令和７年度厚生労働科学研究費補助金（地域医療基盤開発推進研究事業）
分担研究報告書

【分担研究４】LLM に知識グラフを併用し情報抽出の精度を向上させる研究

田淵　貴大・宮本　太郎・川本　章太・井汲　沙織・園部　真也

【要旨】
診療記録を LLM に読解させることは情報抽出や文脈判定において有用であるが、LLM はハルシネーションによる誤答を返すことがあるという課題がある。知識グラフを併用することで LLM の誤答を抑制し、タスクによっては人間とほぼ同等の精度で情報を抽出できる。

【目的】
診療記録を LLM に読解させることは情報抽出や文脈判定において有用であるが、LLM はハルシネーションによる誤答を返すことがあるという課題がある。これは、非構造化データであるテキストの読解において柔軟性を持たせるために、ベクトルへの埋め込みという技術を用いており、そのベクトルに基づいて回答を生成する過程でランダム性が含まれることによるハルシネーションのためである。スクリーニングをはじめ、感度が重視される大まかな調査においては、このような誤答を生じても前後のワークフローでその影響を吸収できるため許容できるが、対象選定や情報抽出をはじめ、特異度が重視される厳密な調査においては、これらの誤答は致命的である。
本研究では、LLM に知識グラフを併用することでこれらの誤答を抑制する効果があるか検証する。

【方法】
東北大学病院の乳癌患者 50 名を対象とした。臨床医が診療録を読解して 24 項目の臨床情報について判断し、これをゴールドスタンダードとした。
推論用 LLM として qwen14b モデルおよび qwen7b モデルをローカルに構築し使用した。
対照群：診療録を推論用 LLM に入力し、24 項目の臨床情報について判定させた。
実験群：あらかじめ診療録を LLM（Claude v5）へ入力し、診断、薬剤、副作用、否定所見、および時系列・因果関係を含む知識グラフを、患者ごとに構築させた。この知識グラフを推論用 LLM に入力し、24 項目の臨床情報について判定させた。
対照群と実験群の判定をゴールドスタンダードと比較し、一致率を macro F1 スコアで算出した。
さらに、実験群の推論用 LLM として qwen2.5 32B base を使用して、臨床情報の項目別に、一致率を macro F1 スコアで算出した。
同様の手法を用いて、他施設（北海道大学病院、岡山大学病院、名古屋大学医学部附属病院）の乳癌患者を対象として試験的に解析した。
観察研究であり、オプトアウト方式で実施した。

【結果】
macro F1 スコアは、qwen14b モデルにおいては、対照群が 0.550 で、実験群が 0.712 であった。また qwen7b モデルにおいては、対照群が 0.595 で、実験群が 0.689 であった。また、がんのサブタイプ（HER2、TNBC 等）、ステージ、治療歴（手術、化学療法等）の項目では、qwen2.5 32B base による実験群の推論において、macro F1 スコアが 0.94 から 1.00 という極めて高い精度を達成した。
同様の傾向が各施設の解析で確認された。

【考察】
ベクトルへの埋め込みは、非構造化データを非構造状態のまま扱おうとするのに対し、知識グラフの構築は、非構造化データを構造化して扱おうとする。知識グラフに基づく LLM の推論は、ベクトルへの埋め込みに基づく LLM の推論と比べ、回答を生成する過程においてランダム性が含まれにくいため、誤答を減少させることができる。

【結論】
医療記録の読解において、LLM と知識グラフを併用することで、誤答を抑制し、LLM を単独で用いた際よりも高い精度で情報を抽出できる。タスクによっては人間とほぼ同等の精度で情報を抽出できる。

【論文発表】
該当なし

【学会発表】
該当なし

【知的財産権の出願・登録状況】
該当なし