論文 / 著書情報
Article / Book Information

題目(和文)	確信度を考慮した言語モデルの関係知識評価
Title(English)	
著者(和文)	吉川和
Author(English)	Hiyori Yoshikawa
出典(和文)	学位:博士(工学), 学位授与機関:東京科学大学, 報告番号:甲第368号, 授与年月日:2025年3月26日, 学位の種別:課程博士, 審査員:岡崎 直観,徳永 健伸,宮崎 純,村田 剛志,篠田 浩一
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Institute of Science Tokyo, Report number:甲第368号, Conferred date:2025/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名		吉川 和	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	岡崎 直観	教授	審査員	篠田 浩一	教授
	審査員	徳永 健伸	教授			
		宮崎 純	教授			
		村田 剛志	教授			

論文審査の要旨（2000 字程度）

本論文は，「確信度を考慮した言語モデルの関係知識評価」と題し，和文 5 章から構成されている．事前学習済み言語モデルの性能の向上により，言語モデルが学習時に獲得した常識や実世界の物事に関する知識を活用し，知識を要するタスクを追加の訓練なしで解くことが可能となっている．一方で，言語モデルが具体的にどのような知識を獲得し，活用可能であるかを確認することは困難であり，誤りを含む出力を用いることによるリスクがある．言語モデルの知識評価に用いられる既存のベンチマークは，Wikipedia 等の知識源をもとに作成され，実世界の知識に関する穴埋め形式の問題を言語モデルに与え，出力内容を評価することで知識の有無を間接的に評価する．しかしながら，こうした評価方法では個別事例の知識の有無を直接的に評価できないため，言語モデルの応答を抑制してハルシネーションを防止するといった用途を想定していない．本論文では，既存の知識評価ベンチマークに確信度推定の観点を導入し，誤りリスクを考慮した言語モデルの知識評価を行う研究に取り組んでいる．

第一章「序論」では，研究背景として，近年の言語モデルの発展と利用の拡大について述べた後，言語モデルの出力誤りの判別が困難であるという課題とその社会的リスクについて論じている．その後，これに対するアプローチとして，本研究で導入する確信度を考慮したモデル知識評価の概要を説明している．提案手法では既存の予測精度に基づく言語モデルの知識評価ベンチマークに選択的予測の枠組みを導入し，確信度推定に基づく予測誤りの判別能力を考慮したモデル評価を行う．また，言語モデルの入出力と内部状態に基づく確信度指標および言語モデルの訓練データを用いる確信度指標をそれぞれ設計し，性能を評価している．

第二章「準備と関連研究」では，まず，主に Transformer モデルを基礎とした近年の事前学習済み言語モデルの発展と，事前学習済みモデルの知識獲得や活用，評価に関する既存研究について述べ，その課題について議論している．次に，言語モデルの出力誤りや確信度推定に関する近年の研究について説明し，これらの関連研究の中での本研究での位置づけを説明している．最後に，本研究で用いる選択的予測について説明している．

第三章「選択的予測に基づく言語モデルの知識評価」では，既存の言語モデル知識評価ベンチマークである LAMA probe に選択的予測に基づく評価を導入し，言語モデルの入出力や内部状態を用いる確信度指標を用いた評価を行っている．選択的予測を導入する効果としては，誤りリスクの高いデータセットと低いデータセットの差異を定量的に区別できるようになることを例に挙げている．これに加え，既存の知識評価の課題であった予測やデータセットの偏りによる影響を低減する効果があることを，予測の偏りを示す指標と各評価指標との相関を比較することで確認している．確信度指標ごとの比較では，モデルの予測尤度を直接用いる指標の性能が一貫して高い一方で，各指標が特定の文脈や予測単語に対し偏った評価をする傾向があることを分析により確認している．

第四章「訓練データに基づく確信度指標」では，言語モデル出力の確信度推定として新たに言語モデルの学習に用いた訓練データを用いた方法を導入している．提案手法は予測事例と関連のある訓練データ中の事例を検索し，関連事例を用いて確信度を計算するものである．関連事例を検索するための訓練データの保存・検索方式としては，トークンレベル文脈表現，文レベル分散表現，テキスト一致に基づく手法を実装し，それぞれに対応する複数の確信度推定手法を提案し，BERT モデルを用いた実験により評価を行っている．実験の結果，提案する訓練データに基づく確信度推定方法が選択的予測の性能改善に効果的であり，特に訓練データを用いない確信度指標と組み合わせて用いることで高い効果が得られることを確認した．分析では，異なる検索方式のうち，テキスト一致検索に基づく確信度指標が最も性能への貢献が大きく，モデルの予測尤度に基づく指標の相互補完的な効果をもつことを確認した．また，効果が低かった分散表現に基づく検索による確信度推定

については，学習済み言語モデルの分散表現を用いた関連事例検索の性能が低いことが主な要因であることを確認し，検索性能向上による性能改善の可能性を示唆している．

第五章「結論」では，本論文のまとめと今後の展望を述べている．

以上の通り，本論文では，学習済み言語モデルの知識評価に選択的予測を導入し，出力の誤りリスクを考慮した評価を行うことを提案し，複数の確信度指標を用いて学習済み言語モデルの知識評価を行った．提案する評価手法は従来の予測精度に基づく評価で捉えられなかったモデルの誤りリスクを考慮した定量的評価を行うことができ，言語モデルの出力誤りによるリスクが大きい応用分野における言語モデル利用において，リスク評価の新たな指標として用いることが期待できる．また，学習済み言語モデルの入出力や内部状態，訓練データにアクセスできる状況を想定した複数の確信度指標の比較評価を行い，訓練データを用いない場合はモデルの予測尤度が，訓練データを用いることができる場合は関連事例検索に基づく確信度指標との組み合わせが知識評価において有効であることを確認した．この結果は言語モデル知識評価における確信度推定の新たな知見となるだけでなく，大規模言語モデルの訓練データにアクセスできることの有用性を示唆するものであり，言語モデル利用や開発に関する研究・応用の発展に寄与する．よって，本論文は博士（工学）の学位論文として十分価値あるものと認める．