報告書

1 今週の進捗

- 情報知識学会での質問内容
- 今後について

2 情報知識学会でのポスター発表

2.1 ナレッジグラフを利用する利点

報告書のような文章で保存すると莫大なデータ量になるが, それらをナレッジグラフとして保存することでデータ量の削減につながる. ナレッジグラフから文章を生成するタスクにも関係してくる.

時間経過による関係性の変化を見ることができれば新たな災害に過去の災害のナレッジグラフから傾向を予測することができる. 予測については検討していない.

2.2 重要度の設定について

重要度を GPT に出力させるのではなく, 自分で計算したほうがよいのではないか. 例えば, Mecab, 多次元空間で類似度を計算してその出現頻度で重要度を設定すればいい. また, ナレッジグラフの中心 (着目する情報) を変更すると重要度も変化されると面白そう. また, 県や国が出す報告書からそれぞれナレッジグラフを生成すると, 県ではこの情報を重要だと考えている, 国ではこの情報を重要だと考えている, のように重要とする情報の変化が見れたら面白そう.

2.3 その他

エンティティをカテゴリやイベント, 時系列ごとに分けて出力すれば関係性をみることができる. トリプル数の変化などもみる.

評価としてどのようにするのか. 現在は定性的な評価だが, サンプリングしてこの部分のナレッジグラフはこういうタスクに使える, のように評価できる.

“地震-地震の震源地-震源地” のようなトリプルは不必要.

2.4 小説データセット

時系列で変化するデータセットとして小説にも適応したと言うと, 浅間香織さんの小説を勧められた. データセットとして扱うことが可能かどうかは分かっていないため調査してみたい. まず小説をデータセットとして使う場合は簡単なおとぎ話から始めようと考えている.

3 今後について

ローカルの LLM を用いて文章からのエンティティ抽出およびエンティティ間の関係性予測をしたい. 以前から言っている DPO については早めに実装し, 結果をみたい. また, Neo4j についても今もっているトリプルのデータを試したい.

4　今後

- DPO のためのデータセット作成および実装
- Neo4j の実装

参考文献