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)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学系 知能情報コース	申請学位（専攻分野）： Academic Degree Requested	博士（工学） Doctor of
学生氏名： Student's Name	吉川 和	審査員主査： Chief Examiner	岡崎 直観

要旨（和文 2000 字程度）
Thesis Summary（approx.2000 Japanese Characters ）

近年，事前学習済み言語モデルの性能が著しく向上し，実社会での利用が急速に進んでいる．言語モデルの大規模化に伴い，言語モデルは学習の過程で訓練データから言語知識だけでなく常識や実世界の物事に関する知識を習得し，さまざまなタスクに活用できるようになった．こうした知識獲得は，言語モデルがより人間に近い自然な対話や現実に即した推論，意思決定支援などを行う助けとなっているだけでなく，実世界に関する知識を要する質問に外部の知識源を参照することなく回答するなど，言語モデルそのものを知識ベースの代替とみなすような使われ方も広がりつつある．一方で，言語モデルの出力には事実としての誤りが含まれることも多い．言語モデルが誤った内容を含む自然な文章を容易に生成できてしまうことにより，誤情報の拡散や正常な意思決定の阻害のリスクも重大化している．こうした背景から，言語モデルのもつ実世界に関する知識を正しく評価し，誤りを含む出力を検知・防止するための仕組みの構築が喫緊の課題となっている．事前学習済み言語モデルが獲得した知識はニューラルネットワークモデルのパラメータとして非明示的に保持されているため，モデルが具体的にどのような知識を保持しているかを直接確認することができない．そこで，言語モデルに特定の入力を行った際の出力やモデルの振る舞いを調べることで間接的に知識評価を行う方法が提案されている．LAMA probe はその代表的なもので，”Dante was born in ____.”のような特定の知識に関する穴埋めタスクを事前学習済み言語モデルに解かせることで，モデルが対象となる事実に関する知識を保持しているか否かを間接的に評価するベンチマークである．しかしながら，LAMA probe による知識評価には，モデルの予測の偏りによる偶然の正解が過大評価されてしまう懸念がある，個別の出力を信頼すべきか否かの判別可能性が考慮されていないといった課題がある．本研究ではこうした課題を解決するため，LAMA probe に選択的予測を導入し，確信度を考慮したモデルの知識評価を行う枠組みを提案する．選択的予測では，言語モデルの個々の出力に対し何らかの確信度指標に基づく確信度を計算するシステムを想定し，システムがより多くの質問に正答できるだけでなく，誤った出力の可能性が高い場合にはそれを検知できるかどうかをあわせて評価する．第一の研究では，選択的予測の導入が前述の課題を改善できるかを確認するため，言語モデルの内部状態と入出力のみを使って計算可能な確信度指標を複数設計した上で，選択的予測に基づく LAMA probe によるモデル評価を行った．複数のマスク言語モデルを対象にした実験では，選択的予測に基づく評価が従来の予測精度に基づく評価と比較し，モデルの予測や評価データの偏りによるモデル知識の過大評価の影響を低減できることが示唆された．評価に内在する偏りの是正方法として評価データセットを改善する従来のアプローチとは異なり，提案手法は評価データに依存せずこうした問題を緩和可能である．異なる確信度指標間の比較では，評価対象の知識の種類や言語モデルによって差はあるものの，単純な予測尤度に基づく確信度指標が一貫して良い性能であった．そこで第二の研究では，より多くの情報を利用することで言語モデル出力の確信度推定の性能を向上することができるかを焦点とし，言語モデルの学習時に用いられた訓練データに基づく確信度指標の設計・評価を行った．訓練データを公開している大規模言語モデルが増えつつある一方，従来の言語モデル出力の確信度推定はモデルの入出力やパラメータへのアクセスを前提としたものがほとんどであり，訓練データへのアクセスを想定した確信度推定の研究は現在のところ発展していない．訓練データを用いる確信度指標としては，入出力情報と関連する記述を訓練データ中から検索し，それらの事例を付加情報として用いる方法を複数設計した．実験には英語 Wikipedia データを用いて訓練した BERT モデルを用い，入出力と類似の訓練事例検索としては，訓練データに対してモデルがエンコードする文脈表現ベクトル・文ベクトルに基づくベクトル類似度検索とテキスト一致検索の 3 種類の検索方式を用いた．実験の結果，テキスト一致検索に基づき得られた関連事例を文脈に追加して再予測を行う方法が，単体で尤度ベースの確信度に匹敵する性能を達成した．さらに，訓練データを用いる確信度指標と訓練データを用いない指標とを組み合わせることにより，確信度推定の性能を改善できることを確認した．

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note：Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東京科学大学リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Science Tokyo Research Repository Website (T2R2).

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学系 知能情報コース	申請学位（専攻分野）： Academic Degree Requested	博士（工学） Doctor of
学生氏名： Student's Name	吉川 和	審査員主査： Chief Examiner	岡崎 直観

要旨（英文 300 語程度）
Thesis Summary（approx.300 English Words ）

Recently, the performance of pre-trained language models has improved significantly, and their use in the real world is rapidly increasing. It is known that pre-trained language models acquire from their training data not only linguistic knowledge but also common sense and knowledge about real-world entities, which can then be used for various tasks. On the other hand, the output of language models often generates fluent sentences that contain erroneous content, which increases risks such as spreading misinformation and inducing errors in decision-making. For this reason, a mechanism is urgently needed to assess the knowledge stored in language models and to detect their potentially erroneous output.

This study proposes an evaluation framework of the knowledge stored in language models. Our work is built on a benchmark for language model evaluation, the LAMA probe, which employs cloze tasks to assess the amount of knowledge stored in language models. The LAMA probe evaluation has issues, such as concerns about overestimating the models' knowledge due to lucky guesses caused by biases in the model predictions and not considering the discriminability of correct and erroneous answers. To address these issues, we introduce the selective prediction framework to the LAMA probe. Selective prediction assumes a system that calculates a confidence score for each output of a language model and evaluates not only the number of correct answers the model can make but also whether the system can detect when there is a high chance of incorrect output.

In the first study, we evaluated multiple masked language models using the LAMA probe under the selective prediction setting. Experiments suggest that selective prediction-based evaluation can reduce the effect of overestimation of models' knowledge due to biases in model prediction and evaluation data compared to the conventional evaluation based on prediction accuracy. In the second study, we focused on whether the performance of confidence estimation can be improved by using more information and designed confidence measures based on training data used during language model pre-training. The experiments using the BERT model showed that the performance of confidence estimation can be improved by combining the confidence measures using training data with those not using training data.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東京科学大学リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Science Tokyo Research Repository Website (T2R2).