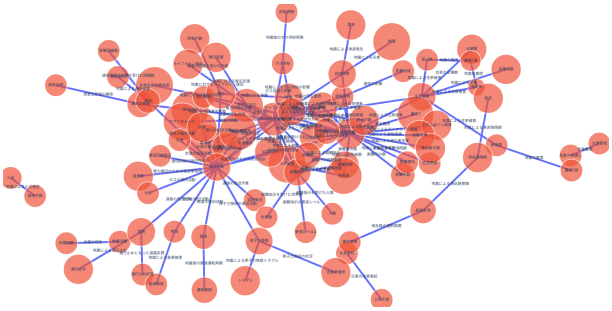


図 4: 統合した Knowledge Graph

参考文献

[1] 川村隆浩, 江上周作, 田村光太郎, 外園康智, 鵜飼孝典, 小柳佑介, 西野文人, 岡嶋成司, 村上勝彦, 高松邦彦, 杉浦あおい, 白松俊, 張翔宇, 古崎晃司. 第 1 回ナレッジグラフ推論チャレンジ 2018 開催報告 ―説明性のある人工知能システムを目指して―. 人工知能 34 巻 3 号, 2019.

[2] OpenAI. Gpt-4 technical report. 2024.

[3] 内閣府ホームページ: 災害情報. https://www.bousai.go.jp/updates/r60101notojishin/r60101notojishin/index.html.