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
Type(English)	Doctoral Thesis

博士論文

確信度を考慮した言語モデルの関係知識評価

吉川 和

東京科学大学 情報理工学院
情報工学系 知能情報コース

本論文は東京科学大学情報理工学院に
博士（工学）授与の要件として提出した博士論文である.

審査委員：
岡崎 直観 教授　（主指導教員）
徳永 健伸 教授
村田 剛志 教授
篠田 浩一 教授
宮崎 純 教授

概要

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．

キーワード

言語モデル，知識獲得，知識活用，確信度，不確実性，選択的予測，評価

Abstract

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 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.

Keywords:

Language Models, Knowledge Acquisition, Knowledge Utilization, Confidence, Uncertainty, Selective Prediction, Evaluation

謝辞

本研究の推敲および博士論文の執筆にあたり，多くの方からご指導・ご支援を賜りました．この場を借りて厚く御礼申し上げます．

主指導教員の岡崎直観教授には，4 年間にわたり非常に多くのご指導をいただきました．研究を進めるにあたっては至らない点が多く常に綱渡りのような状態で，大変なご迷惑をおかけしたかと思います．研究の方針や論文執筆でつまずいた際に的確な助言で背中を押していただき，なんとか研究を前に進めることができました．自然言語処理の分野に大きな動きがあった数年間に，岡崎研究室の一員としてこの分野に関われたことは非常に価値のある経験でした．心より感謝申し上げます．

徳永健伸教授には，入学前から進学に関する相談をさせていただいたり，入学以降も研究内容に対して的確な助言をいただくなど，様々な場面でお世話になりました．本研究に関しては，研究のスコープのや研究の発展の方向性についてのご質問・助言をいただき，研究の位置づけを整理するために役立つ視点を得ることができました．村田剛志教授には，提案手法を運用する上での妥当性や，他のアプローチとの比較についての的確なご指摘をいただきました．本研究の意義と提案するアプローチの妥当性を広い視点から見直すきっかけとなりました．篠田浩一教授には，研究発表において不明瞭な点を明確化させるための重要な助言をいただきました．検索に基づく真偽判定やモデルによる確信度の言語化など，類似の研究テーマと本研究がどのように差異化できるかを検討することができました．宮崎純教授には，最近の大規模言語モデルの技術的発展も踏まえた研究の位置づけや，確信度推定の定義や運用方法について踏み込んだご質問・ご指摘をいただきました．

Simone Teufel 教授には，英語ライティングに関する集中講義を行っていた

だいただけでなく，EACL に投稿した論文を丁寧に添削していただきました．特に論文の導入部分の英語表現を大きく改善することができ，非常に勉強になりました．

秘書の小西由希子さん，雲財祐子さん，古谷奈緒子さんには研究生活を送る上で非常に多くのサポートをいただきました．小西さんには日頃の打合せ調整や大学事務とのやり取り，博論審査までのスケジュール管理などさまざまな面で大変お世話になりました．雲財さんは計算資源の管理や支払い手続きなど，研究をスムーズに進める上で無くてはならないご支援をいただきました．古谷さんには海外出張の際に特にお世話になりました．細かい点まで何でもご相談させていただき大変助かりました．懇親会などでもいつも気さくに話しかけてくださり楽しかったです．

岡崎研の研究員や学生の皆様にも大変お世話になりました．高瀬翔元助教はコロナの影響で人が少ない時期からいつも研究室にいらっしゃり，継続的にトップレベルの成果を挙げられている姿に刺激を受けていました．また，国際会議論文の投稿の際には〆切が近い中で的確なレビューをしていただき大変助かりました．金子正弘さんは論文の書き方などのノウハウを積極的に共有してくださったり，外部のコミュニティや海外に積極的に出る姿勢で研究室に多くの影響を与えてくださっていました．EACL でご一緒できたのは心強かったです．平岡達也さんは博士課程の先輩として大きな足跡を残してくださり，研究生活において多くの場面で助けられました．また，短い間ではありますが会社の同僚としてもご一緒できてよかったです．丹羽彩奈さんとは入学前から学会などでご一緒し，岡崎研の様子を聞かせていただいたりしていました．研究を進めながら対外的にも様々な活動に取り組まれている姿勢からはいつもエネルギーを貰っていました．水木栄さんには入学当初から講義やカリキュラムに関する相談に乗っていただきました．また，セミナーではいつも的確かつ丁寧なコメントをいただき，大変助けられました．プレゼンの構成力や分かり易さも素晴らしく，お手本にさせていただいています．飯田大貴さんとは研究の議論をしたり，参考になる論文や実装を共有していただいたりと度々お世話になりました．また，社会人博士の先輩としても様々な情報共有をしていただき，とても助かりました．村岡雅康さんは企業研究員という近い立場で業務や博士課程と家庭をしっかりと両立させており，刺激をいただいていました．Vijay Daultani さんとは同じ時期に入学した同期として

いろいろな情報共有をさせていただきました．Marco Cognetta さん，An Wang さんには国際会議に投稿する論文をレビューしていただいたり，研究や生活のことに関して楽しく雑談させていただきました．Ma Youmi さんは Web 運営をはじめ，研究室の運営をさまざまな面で支えてくださり，感謝いたします．EACL 出張では色々なお話ができてとても楽しかったです．ここでお名前を挙げきれなかった皆様にも日頃の活動や議論を通じて大変お世話になりました．改めて感謝申し上げます．

富士通研究所の皆様には，博士課程入学から在籍中まで多くの支援をいただきました．博士号取得を薦めてくださった岩倉友哉さん，在学中に様々な面でご支援・ご配慮くださった高橋哲朗さん，梅田裕平さん，河東孝さんをはじめ，同僚の皆様に感謝いたします．

最後になりますが，博士課程進学を応援してくださった父，母，兄弟，親戚，友人と，いつも支えてくださる中村尚貴さんに感謝いたします．

目次

図目次

表目次

第 1 章

序論

1.1　言語モデルの高度化とその影響

近年，大規模なテキストデータによる学習に基づく言語モデルの性能が飛躍的に向上している．特に，Transformer モデル (Vaswani et al. 2017) の出現以降は，これを基礎とする言語モデルを中心に大規模化と高性能化が進んでおり，現在では人間によるものと区別のつかない文章の生成が可能なレベルに到達している (Devlin et al. 2019, Radford et al. 2019, Touvron et al. 2023)．また，2022 年に ChatGPT (OpenAI 2022) のサービスが開始されたことにより，専門的な知識を持たない一般のユーザも言語モデルを対話形式で利用することが可能となり，大規模言語モデルへの注目が集まるとともに，利用範囲が急速に広まった．

言語モデルの高度化は，自然言語処理におけるさまざまなタスクへのアプローチにも変化をもたらした．従来の機械学習に基づく自然言語処理では，解くべきタスクに応じた訓練データを用いて専用のモデルを学習する必要があった．これを転換させたのが，BERT (Devlin et al. 2019) をはじめとする事前学習済みモデルの出現である．BERT は大規模なテキストデータに基づく事前学習により汎用的な言語知識をもつ基盤モデルを構築し，少量の訓練データを用いたファインチューニングによりモデルを個別タスクに適応させる方式により，さまざまな言語理解タスクで高い性能を達成した．その後，モデルと訓練データのさらなる大規模化により，追加の訓練データを用いることなく，単一のモデルに少数の例やタスク指示を入力することにより多様なタスクを解かせることが可能となった (Radford et al. 2019, Brown et al. 2020)．

言語モデルが訓練データから獲得する知識は言語理解に関わるものだけではなく，常識や実世界の物事に関する知識も含まれる．これにより，質問応答のように実世界に関する知識を必要とするタスクも，学習の過程で知識獲得ができているならば，外部の知識源を与えることなく言語モデルが直接回答することが可能な場合がある (Petroni et al. 2021, Roberts et al. 2020)．このように，言語モデルの大規模化と高度化は，文章の読解，生成，知識獲得と活用など，さまざまな面で自然言語処理における言語モデルの役割を大きく広げることとなった．

1.2 言語モデルの出力誤りと知識評価

言語モデルが訓練データから得た実世界の知識を活用して高度なタスクを行うことができるようになった一方で，出力には事実と異なる誤った内容が含まれることもある．文生成の質の向上により，誤った内容を含む文章でも人が書いたものと見分けがつかない場合が多い．言語モデルの実用化が急速に進む中，このことは誤情報の拡散や Web 検索汚染，意思決定への影響といった社会的リスクに関わる重要な問題のひとつとなっている (Huang et al. 2024)．

こうした問題に対応するためには，言語モデルが具体的にどのような知識を獲得しており，逆に何を知らないのかを確認する手段があることが望ましい．しかしながら，言語モデルの獲得した知識はモデルパラメータとして非明示的に保持されているため，モデルのもつ知識の内容を直接確認することができない．また，商用の大規模言語モデルでは，そもそもモデルへのアクセス手段が API を介するものに限定されていたり，訓練データが公開されていない場合も多い．このような状況の中，事前学習済み言語モデルが訓練データとして用いた文書を特定する方法 (Zhang et al. 2024, Wei et al. 2024) や，モデルの獲得した知識を抽出する方法 (Bosselut et al. 2019) が模索されている．

事前学習済み言語モデルの知識評価のためのベンチマークタスクとして代表的なものに LAMA probe (Petroni et al. 2019) がある．LAMA probe では，知識ベースをもとに作成した作成した穴埋め形式の問題を言語モデルに与えることで，言語モデルが事前学習で得た知識をもとに正しい予測ができるかを評価する．正しい予測ができた場合には言語モデルが当該事実を「知っている」という仮定のもと，モデルの知識量が予測精度に基づき評価される．

しかしながら，この仮定はさまざまな要因から正確ではないことが指摘されている．例えば，言語モデルに与える入力文（プロンプト）の違いにより，同じ内容であってもモデルが正解を予測できる場合とできない場合がある (Jiang et al. 2020)．また，人名と出身地のように，一方のエンティティの表層情報から知識がなくとも容易に推測できてしまうような場合があり，モデル知識の過剰評価が生じる懸念がある (Poerner et al. 2020)．こうしたモデルの文脈依存性や予測の偏りは，入力表現の多様化や評価データの改善による対処が可能であるものの，予測精度に基づく知識評価の枠組みでは避けられない問題である．

もう一つの課題は，予測精度に基づくモデルの知識評価が個々のモデル出力の信頼性に関する判断材料を提供しないことである．言語モデルの利用範囲が拡大する中，言語モデルの出力誤りが重大な社会的リスクに繋がる場面が増えている (Hao et al. 2024, Ji et al. 2023)．こうした状況下では，仮に言語モデルが高い割合で正しい知識に基づいた出力を行っていたとしても，誤りの可能性のある出力が一定数含まれるリスクを無視できない．この懸念を解消するためには，言語モデルの個別の出力が誤りを含む可能性を適切に検知し，誤りリスクを回避することが容易であることが望ましい．予測精度のみに基づく知識評価には誤りリスクの観点が欠けており，実用上の要請との乖離がある．

1.3 アプローチ: 確信度を考慮したモデル知識評価

本研究では，以上に挙げた言語モデル知識評価の課題を改善し，予測を信頼すべきかの判断材料を提供することを目的とし，次のことに取り組んだ．

まず，言語モデルの出力誤りのリスクを考慮した知識評価として，選択的予測 (El-Yaniv and Wiener 2010, Geifman and El-Yaniv 2017) を導入した枠組みを提案した．選択的予測においては，言語モデルによる予測が信頼できるものであるかを何らかの確信度指標に基づき判断し，予測の信頼性が高いとみなされたときのみ出力を行うシステムを考える．システムは，不確かな出力を確信度指標に基づき抑制しつつ，できるだけ多くの正しい出力を行うことができるかによって評価されるため，予測精度に基づく評価では捉えられない個々の出力の信頼性判断を考慮した評価が可能となる．本研究では既存の言語モデル知識評価タスクである LAMA probe に対し選択的予測の枠組みを導入し，複数のマスク言語モデル

の評価を行った．確信度指標としては，モデルの予測尤度を用いる最も単純な指標をはじめとする，モデルの入出力や内部状態を用いて計算できる複数の指標を設計して用いた．実験では，提案する評価手法が従来の予測精度に基づく評価に比べ，モデル予測の偏りや評価データの偏りに影響されたモデル知識の過大評価を低減していることを確認した．したがって，提案手法は LAMA probe の二つの課題であったモデル予測の偏りの影響と個々の出力の信頼性考慮の双方を改善するアプローチとなり得る．提案手法は新たな評価尺度の導入であるため，データセットや入力文の改善を伴う既存手法と異なり評価データにかかわらず適用することが可能で，既存手法と組み合わせることも可能である．

次に，選択的予測において用いられる言語モデル出力の確信度指標の改善に取り組んだ．具体的には，言語モデルの訓練データにアクセスできる状況を想定し，訓練データ中の関連事例に基づく確信度指標を設計し，評価した．言語モデル出力の確信度推定には，これまで出力テキストや予測尤度，モデルパラメータといった情報が用いられているが，訓練データを参照できる状況下での確信度推定については十分な検証が行われていなかった．本研究では，英語 Wikipedia 全文を用いて自前で訓練した BERT モデル (Devlin et al. 2019) を用い，訓練データからの関連事例の検索と関連事例を用いた確信度推定について，複数の方法を検討し評価した．実験では，入出力内容と関連する事例を適切に検索できた場合に，これに基づく確信度指標による性能改善が見込めること，訓練データを用いない確信度指標との組み合わせが性能改善に大きく寄与することを確認した．

1.4 本研究の貢献

本研究の貢献は次のとおりである．

- 言語モデル出力の知識評価に選択的予測の枠組みを導入し，言語モデルの知識を個別の出力内容の信頼性を考慮して評価する方法を提案した．
- 選択的予測に基づく確信度評価が，従来の予測精度に基づく評価と比較して言語モデルの予測の偏りや評価データの偏りに起因するモデル知識の過大評価を抑制できることを確認した．
- 言語モデルの事前学習に用いた訓練データを用いる確信度指標を新たに提

案した．提案する指標は，言語モデルの入出力内容と関連するテキストを訓練データ中から検索し，関連事例に基づき確信度指標の計算や補正を行う．

- 訓練データに基づく確信度指標の有効性を検証するため，英語 Wikipedia データを訓練データとして自前で訓練した BERT モデルを用いた評価を行い，提案した指標が確信度推定の性能改善に寄与することを確認した．

1.5 論文の構成

2 章では，本論文で用いる概念や関連研究について説明する．まず，Transformer モデルとそれを基礎とする言語モデルの代表的なものにを説明し，言語モデルの大規模化にともなう発展について概観する．次に，言語モデルの訓練データからの知識獲得と活用，言語モデルの知識評価について述べる．さらに，言語モデルの誤りへの対処や確信度推定にかかわる関連研究について述べる．3 章では，本研究で提案する選択的予測に基づく言語モデルの知識評価について述べる．言語モデルの知識評価ベンチマークである LAMA probe に選択的予測を導入し，言語モデル出力に対する複数の確信度指標を導入する．実験では，複数のマスク言語モデルを対象に，選択的予測の設定のもとで LAMA probe の評価を行い，従来の予測精度による評価との比較を行う．4 章では，確信度指標の性能向上を目的として，言語モデルの訓練データに基づく確信度指標を提案する．ここでは，言語モデルの入出力をもとに訓練データから関連する事例を検索する方法と検索結果に基づく複数の確信度指標計算方法を導入する．実験では，訓練データを用いた確信度指標を LAMA probe で評価し，訓練データを用いない確信度指標との性能比較や組み合わせによる効果を検証する．最後に 5 章で本研究のまとめと今後の展望を述べる．

第 2 章

準備と関連研究

2.1 事前学習済み言語モデル

近年の自然言語処理においては，個別のタスクに応じて訓練データを用意し機械学習モデルを作成するのではなく，大規模なコーパスを用いて単一の事前学習済みモデルを訓練し，さまざまなタスクに転用する流れが主流となっている．こうした近年の自然言語処理の基礎となる事前学習済み言語モデルについて説明する．

近年の事前学習済み言語モデルの大部分は Transformer (Vaswani et al. 2017) モデルを基礎としている．本節ではまず 2.1.1 節で Transformer モデルについて説明し，2.1.2 節，2.1.3 節で Transformer のエンコーダのみ・デコーダのみに基づく代表的な事前学習済み言語モデルをそれぞれ紹介した後，2.1.4 節で近年の大規模言語モデルについて述べる．

2.1.1 Transformer

図 2.1aに Transformer モデルの概観を示す．Transformer モデルは，入力系列 $(x_1, \ldots, x_n)$ をベクトル表現の系列 $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ に変換するエンコーダと，$\mathbf{Z}$ を参照しながら出力系列 $(y_1, \ldots, y_m)$ を生成するデコーダの 2 つのモジュールから構成されるエンコーダ-デコーダモデルである．エンコーダとデコーダはそれぞれ**注意機構** (Attention) と呼ばれるモジュールをもつ複数のネットワーク層により構成される．

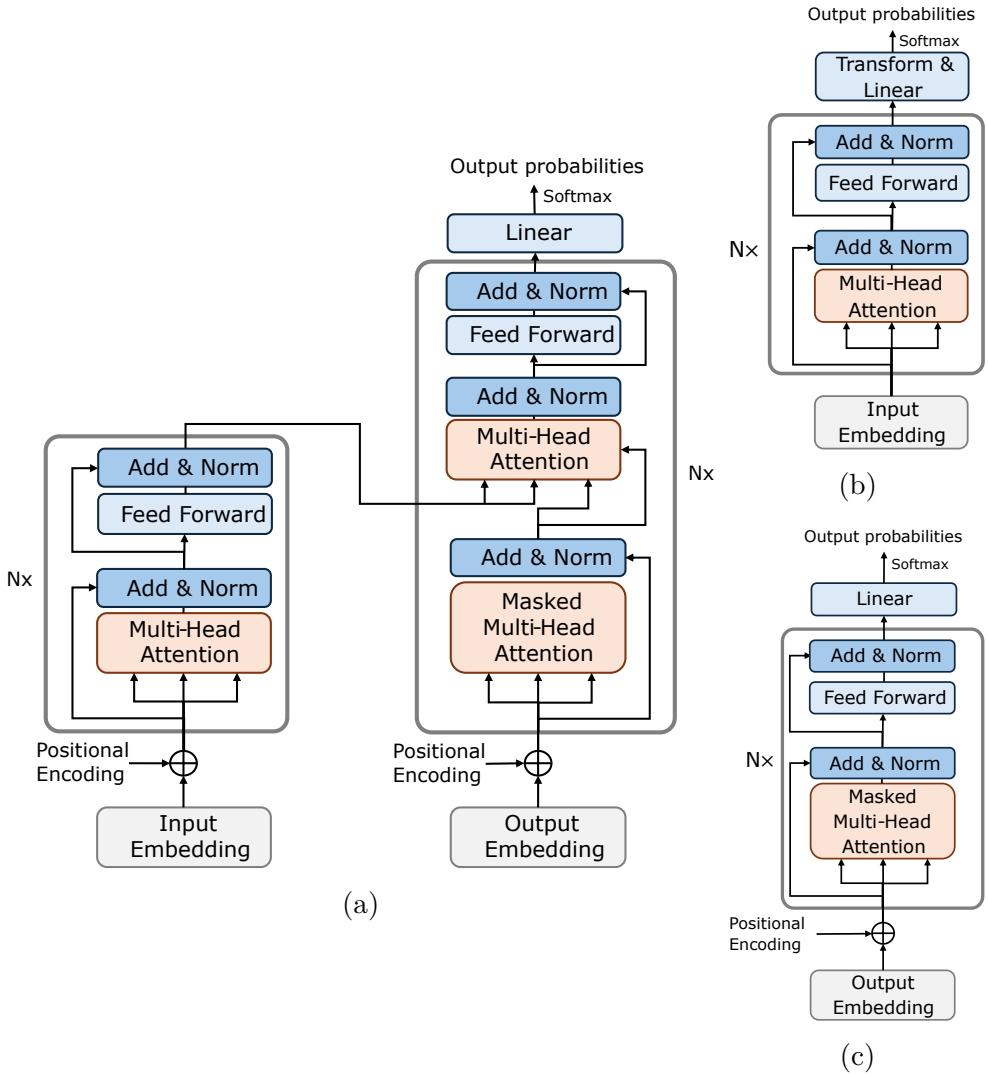


図 2.1: (a) Transformer モデル (Vaswani et al. 2017)，(b) Encoder-only モデル (BERT) (Devlin et al. 2019)，(c) Decoder-only モデル (GPT) (Radford et al. 2018).

入力埋め込み表現

エンコーダおよびデコーダは，トークン分割された入力系列を学習済みの埋め込み行列により d_{model} 次元ベクトルに変換したベクトル系列を入力とする．Transformer モデルのエンコーダやデコーダは入力系列の位置関係を区別する仕組みを持たないため，入力埋め込みにおいて各トークンの系列中の位置を区別するため

の位置エンコーディングを加算する．位置 t の位置エンコーディングは次の式により計算される：

$$PE_{(t,2i)} = \sin(t/10000^{2i/d_{\text{model}}}), \tag{2.1}$$

$$PE_{(t,2i+1)} = \cos(t/10000^{2i/d_{\text{model}}}). \tag{2.2}$$

注意機構

Transformer モデルの要である**注意機構**の Attention 関数は，行列 $Q \in \mathbb{R}^{n_1 \times d_k}, K \in \mathbb{R}^{n_2 \times d_k}, V \in \mathbb{R}^{n_2 \times d_v}$ を入力とし，次の式により $n_1 \times d_v$ 行列を出力する:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}})V. \tag{2.3}$$

上式は直感的には，クエリ系列 Q により K, V で表現される key-value ストアを参照する操作と捉えることができる．Q 中の各クエリベクトル $\mathbf{q} \in \mathbb{R}^{d_k}$ について，n_2 個の key ベクトル $K = (\mathbf{k}_1, \ldots, \mathbf{k}_{n_2})$ との内積をとり関連度を計算する．その後，この内積を各 key に対応する value の重みとし，value ベクトル系列 V の重み和をとる．

Transformer の注意機構では，d_{model} 次元の入力 Q, K, V に対し h 通りの異なる線形変換を施すことで複数の Attention を計算し，結果を組み合わせる:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.4}$$

$$\text{where} \qquad \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \tag{2.5}$$

$W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}, W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ はモデルパラメータである．式 2.5で表されるヘッドを複数用いることにより，系列間の多様な関連性を捉えることが可能となる (multi-head attention)．

エンコーダ

エンコーダは，注意機構を含む self-attention 層とフィードフォワード層の 2 つのサブレイヤーからなるエンコーダ層が重なった構成となっている．各サブレイヤーの間では残差接続 (He et al. 2016) とレイヤー正規化 (Ba et al. 2016) が行わ

れる（図 2.1a中では `Add & Norm` で示される）．

エンコーダ層の注意機構では，入力 Q, K, V は全て同一の入力ベクトル系列 $\mathbf{X}$ である．クエリと key-value ペアのいずれも入力 $\mathbf{X}$ 自身に基づく注意機構は特に自己注意機構 (self-attention) と呼ばれる．

Self-attention 層の出力はフィードフォワード層に与えられ，系列中の各位置で独立に計算される:

$$\mathrm{FFN}(\mathbf{x}) = \max(0, \mathbf{x}W_1 + \mathbf{b}_1)W_2 + \mathbf{b}_2. \tag{2.6}$$

$W_1, \mathbf{b}_1, W_2, \mathbf{b}_2$ は層ごとに設定されるモデルパラメータである．

デコーダ

デコーダは，既に生成済みのトークン系列 $(y_{\mathrm{sos}}, y_1, \ldots, y_{t-1})$ を入力として次のトークン y_t を生成することを逐次的に繰り返すことで出力系列を生成する．なお，y_{sos} は出力文の先頭を表す特別なトークンである．デコーダを構成するデコーダ層は，2 種類の注意機構を含む層とフィードフォワード層の 3 つのサブレイヤーからなる．デコーダ最終層の出力が線形変換と Softmax 関数により次トークンの予測確率分布に変換される．

入力系列を処理するのはエンコーダと同様 self-attention 層である．エンコーダとは異なり，デコーダは位置 t の推論時にはそれまでに生成済みの $t-1$ 番目までのトークン系列のみしか知ることができない．そのため，注意機構においても t 番目以降の系列の情報を参照することがないよう，当該箇所にマスクをかけたマスク付き自己注意機構を用いている．

次に，エンコーダ側の情報を参照するための第二の注意機構として encoder-decoder attention 層が挿入される．ここでのクエリはデコーダの self-attention 層の出力， key と value はエンコーダ出力に基づいて構成される．これにより，各トークンの生成時に入力系列全体から必要な情報を参照することが可能となる．最後にエンコーダ層と同様，式 2.6のフィードフォワード層が計算される．

2.1.2 Encoder-only モデル

BERT (Devlin et al. 2019) は，Transformer のエンコーダ部分のみを用いるモデルの一つである．図 2.1bにモデルの概観を示す．BERT は Transformer のデコーダが行うような逐次的なテキスト生成は行わず，代わりにマスク予測と次文予測の 2 つのタスクにより訓練される（事前学習）．大規模なテキストデータを用いて事前学習を行った BERT モデルを言語理解や質問応答といった下流タスクのデータを用いてファインチューニングすることで，さまざまなタスクで発表当時の最高性能を達成した．

BERT は単語や文の生成確率を直接モデル化していないため，本来の定義上の言語モデルには該当しない．しかし，後述するマスク予測に基づく穴埋めにより単語や文の予測が可能であり，マスク言語モデルとも呼ばれ，広義の言語モデルとして扱われることがある．本論文中でも，特に断りがない限り encoder-only モデルを言語モデルの一種として取り扱う．

入力

BERT は最大 2 文の文対を入力として受け付ける（ここでの「文」は任意のテキスト片を想定しており，本来の意味での文でなくてもよい）．入力は Transformer と同様単語埋め込み表現のベクトル系列である．入力の先頭を示す `[CLS]`，文区切りを示す `[SEP]` の 2 種類が特別なトークンとして用いられ，“`[CLS]` 文 `A` `[SEP]` 文 `B`” のように入力される．入力埋め込みにはトークン埋め込みと位置エンコーディングに加え，文 A と文 B を区別するためのセグメント埋め込みが加算される．

マスク予測

BERT モデルの事前学習における第一のタスクであるマスク予測は，入力文の位置 t の単語を周辺文脈から予測させることで，モデルに前後の文脈を考慮した意味表現を獲得させることを目的とする．BERT モデルは注意機構によって位置 t 自身を含む入力文全体を参照することが可能なため，学習においては予測対象の単語を入力として参照することができないよう対象位置を特別なトークン `[MASK]` によって置き換えたマスク済み入力文を用い，周辺文脈に基づいてマスクの穴埋めをさせる．予測対象となるマスク位置は各入力文の 15%をランダムに選択する．

しかしながら，この方法では学習時の入力には [MASK] トークンが含まれ，ファインチューニング時には本来の入力トークンのみが含まれるという齟齬が生じる．これを緩和するため，選ばれたマスク位置のうち 80%は [MASK] トークンに，10%はランダムに選ばれたトークンに置き換え，残りの 10%は元のトークンをそのまま用いるよう入力を作成する．

BERT モデルによる位置 t のトークン予測では，エンコーダ最終層の同位置 t の隠れベクトルを変換し，語彙全体に対する確率分布を計算する．訓練時には正解単語との交差エントロピーが損失として用いられる．

次文予測

第二のタスクである次文予測は，質問応答における質問文と回答文のような 2 文の関係性を捉えることを目的としたタスクで，入力文対の文 A と文 B がコーパス中で隣接関係にあるかを判定する二値分類の形式をとる．学習時のデータは，文 A に対し 50%の確率で実際に続く文を，50%の確率でコーパス中のランダムな文を文 B として選択し，文対を作成することで容易に構築可能である．次文予測には，文頭の [CLS] トークンに対応するエンコーダ最終層の隠れベクトルが用いられる．

2.1.3 Decoder-only モデル

GPT (Radford et al. 2018) は Transformer モデルのデコーダ部分のみを用いた言語モデルである．図 2.1cにモデルの概観を示す．デコーダ層はエンコーダ情報を受け取らないため encoder-decoder attention を持たず，マスク付き自己注意機構とフィードフォワード層の 2 つのサブレイヤーからなる．

GPT モデルはデコーダに基づくため，Transformer モデルと同様の自己回帰型の言語生成を行うことができる．BERT と同様に事前学習により多様なタスクに利用できる言語知識を獲得するが，事前学習においては単に言語モデリングの目的関数を用いる：

$$L(Y) = \sum_i \log P(y_i \mid y_{i-k}, \ldots, y_{i-1}; \Theta). \tag{2.7}$$

なお，$Y = (y_1, \ldots, y_n)$ は訓練データの正解系列全体，k はモデルが参照する文脈

の窓幅で，確率 P はパラメータ Θ に基づく GPT モデルにより推定された次単語の生成確率である．

2.1.4 大規模言語モデル

とりわけ Transformer ベースのモデルを中心とする言語モデルの性能はモデルパラメータ数の増加にともない向上することが示唆されており (Kaplan et al. 2020)，実際により大規模かつ高性能な言語モデルがこの数年で数多く開発されている．Decoder-only モデルである GPT の後継として，2019 年に GPT-2 (Radford et al. 2019)，2020 年に GPT-3 (Brown et al. 2020) が発表された．GPT-2 は 4500 万のウェブページから収集したテキストデータを用いて学習され，パラメータ数は GPT の 10 倍以上となる 15 億へと増加しており，多様なタスクを単一モデルが追加の訓練なしで解く zero-shot や few-shot 設定での性能を大きく向上させている．GPT-3 はさらに大規模となる 1750 億のパラメータを持ち，数個の例を与えて目的のタスクを解かせる in-context learning 方式により，従来モデルをファインチューニングするのと同程度かそれ以上の性能を達成できることが報告されている．GPT 系列のモデルと対話インターフェースを介してやり取りできる ChatGPT (OpenAI 2022) が登場したことで，大規模言語モデルは一般のユーザにまで広く認知され，普及することとなった．

ChatGPT をはじめとする商用モデルの多くは API を介してのみアクセスできる仕様で，モデルのパラメータやアーキテクチャの詳細が非公開となっている．一方，Llama (Touvron et al. 2023)，BLOOM (Workshop 2023)，Mistral (Jiang et al. 2023) など，実装やパラメータを公開したモデルも増加している．国内では，Swallow (Fujii et al. 2024) や LLM-jp (LLM-jp 2024) といった日本語処理能力の向上に注力した大規模言語モデルの開発が進められている．

2.2 事前学習済みモデルの知識活用と評価

2.2.1 事前学習による知識獲得と活用

2.1 節で述べたように，事前学習済みモデルは大規模な訓練データを用いて獲得した知識をさまざまなタスクに転用することができる．ここでの知識は文章読解に必要な言語知識にとどまらず，実世界の事象に関する知識も含まれる．これに

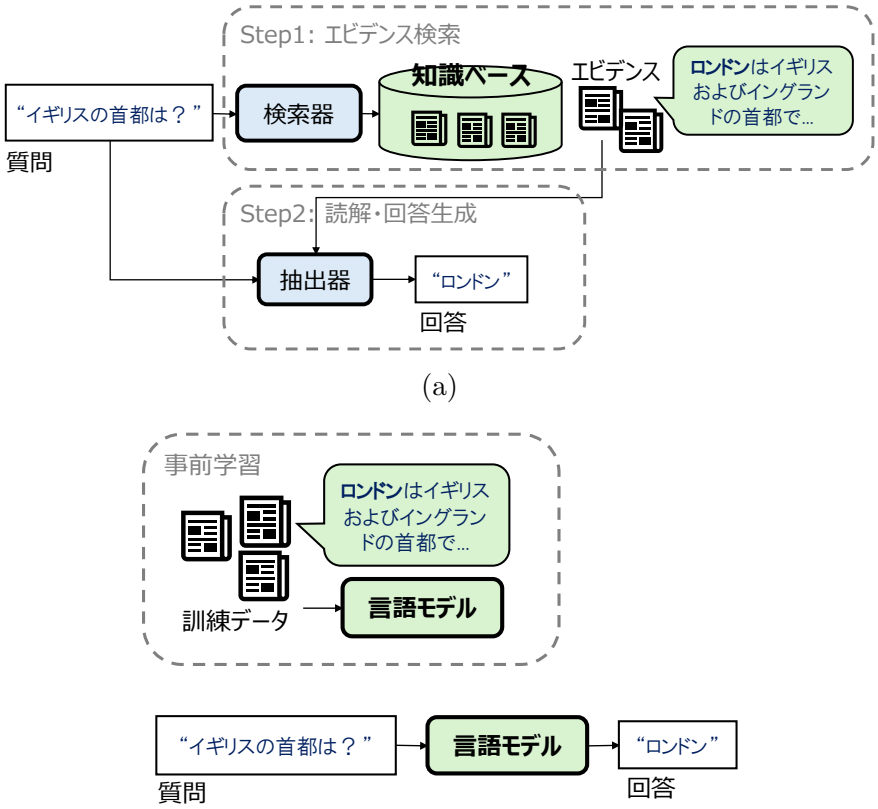


図 2.2: 知識を要するタスクに対するアプローチの変化．(a) は従来のアプローチ．質問をもとに外部の知識ベースから関連する文書を検索し，検索された文書を読解することで回答を行う．(b) は事前学習モデルのみを用いるアプローチ．言語モデルが事前学習の際に獲得した知識を用いて，外部の知識ベースを参照することなく回答を生成する．

より，実世界に関する知識を必要とする質問応答などのタスクも，場合によっては事前学習モデルのみで解くことが可能となった．

図 2.2は知識を要するタスクに対する従来型 (a) と事前学習済みモデルに基づく (b) アプローチの比較である．従来型のアプローチは，入力内容に関連する文書を外部の知識ベースから検索するステップと，検索した文書を読解することで入力に対する回答を生成するステップから構成される．比較的最近のものとしては，ベクトル検索と系列変換モデルを用いた RAG と呼ばれるアプローチがある (Lewis et al. 2020)．一方，事前学習済みモデルに基づくアプローチでは，モ

デルが事前学習で獲得した知識をもとに，外部知識を参照することなく直接回答を生成する．2020 年には，BART モデル (Lewis et al. 2020) のみを用いて実世界の知識を要するタスクが解かれ，一部のデータセットでは検索ベースの手法と同等以上の性能を達成したことが報告されている (Petroni et al. 2021)．同時期にRoberts et al. (2020) も，Transformer ベースの系列変換モデル T5 をファインチューニングしたモデルを用いて，外部知識の参照なしで質問応答タスクを解く試みを行っている．

2.2.2 言語モデルからの知識抽出と評価

事前学習済み言語モデルの知識活用能力が注目される中，モデルが獲得した知識の具体的な内容を抽出・評価する取り組みも進んだ．Transformer ベースの事前学習済みモデルは，獲得した知識を膨大なモデルパラメータとして非明示的に保存しているため，具体的に獲得している知識の内容や量を直接確認することができない．そこで，特定の入力に対するモデルの出力を調べることで間接的にモデルの知識を抽出・評価するアプローチが主流となっている．COMET (Bosselut et al. 2019) は，GPT モデルを常識に関する知識の穴埋めタスクに対しファインチューニングしたモデルを用いて，常識的知識の知識ベースを自動構築するフレームワークである． Petroni et al. (2019) は，事前学習済み言語モデルが学習の過程で獲得した，実世界の事物（エンティティ）や常識といった知識を評価するためのベンチマークとして LAMA probe を提案し，異なる言語モデル間の比較評価を行えるシンプルなフレームワークを提供した．本節では LAMA probe に注目し，評価の枠組みや議論について述べる．

LAMA probe

LAMA probe では主に（`subject`，`relation`，`object`）の三つ組で表現できる関係知識を評価対象とする．言語モデルが評価対象の事実に関する知識を持っているかどうかは，この事実を表す穴埋め問題に回答できるかどうかで判定する．例えば，(Dante, place of birth, Florence) という事実 (“place of birth” は `subject` の出生地が `object` であるという関係を表す) に対応する言語モデルへの入力として，“Dante was born in _____ .” といった文を与える．言語モデルが空欄

表 2.1: LAMA データセットの事例サンプル．SQuAD サンプルには subject, relation ラベルが付与されていない．

データセット	relation	subject	object	言語モデルへの入力文
Google-RE	place of birth	Marvano	Belgium	Marvano was born in ___ .
	date of birth	Cartrain	1991	Cartrain (born ___).
T-REx	official language	Brunei	Malay	The official language of Brunei is ___ .
	occupation	Boethius	philosopher	Boethius is a ___ by profession .
ConceptNet	HasPrerequisite	rinsing	water	Rinsing requires ___.
	IsA	cranberry	fruit	Cranberry is ___.
SQuAD	-	-	red	Phycoerytherin has ___ color.
	-	-	Edinburgh	Under the terms of the Scotland Act of 1978, an elected assembly to be set up in ___.

を正しい言葉で埋めることができたとき，言語モデルがこの事実に関する知識を持っていると判定する．

LAMA probe では異なる種類の知識を評価する 4 つのデータセットをもとに構築された評価データを用いる．以降，これらのデータセットを総称して LAMA データセットと呼ぶ．以下で各データセットの概要と入力文の作成方法について述べる．

Google-RE Google-RE (Google 2013) コーパスは，Wikipedia から人手で抽出されたエンティティ間の関係のデータセットである．LAMA probe では，“place of birth”（出生地），“date of birth”（誕生日），“place of death”（没地）の 3 種類の関係についての約 5500 の事実を取り扱う．入力文は，各関係ごとに共通のテンプレートを人手で作成し，これに各事実のエンティティを当てはめることで作成する．例えば “place of birth” に対応するテンプレートは “`<subject>` was born in _____ .” （`<subject>` には `subject` エンティティが入る）となり，下線部の正解が `object` エンティティとなる．

T-REx T-REx (Elsahar et al. 2018) は，Wikidata (Vrandečić and Krötzsch 2014) をソースとしたエンティティ間関係のデータセットである．Wikidata 中の三つ組が Wikipedia 本文中の記述と自動的に紐づけされており，紐づけの

精度は 97%を超えたと報告されている．LAMA probe ではこの内 41 種類の関係についてそれぞれ最大 1000 件，計 3 万 4000 件の事実を評価データとして用いている．Google-RE 同様，関係の種類ごとに人手で作成したテンプレートを用いて入力が作成される．

ConceptNet ConceptNet (Speer and Havasi 2012) は，一般的な物事に関する常識的知識を自然言語による表現に基づき表現した巨大な意味グラフである．含まれる関係知識は，(cake, IsA, dessert)（ケーキはデザートである）のような言葉の定義から導けるものから，(jazz, AtLocation, new orleans)（ジャズはニューオーリンズでみられる）といった人間の一般的な知識に基づくものまで多岐にわたる．ConceptNet は Open Mind Common Sense (OMCS) プロジェクト (Singh et al. 2002) によって Web を通じて収集した文を主要な知識源としている．LAMA probe では ConceptNet の英語データの一部である 16 種類の関係を含む約 1 万 1000 件の三つ組が用いられる．入力文としては，三つ組に対応する OMCS の文が `subject` と `object` を含んでいるため，これを使用し，`object` をマスクして予測対象とする形をとっている．

SQuAD SQuAD (Rajpurkar et al. 2016) は質問応答の代表的なデータセットの一つで，問題は Wikipedia 記事をベースに作成されている．元の SQuAD は対象となる Wikipedia 記事を背景情報として提示し，質問の回答を記事中から特定する読解問題であったが，LAMA probe では背景情報を与えずにエンティティに関する知識を問うために用いる．LAMA probe においては，回答が一単語である 305 件の質問を評価データとして利用する．入力文は穴埋め形式に適した表現に人手で作り変えられたものが用いられる．なお，SQuAD データセットのみ質問内容が三つ組で表現される関係知識に限定されておらず，このデータセットに関しては `subejct` や `relation` が定義されていない．

表 2.1に各データセットの事例サンプルを示す.

LAMA probe は GPT や Transformer-XL (Dai et al. 2019) のように前の文脈のみを参照する一方向言語モデルと，BERT や ELMo (Peters et al. 2018) のように前後の文脈を参照する双方向言語モデルを評価対象として想定する．また，

評価の複雑化を防ぐ目的で，データセット中の予測対象のエンティティは全て一単語で表現可能なものとなっている．一方向言語モデルの場合は入力文の予測対象位置の直前までの文脈を与えた際の予測を用い，双方向言語モデルの場合は各モデルの予測方式にのっとり，マスクした予測対象箇所を除く前後全ての文脈に基づく予測結果を用いる．BERT モデルの場合，予測対象位置の単語を `[MASK]` トークンに置き換えたものを入力とし，2.1.2 節のマスク予測の要領で予測を行わせる．

評価指標としては，モデルの予測上位 k 件のうちに正解単語が含まれている割合を示す precision at k (P@k) の平均を用いる．ただし，主結果は $k = 1$ で評価しているため，この場合は最上位の予測を用いた予測精度と等価になる．

LAMA probe に関する議論

LAMA probe では言語モデルへの入力文として人手で作成したテンプレートに沿った固定の表現を用いている．これは，入力文を言い換えることで言語モデルが正しい予測を出力できるケースを無視しており，モデルの能力の過小評価につながる可能性がある．Jiang et al. (2020) は実際に LAMA probe の入力文の言い換えを自動生成し結果を組み合わせることで，モデルの予測精度を向上させることができると報告している．

一方，LAMA probe がモデルの知識を過大に見積もっている可能性も指摘されている．Poerner et al. (2020) は BERT モデルが LAMA probe の予測において，例えばイタリア人風の人名に対しイタリア語を話すと推測するといったように，エンティティの表層に強く依拠した推測を行っている可能性を指摘した．これを確かめるために作成された LAMA-UHN データセットは，LAMA データセットからエンティティの表層からの推測が用意な知識を除外したものである．実際，BERT の平均正答率はこのサブセット上で大きく低下することが確認された．Kassner and Schütze (2020) は LAMA probe の入力を否定表現に置き換える（“Birds can `[MASK]`” → “Birds cannot `[MASK]`”），誤答を入力に付加することで誘導する（“Birds can `[MASK]`” → “Talk? Birds can `[MASK]`”）という 2 つの操作に対するモデルの振る舞いを調べ，BERT モデルがいずれの操作に対しても脆弱で，元の表現と否定表現とで同じ回答を行ってしまったり，誘導に従って誤答を生成してしまうことを確認した．Cao et al. (2021) は入力文の言い回しが予測に

与える影響についてさらに踏み込んで調査した．特定の関係についてLAMAと同様のテンプレートを用いて入力を作成したとき，モデルの予測分布が実際の正解分布によらず同様の偏りを見せることを確認し，言語モデルが個別のエンティティに関する知識よりも，テンプレートに強く影響を受けて予測を行う傾向があることを指摘した．

前述のとおり，LAMA probeにおける評価対象のうち，実世界の事実に関する知識に関連するものは主に英語Wikipediaを情報源とするデータセットに基づく．多くの主要な知識評価ベンチマークが同様にWikipediaを知識源としており，一般的なドメインにおける知識評価としては標準的なものといえる (Petroni et al. 2021, Thorne et al. 2018, Kwiatkowski et al. 2019, Yang et al. 2018)．一方で，より専門性の高いドメイン特化の知識はLAMA probeではカバーされていない．ドメイン特化の知識評価としては，生物医学分野の情報源をもとにしたBioLAMA (Sung et al. 2021)が開発されている．また，Kassner et al. (2021)はLAMAデータセットを多言語化し53言語で評価可能なMultilingual LAMAを作成した．ただし，データセットは元のLAMAデータセットの自動翻訳により構築されており，非英語圏に固有の知識が十分にカバーできていない可能性があることに注意が必要である．また，LAMAデータセットを構成する各データセットは2018年以前の知識に基づいている．したがって，最新の情報を含むテキストで学習された言語モデルの評価において，情報の更新に伴う不整合により本来正しい知識が過小に評価される可能性がある．

2.3 言語モデルの出力誤り

言語モデルの生成の質が向上する一方で，言語モデルが事実と異なる誤りを含む内容を自然な文章として生成してしまうハルシネーションと呼ばれる問題に注目が集まっている (Huang et al. 2024)．テキスト生成におけるハルシネーションの問題は事前学習済み言語モデルの発展以前から研究されており，文書要約のような条件付き生成の文脈での研究が主流であった．条件付き生成におけるハルシネーションは，モデルの出力が入力となる文章に忠実でなく，書かれていない内容を含んでしまうような場合を指す (Maynez et al. 2020)．これに対しては，入力文章への忠実性を向上することが主なアプローチとなる (Cao et al. 2018, Zhu et al.

2021)．一方，2.2.1 節で述べたように，大規模言語モデルの生成ではモデルが事前学習で獲得した知識をもとに，入力にない情報を出力する使われ方が一般的である．こうした文脈では，出力内容に事実としての誤りが含まれていることを指してハルシネーションと呼ばれることが多い．従来の入力文章に忠実でない誤りは内在的ハルシネーション (intrinsic hallucinations)，事実としての誤り，言いかえれば外部の情報ソースを用いて事実であることが確認できない内容を含む生成は外在的ハルシネーション (extrinsic hallucinations) と呼ばれる (Huang et al. 2024)．

実世界の事実に関する誤りは，訓練データ中に高頻度で出現するエンティティを誤って予測してしまう場合や，“Ikeda” という名前から日本出身であると推定するといったような属性に依存する偏りに起因する誤りなどに細分化される．また，事実に関する出力誤りの要因はモデルの知識不足だけでなく，入力文の曖昧性により正解が絞り込めない場合や訓練時の知識が古くなっている場合もある (Zhang and Choi 2021)．後者の問題への対処としては，言語モデルの知識編集や継続学習に基づくモデルの更新，外部の知識源を用いた知識拡張などがある (De Cao et al. 2021, Zhang et al. 2023)．

言語モデルの出力誤りの対処は主に，モデルの文生成時に誤りを抑制する方法と，生成された文を評価し誤りを判別する方法に分けられる．Li et al. (2023) は，モデルの内部状態を特定の方向に操作することで，モデルが不確かな出力を抑制するよう出力をコントロールする方法を提案した．別の方法としては，言語モデルへの指示により出力に思考過程を含めるよう誘導したり (Wei et al. 2023)，自身の出力に対するフィードバックを行わせて出力内容を改善する (Madaan et al. 2023) など，言語モデルへの指示方法を通じた出力の正確性改善の試みがある．また，外部の知識源を検索し参照しながら出力する Retrieval-Augmented Generation (RAG) と呼ばれる生成手法の利用も増加しており (Gao et al. 2024)，知識源中の根拠となる記述に忠実な生成を促進することで誤りを含む生成を抑制する効果が期待される．言語モデルによる生成後に評価を行うアプローチとしては，生成された内容が外部の知識源を用いて根拠づけできるかを評価する取り組みがある (Rashkin et al. 2023, Bohnet et al. 2022)．

2.4 確信度推定

言語モデルの出力誤りに対するアプローチの一つに，言語モデルの生成内容に対する確信度推定がある．生成された内容が正しい確率が高いときにスコアが高くなるような確信度指標を設計することで，利用者がモデル出力を信頼すべきかの判断材料としたり，アプリケーションにおける出力制御に用いることができる．本節では，統計的機械学習や深層学習における確信度推定の研究を概観した後，学習済み言語モデル出力の確信度推定の課題と近年の研究について述べる．

2.4.1 機械学習・深層学習における確信度推定

機械学習においてモデル予測の不確実性を捉えることは重要であり，誤りリスクが重大な医療画像解析 (Kurz et al. 2022) や自動運転のための物体認識 (Feng et al. 2018)，少量のラベル付きデータに基づく時系列予測 (Rußwurm et al. 2020) など応用は多岐にわたる．また，予測に基づく意思決定だけでなく，能動学習におけるデータ選択 (Lewis and Gale 1994) などにも確信度推定が用いられる．

不確実性の分類

Senge et al. (2014) は，機械学習における予測の不確実性の要因を偶然的不確実性 (aleatoric uncertainty) と経験的不確実性 (epistemic uncertainty) の 2 つに分類した．この分類は後の多くの研究で採用されている (Hüllermeier and Waegeman 2021)．偶然的不確実性は，データに起因しモデル側での対処が不可能な不確実性で，データ不確実性とも呼ばれる．入力データのノイズが大きい，低解像度の画像のように識別に必要な情報をデータから判別できないといった状況ではデータ不確実性が大きいといえる．これに対し，モデルの設計や学習方法，訓練データの選択やラベル付け方法といったモデル側に起因する不確実性は経験的不確実性，またはモデル不確実性と呼ばれる．モデル予測の不確実性にはデータ不確実性とモデル不確実性が複合的に影響する．

別の観点では，モデルの学習時と推論時のデータ分布に着目した分類も可能である． Gawlikowski et al. (2023) は，不確実性を推論時の入力が訓練データと同一の分布に従う場合 (in-domain)，学習時と異なる状況でデータを取得するなど学習時と推論時のデータ分布が異なる場合 (domain-shift)，学習時に想定して

いない全く未知のデータが入力される場合 (out-of-domain) に分類し既存研究のアプローチを整理している．

確信度推定

深層学習に基づく分類モデルでは，分類対象のクラスに対応するロジット z にソフトマックス関数

$$\mathrm{softmax}(z)_j = \frac{\exp(z_j)}{\sum_{k=1}^{K} \exp(z_k)} \tag{2.8}$$

を適用し，確率分布を直接算出することが一般的である．したがって，最も単純な方法としてはこの予測確率を直接確信度として用いることができる．しかしながら，深層学習モデルの予測確率は実際より過大に見積もられる傾向があることが指摘されている (Guo et al. 2017)．また，学習時の分布外のデータに対しても意図せぬ過大な確信度を与える可能性がある．

こうした問題への有効なアプローチの一つは，分布外のデータを特定して除外したり，確信度を抑制する方法である．Liang et al. (2018) は入力データへの摂動と temperature scaling を用いて分布外のデータを特定する方法を提案した．Papernot and McDaniel (2018) の手法は，ニューラルネットワークモデルへの入力に対する各中間層の表現パターンを，類似する学習データ中の事例と比較することでデータが学習時の分布から外れているかを判別する．Sensoy et al. (2018) はクラス確率分布を Dirichlet 分布で表現することで，分布外のデータに対する予測確率の挙動を改善している．これらの手法はモデルの予測確率分布に干渉し，単純なソフトマックス関数による予測確率の改善を試みている．別のアプローチとして，予測とは別に確信度推定専用のモジュールを用意し，確信度推定を直接学習する方法がある (Raghu et al. 2018)．

ベイジアンニューラルネットワークは，モデルパラメータ θ を確率変数として扱い，データ不確実性とモデル不確実性を明示的に区別したモデル化を行う．訓練データ中の入出力の組 (x, y) に対し，モデルパラメータ θ の事後分布は次のように表せる:

$$p(\theta|x, y) = \frac{p(y|x, \theta)p(\theta)}{p(y|x)} \propto p(y|x, \theta)p(\theta). \tag{2.9}$$

分母の $p(y|x)$ は正規化定数で,

$$p(y|x) = \int p(y|x,\theta)p(\theta)d\theta \tag{2.10}$$

が成り立つ．新たな入力 x^* に対する y^* の予測確率は以下のように，モデル予測 $p(y^*|x^*,\theta)$ をパラメータの事後分布 $p(\theta|x,y)$ で周辺化した形で表せる:

$$p(y^*|x^*,x,y) = \int p(y^*|x^*,\theta)p(\theta|x,y)d\theta. \tag{2.11}$$

式 2.11の右辺を厳密に求めることはできないため，実用上は何らかの近似手法を適用する必要がある．推論時にモデルに Dropout を適用する MC Dropout (Gal and Ghahramani 2016) も近似手法の一種とみなすことができ，実装の容易さから広く用いられている．

以上に挙げた手法はいずれもモデルの学習時点で確信度推定を考慮した設計を行うか，学習済みモデルとは別の確信度推定モデルを追加の訓練データにより学習する必要がある．所与の学習済みモデルに対し確信度推定を行う方法としては，推論時の入力データに何らかのデータ拡張を適用し，複数の入力に対する予測結果の整合性を見る方法がある (Ayhan and Berens 2018, Wang et al. 2019)．

2.4.2 言語モデル出力の確信度推定

言語モデル出力における確信度推定には次のような特徴がある．

- 言語モデル生成では文脈情報を入力として単語系列を予測する．語彙サイズは大きいもので数十万トークンになる場合があり，系列長に応じて組み合わせは指数的に増加する．一般的な分類タスクと比較して非常に大きな出力空間を扱う必要がある．
- 単語や概念によってコーパス中の出現頻度に大きな差があり，コーパス中のごく少数にしか出現しないロングテールの単語や概念が多数存在する (Kandpal et al. 2023)．コーパス中に記載されている内容であっても，出現頻度が少ない場合には言語モデルが知識として記憶できていない場合がある．

- 同じ意味を表す複数の表現が存在するなど，ある文脈に対し適切な出力が一意に定まらない場合が多い．
- 言語モデルはラベルの無いコーパスを用いて自己教師あり学習により訓練される．したがって，教師ラベルに基づく分類を前提とした確信度計算方法を単純に適用できない場合がある．
- 既存の学習済み言語モデルに対する確信度推定を考える場合，確信度を考慮した学習方式を伴うアプローチの適用可能性は限定的である．ただし，少量のモデルパラメータのみの変更を伴うファインチューニング (Han et al. 2024) などを用いることにより，モデルの多様性を確保することが可能な場合がある．

こうした特性から，他分野で有効な確信度推定手法が言語モデル出力の確信度推定でも有効かは必ずしも明らかではない．また，近年では言語生成の特性を活かした確信度推定方法も数多く提案されている．ここでは近年の言語生成における確信度推定のアプローチについて説明する．

言語モデルの確信度指標として最も素朴なものはモデルが出力する単語予測確率を用いる方法である．BERT などの事前学習済みモデルでは，訓練データと評価タスクのドメインが一致する場合にはモデルの予測確率が実際の正答率をよく反映しているが，異なるドメイン間では乖離が大きくなるという報告がある (Desai and Durrett 2020)．

言語モデルから得られる情報の範囲はモデルの利用者や開発者といった立場によって異なり，利用する情報の範囲に応じた確信度推定手法が開発されている (Geng et al. 2024)．モデルの出力テキストのみを参照できるブラックボックス条件下では，“Obviously”, “I’m not sure” といった言語表現や “90%” のような数値表現を用いて言語モデルに出力の確信度を表現させ，これを手がかりとして用いる方法が考えられる．しかし，特に調整を行わない場合，言語表現に基づく確信度は実際の正答率と比べて過剰に高くなる傾向があり，何らかの調整が必要であるとされる (Mielke et al. 2022, Xiong et al. 2024)．ただし，より大きいモデルサイズにおいては言語表現に基づく確信度の信頼性が高くなるという報告もある (Kadavath et al. 2022)．同一の入力に対する複数の出力を生成して出力間の一貫性を確認する方法も有効とされる方法である (Manakul et al. 2023)．

モデルの内部状態を参照可能な条件下では，モデルの埋め込みや注意機構の内部状態を特徴量とし，少量の追加データを用いてモデルが正答できる入力を判別するといった方法が提案されている (Ren et al. 2023, Kadavath et al. 2022)．いずれも訓練内外のドメインのデータや正誤ラベル付きデータといった追加のデータを必要とする方法である．別のアプローチとしては，言語モデルに自身の過去の出力を含むプロンプトを繰り返し与えて振舞いの推移を見ることで，言語モデルの知識不足に起因する誤りを特定する方法が提案されている (Abbasi-Yadkori et al. 2024)．

2.5 選択的予測

選択的予測 (El-Yaniv and Wiener 2010, Geifman and El-Yaniv 2017) は，機械学習において分類モデルによる予測の出力可否を選択的に判断する仕組みを伴う枠組みである．入力空間 $\mathcal{X}$ におけるラベル空間 $\mathcal{Y}$ への分類問題を考える．**選択的分類器** (f, g) は，元となる分類モデル $f : \mathcal{X} \to \mathcal{Y}$ と**選択関数** $g : \mathcal{X} \to \{0, 1\}$ から構成される．入力事例 $x \in \mathcal{X}$ が与えられたとき，選択関数はシステムが予測 $f(x) \in \mathcal{Y}$ を出力するかを判定する:

$$(f, g)(x) := \begin{cases} f(x) & \text{if } g(x) = 1 \\ \textit{don't know} & \text{if } g(x) = 0 \end{cases}. \tag{2.12}$$

Geifman and El-Yaniv (2017) は，確信度指標に基づく選択関数を用いるリスク保証付きの選択 (Selection with Guaranteed Risk; SGR) を導入した．選択関数 $g_\beta(x)$ を次のように定める:

$$g_\beta(x) = \begin{cases} 1 & \text{if } \phi(x) \geq \beta \\ 0 & \text{if } \phi(x) < \beta \end{cases}. \tag{2.13}$$

ここで， $\phi(x) : \mathcal{X} \to \mathbb{R}$ は**確信度関数**である．システムは，確信度の値 $\phi(x)$ が閾値 $\beta \in \mathbb{R}$ を超えたときに予測を出力する．閾値 β によりシステムが出力するエラーの許容リスクを調整することができ，β が大きければ大きいほど，システムが予測する事例の数が少なくなり，誤った予測のリスクを低減する．このよう

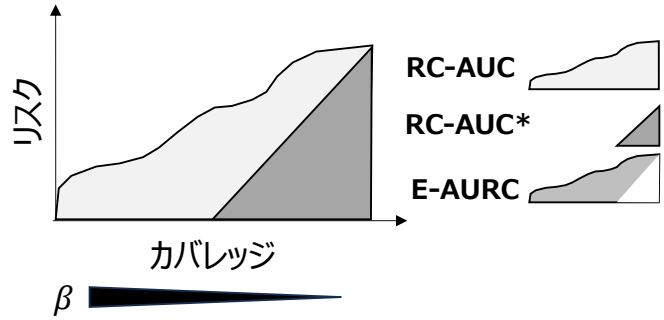


図 2.3: 評価指標 RC-AUC および E-AURC の関係図.

に，リスク保証付き選択の設定のもとでは，選択的分類器が誤った出力を行うリスク

$$\mathrm{Risk}(f, g_\beta) = (N_{\mathrm{pred}}(f, g_\beta) - N_{\mathrm{corr}}(f, g_\beta))/N_{\mathrm{pred}}(f, g_\beta) \tag{2.14}$$

とシステムが回答可能な予測の割合（カバレッジ）

$$\mathrm{Coverage}(f, g_\beta) = N_{\mathrm{pred}}(f, g_\beta)/N \tag{2.15}$$

がトレードオフの関係になる．なお，$N, N_{\mathrm{pred}}(f, g_\beta), N_{\mathrm{corr}}(f, g_\beta)$ はそれぞれ全ての入力事例数，選択的分類器 (f, g_β) が予測を出力した事例数，正しく予測された事例数を表す．選択的分類器の性能は，閾値 β を動かすことで得られるリスク-カバレッジ曲線の面積 (RC-AUC) によって評価される．RC-AUC が小さいほど誤った予測を低減しつつ多くの予測が可能である良いシステムであることを意味する．実用上は，適切な閾値 β はユーザ側で決定することになるが，RC-AUC を計算する上では β を決定する必要がないことに注意する．

あるモデル予測に対する RC-AUC の下限値 RC-AUC* （以下，オラクル確信度）は，正解事例で 1, 不正解事例で 0 の値をとるような最適な確信度指標により達成される．RC-AUC から下限値 RC-AUC* を引いた差分 E-AURC (Geifman et al. 2019) は，モデルの予測性能による影響を差し引いて確信度の性能のみを把握するのに役立つ．図 2.3に評価指標の関係図を示す.

第 3 章

選択的予測に基づく言語モデルの知識評価

本章では，言語モデルの知識評価に 2.5 節で導入した選択的予測の枠組みを導入する．事前学習済み言語モデルが訓練時に獲得した知識を評価するベンチマークタスクである LAMA probe (Petroni et al. 2019) では，言語モデルに個別の知識に関して記述した文の穴埋めタスクを解かせる．言語モデルが正しく穴埋めできれば当該知識を保持しているという仮定のもと，モデルのもつ知識量は予測精度に基づいて評価される．しかしながら，こうした知識評価の方法には 2.2.2 節で述べたようないくつかの課題がある．本章では，次の 2 点の課題に着目する．

予測や評価データの偏りの影響　第一の課題は，予測精度に基づく評価が必ずしもモデルの知識量を正確に反映しないことである．図 3.1は，LAMA データセットにおける `place-of-birth` 関係に関する事例サブセットに対する BERT-base モデルの予測トークンの分布を示している．入力文は “`<subject>` was born in `[MASK]` .” という共通のテンプレートにそれぞれの関係知識の `subject` エンティティを入れたものである．この関係に対する BERT モデルの正答率は 14.9%となっているが，正解事例，不正解事例ともに 5 種類の予測単語が全体の半数以上を占めており，いずれも “London” の予測頻度が最も高く，他にも “Paris”，“Chicago” など共通の単語が高頻度で予測されている．このことは，特定のテンプレートに対するモデルの予測が実際の正解分布に依存しない偏りをもつというCao et al. (2021) の観察からも支持されるように，モデルが一部の事例について個別のエンティティに関する知識をもたないまま「推測」を行っている可能性を示唆している．こうしたケースでは，モデルが個別のエンティティの知識をもたずにテンプ

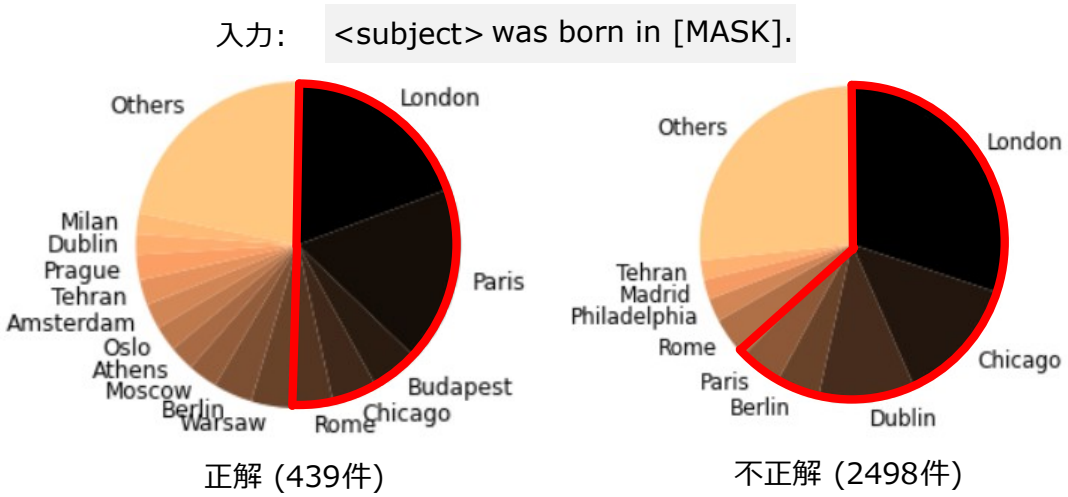


図 3.1: LAMA データセット中の `Google-RE:place-of-birth` 関係のサブセット（全 2937 件）に対する BERT-base モデルの予測単語分布．左が正答，右が誤答事例の単語分布．それぞれ予測頻度上位 5 件の単語を赤枠で示している．

レートから推測した回答が偶然正解することにより，モデルがもつ知識の過大評価が生じ得る．こうした問題への対処としては，モデルへの入力パターンの多様化や推測が容易な事例の除外といった評価データ側を改善するアプローチが主にとられてきた (Jiang et al. 2020, Poerner et al. 2020)．これに対し本研究では，評価の枠組みを改善することにより，データに依存しない方法でこの問題を緩和することを目指す．

誤りリスクの考慮　第二の課題は，予測精度に基づくモデル知識評価には予測の誤りリスクの観点が無いことである．現在，事前学習済み言語モデルの利用範囲は，医療や法律といった高度な専門性を要する分野における意思決定支援にまで拡大しつつある (Jin et al. 2024, Zhan et al. 2024, Jiang et al. 2024)．こうした状況下で，言語モデルが誤った知識に基づいた出力を行い，それに基づいて意思決定が行われた場合のリスクは重大なものになり得る．そのため，言語モデルの出力の正しさだけでなく，出力が誤りである可能性がある場合にそれを検知し起こり得るリスクを回避できるかどうかが重要となる．こうした実応用上の背景を踏まえると，全てのモデル予測を採用した上での予測精度のみに基づいて知識評価を行うだけでは不十分な場合がある．

本章では，LAMA probe の評価に選択的予測を導入することでこれらの課題の解決を試みる．選択的予測の枠組みにおいては，評価時にモデル予測の確信度を考慮することにより，予測結果が疑わしい事例を検知し，誤った出力を抑制することが可能となる．システムは，モデル予測の正しさに加え，確信度指標を用いて誤った予測を適切に除外することができる能力を評価される．実験では，LAMA probe で比較的高い性能を示したマスク言語モデルを対象とし，選択的予測に基づきモデル評価を行うことで，各モデルの予測に対し確信度を適切に推定することが可能かを調べる．確信度指標としては，言語モデルの入出力や内部状態を用いて計算可能な複数の指標を定義し，確信度指標による性能差も検証する．さらに，選択的予測に基づく評価と従来の予測精度に基づく評価とを比較し，予測や評価データの偏りによる影響を調べる．

3.1 言語モデル出力への選択的予測の導入

本章では，主に予測位置の前後双方の文脈を考慮して予測を行うマスク言語モデルを想定した議論を行う．また，確信度指標もマスク言語モデルを想定して設計する．しかし，評価の枠組みや確信度指標の大部分はモデル非依存であり，デコーダを基礎とする言語モデルにも適用可能である．

言語モデルの予測　t 番目の位置の単語がマスクされた入力文

$$c_t := (w_1, \ldots, w_{t-1}, \texttt{[MASK]}, w_{t+1}, \ldots, w_{|W|}) \tag{3.1}$$

に対し，言語モデルは t 番目の単語に対する確率分布 $P_{\mathrm{LM}}(w|c_t)$ を予測する．特に断らない限り，モデルの予測はこの確率分布上で確率最大の単語 w' とする:

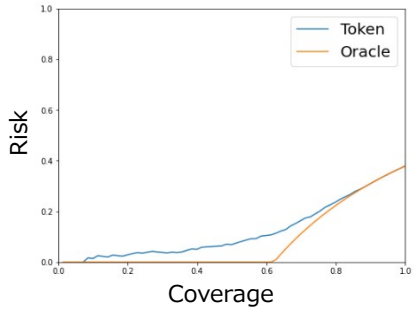
$$w' := \arg\max_w P_{\mathrm{LM}}(w|c_t). \tag{3.2}$$

入力文のマスク位置が予測単語 w' で埋められた文を W' と表記する．

選択的予測に基づく評価　2.5 節で導入した Geifman and El-Yaniv (2017) のリスク保証付きの選択 (SGR) に基づき，選択的予測の枠組みのもとで言語モデルの

P36 (“The capital of <subject> is [MASK] .”)

精度: 0.621　RC-AUC: 0.121



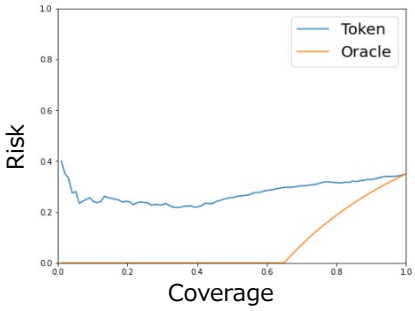
確信度上位の予測

Subject	正解	予測	ϕ_T
Sri Lanka	Colombo	Colombo	-0.001
Bratislava Region	Bratislava	Bratislava	-0.001
Albania	Tirana	Tirana	-0.002
Tirana District	Tirana	Tirana	-0.002
Hiroshima Prefecture	Hiroshima	Hiroshima	-0.002

(a) T-REx:P36

P1412 (“<subject> used to communicate in [MASK].”)

精度: 0.650　RC-AUC: 0.278



確信度上位の予測

Subject	正解	予測	ϕ_T
Adrianus Valerius	Dutch	Latin	-0.490
Muhammad Ali	English	Arabic	-0.575
Gloria Estefan	Spanish	Spanish	-0.587
Imre Nagy	Hungarian	Hungarian	-0.610
Sextus Pompeius Festus	Latin	Latin	-0.619

(b) T-REx:P1412

図 3.2: T-REx データセットの 2 種類の関係に関する事例集合と評価の例．予測に用いたモデルは BERT-base．関係ラベル P36 は “capital”，P1412 は “languages spoken, written or signed” の関係である．各画像左は `Token` 指標およびオラクル確信度 (`Oracle`) に基づくリスク-カバレッジ曲線．右は `Token` 確信度上位 5 件の予測．網掛け行は誤った予測を示す．

評価を行う．すなわち，言語モデル f の予測 $f(c_t) = w'$ に対し，確信度関数 ϕ を用いて計算される確信度スコア $\phi(w', c_t)$ に基づき予測の出力可否を決める選択的分類器を考え，データセット中の全事例の確信度スコアから計算される RC-AUC に基づきシステムの性能を評価する．

選択的予測に基づく評価の具体例として，BERT-base モデルによる T-REx データセットの 2 種類の関係に関する事例集合に対する予測と評価結果を図 3.2に示す．なお，ここでは確信度として単語予測尤度に基づくスコアを用いている（定義は式 3.3を参照）．いずれの事例集合でも予測精度は 0.6 前後であるが，図 3.2aの

事例集合 (P36) では正解である予測の確信度を高く見積もることができており，逆に図 3.2bの事例集合 (P1412) では誤った予測の確信度が高くなっている．選択的予測に基づく評価ではこの差を捉えることができており，確信度を適切に見積もり誤りリスクを低減できる場合に RC-AUC の値がより低くなる．

3.1.1 確信度関数

言語モデルの知識評価タスクにおいて用いる確信度関数を導入する．ここでは，外部の知識源を用いることなく，言語モデルと入出力の情報のみから計算可能な確信度指標を考える．

`Token`

最も素朴な確信度関数としては，予測単語 w' の対数尤度を直接確信度として用いることができる:

$$\phi_{\mathrm{T}}(w', c_t) = \log P_{\mathrm{LM}}(w'|c_t). \tag{3.3}$$

`Sent`

予測単語でマスク位置を埋めた文 W' の文としての確からしさを確信度の指標として用いることを考える．マスク言語モデル出力は対象位置を除く全ての文脈情報を与えたうえで各位置の予測スコアを算出する点で逐次的なデコードを行う通常の言語モデルと異なるため，厳密な文確率を算出することができない．ここでは， Salazar et al. (2020) が提案したマスク言語モデルの疑似対数尤度スコア (pseudo-log-likelihood scores; PLLs) を文長で正規化したものを採用する:

$$\phi_{\mathrm{S}}(w', c_t) = \frac{1}{|W'|} \sum_{u=1}^{|W'|} \log P_{\mathrm{LM}}(w_u|c_u). \tag{3.4}$$

`Gap`

予測確率最大の予測単語 w' と，予測確率が 2 番目に大きい予測単語 w'' の対数尤度の差を比較する．この差が大きければ予測単語の他に有力な候補が存在しな

いことになり，確信度が高いとみなす:

$$\phi_{\mathrm{G}}(w', c_t) = \log P_{\mathrm{LM}}(w'|c_t) - \log P_{\mathrm{LM}}(w''|c_t). \tag{3.5}$$

Reranking

確信度スコアは一部を除き，予測単語 w' 以外の候補単語についても同様の方法でスコアを算出することが可能である．全ての候補単語について異なる指標に基づく確信度スコアを計算したとき，単語予測確率に基づくスコアとは異なる候補単語が予測単語のスコアを上回る場合がある．こうした場合には，計算方法によって予測にぶれが生じることになる．ここでは式 3.2の単語予測確率に基づき上位 K 件の単語 $\mathcal{W}$ を対象に，式 3.4の文レベル確信度 (Sent) で確信度スコアを再計算して並び替えたとき，元の予測単語 w' の順位 $\mathrm{rank}_\psi(w')$ が低下する場合には確信度が低くなるような Reranking 指標を導入する:

$$\phi_{\mathrm{R}}(w', c_t) = \log_2 \frac{K}{\mathrm{rank}_\psi(w')} = \log_2 K - \log_2 \mathrm{rank}_\psi(w'). \tag{3.6}$$

なお，Reranking スコアは言語モデルが訓練時のテキストを直接記憶しているかを評価する暴露スコア (Carlini et al. 2019) を参考にしている．実験では $K = 100$ を用いた．

Dropout

Dropout に基づく確信度は深層学習モデルの不確実性評価に広く用いられている (Gal and Ghahramani 2016)．Dropout (Srivastava et al. 2014) は通常モデルの訓練時に過学習を防ぐ目的で，ニューラルモデルのパラメータの一部を欠落させる方法である．不確実性評価の文脈においては，推論時にモデル中間層のニューロンを一定の割合 p でランダムに選択し，それらのニューロンの出力を 0 で上書きする（dropout マスク）ことで，モデルの推論に摂動を加える．M 通りの異なる dropout マスクを適用することで得られる M 通りの出力の統計量から予測の安定性を測る．$P_{\mathrm{LM}}^{(m)}(w'|c_t)$ を m ($m \in \{1, \ldots, M\}$) 番目の dropout マスクに基づく出力とする．Dropout に基づく指標としては，Kamath et al. (2020) の方法に従い 2 種類の指標を導入する．

DropoutMean (DM)　は M 通りの出力の予測確率の平均をとる:

$$\phi_{\mathrm{DM}}(w', c_t) = \frac{1}{M} \sum_{m=1}^{M} P_{\mathrm{LM}}^{(m)}(w'|c_t). \tag{3.7}$$

DropoutVar (DV)　は予測の負の分散をとる:

$$\phi_{\mathrm{DV}}(w', c_t) = -\frac{1}{M} \sum_{m=1}^{M} (P_{\mathrm{LM}}^{(m)}(w'|c_t) - \phi_{\mathrm{DM}}(w', c_t))^2. \tag{3.8}$$

実験では $M = 30$ 通りの dropout マスクを適用した結果を用い，dropout マスクの割合は各モデルの訓練時のものと同一とした．

TemplateDiff

次に導入するのは，知識評価タスクのうち (subject, relation, object) の三つ組で表現される関係知識を問う入力事例を想定した確信度推定方法である．2.2.2 節で述べたように，LAMA probe のうち三つ組の関係知識を問うデータセットでは，関係の種類ごとに決められたテンプレートに基づき入力が作成されている．たとえば出生地に関する入力文のテンプレートは “<subject> was born in [MASK] .” となり，<subject>に個別の主語にあたるエンティティが挿入される．Cao et al. (2021) は言語モデルによる予測がこうしたテンプレートの影響を強く受け，主語にあたるエンティティによらない偏った生成を行う傾向があることを指摘している．

上記のような背景から，言語モデルによる予測にテンプレートだけでなく主語エンティティの情報が貢献しているかを定量化した確信度指標を定義する．入力文 c_t の主語エンティティにあたる箇所をマスク単語に置き換え，テンプレート情報のみを残した文を W_{temp} とする．たとえば (Dante, place-of-birth, Florence) という三つ組に対応する入力文 “Dante was born in [MASK] .” に対し，置き換え後の文は “[MASK] was born in [MASK] .” となる．言語モデルに元の入力文 c_t と主語を隠した入力文 W_{temp} の両方を与え，予測単語の対数尤度の差分をとったも

のを確信度とする:

$$\phi_{\mathrm{TD}}(w', c_t) = P_{\mathrm{LM}}(w'|c_t) - P_{\mathrm{LM}}(w'|W_{\mathrm{temp}}). \tag{3.9}$$

3.2 実験

選択的予測の導入が言語モデルの知識評価にどのように影響するかを確認するため，LAMA probe を用いた評価を行う．第一に，3.2.1 節で従来の精度評価と選択的予測に基づく評価を予測の偏りに対する頑健性の観点から比較する．第二に，3.2.2 節では 3.1.1 節で導入した異なる確信度関数を適用し，3 種類のマスク言語モデルにおける確信度を考慮した知識評価を行う．

評価対象の言語モデルとしては，モデルサイズの異なる 2 種類の BERT モデル (Devlin et al. 2019) (BERT-base, BERT-large) と RoBERTa-base(Liu et al. 2019) を用いる．いずれも文中の予測位置の前後の文脈を考慮して予測を行うマスク言語モデルで，パラメータサイズはそれぞれおよそ 11 億，34 億，12 億である．LAMA probe のデータセットのうち Google-RE, T-REx, SQuAD の 3 つは Wikipedia が知識源となっており，実験対象のモデルは Wikipedia を訓練データの一部として用いていることから，評価データ中の知識を裏付ける記述が訓練データに含まれていると考えられる．

3.2.1 テンプレート起因のバイアスに対する頑健性

図 3.1でも例示したように，精度に基づく知識評価は偶然の一致による影響を受け，その影響は予測がテンプレートの影響を受けて偏る場合に顕著になる．本節ではこうしたテンプレートに関連する偏りに評価がどの程度影響を受けているかを確認する．ここでは LAMA probe の T-REx データセットを用いる．データセットは 41 種の関係に関する約 3 万 4 千件の事例を含み，関係の種類ごとに入力文のテンプレートが用意されている．

まず，テンプレートに起因するバイアスを定量化するための 2 つの指標として**予測カバレッジ**と**正解カバレッジ**を導入する．予測カバレッジは特定のテンプレートに対するモデル予測の偏りを定量化するものである．あるテンプレートに対するモデル予測に偏りがあるとき，モデルは個々の主語の情報を見ずにテンプ

		BERT-base		BERT-large		RoBERTa-base	
		$\mathrm{Cov}^{\mathrm{A}}$	$\mathrm{Cov}^{\mathrm{P}}$	$\mathrm{Cov}^{\mathrm{A}}$	$\mathrm{Cov}^{\mathrm{P}}$	$\mathrm{Cov}^{\mathrm{A}}$	$\mathrm{Cov}^{\mathrm{P}}$
精度		0.387	-0.247	0.469	-0.244	0.512	-0.224
−RC-AUC	`Token`	0.344 ↓	-0.316 ↓	0.438 ↓	-0.292 ↓	0.484 ↓	-0.277 ↓
	`Sent`	0.355 ↓	-0.290 ↓	0.441 ↓	-0.285 ↓	0.499 ↓	-0.249 ↓
	`Gap`	0.351 ↓	-0.314 ↓	0.430 ↓	-0.294 ↓	0.474 ↓	-0.285 ↓
	`Reranking`	0.350 ↓	-0.286 ↓	0.452 ↓	-0.283 ↓	0.498 ↓	-0.266 ↓
	`DropoutMean`	0.338 ↓	-0.319 ↓	0.433 ↓	-0.293 ↓	0.486 ↓	-0.280 ↓
	`DropoutVar`	0.419 ↑	-0.125 ↑	0.470 ↑	-0.124 ↑	0.456 ↓	-0.166 ↑
	`TemplateDiff`	0.349 ↓	-0.317 ↓	0.427 ↓	-0.299 ↓	0.486 ↓	-0.271 ↓

表 3.1: T-REx データセットにおける各評価指標とテンプレートバイアス指標（正解カバレッジ ($\mathrm{Cov}^{\mathrm{A}}$), 予測カバレッジ ($\mathrm{Cov}^{\mathrm{P}}$)）の相関．分かり易さのため，RC-AUC については正負を反転した値との相関を記載している．

レートに影響を受けた生成を行っている可能性がある．$\mathcal{D}_r = (\{(s_i, o_i)\}_{i=1}^{N_r}, t_r)$ を N_r 個の三つ組関係 (s_i, r, o_i) を含む事例集合とする．また，関係 r に対応するテンプレートを t_r とする．i 番目の事例 (s_i, o_i) に対応する入力文を $t_r(s_i)$ とする．まず，各事例集合 $\mathcal{D}_r$ に対し評価対象の言語モデルで予測を行い予測単語の頻度上位 k 件 $\mathcal{W}^{\mathrm{freq}}(r)$ を得る．ここでは $k = 5$ とした．これらの単語が予測全体に占める割合を予測カバレッジとする:

$$\mathrm{Cov}^{\mathrm{P}}(r) = \frac{|\{i \mid f(t_r(s_i)) \in \mathcal{W}^{\mathrm{freq}}(r)\}|}{N_r}. \tag{3.10}$$

予測カバレッジが大きいことは，モデル予測はテンプレートの影響を受け偏っていることを示唆する．

正解カバレッジは $\mathcal{W}^{\mathrm{freq}}(r)$ が $\mathcal{D}_r$ の正解単語に占める割合を指す:

$$\mathrm{Cov}^{\mathrm{A}}(r) = \frac{|\{i \mid o_i \in \mathcal{W}^{\mathrm{freq}}(r)\}|}{N_r}. \tag{3.11}$$

正解カバレッジは評価データの偏りを表す．評価データの正解事例が数種類の単語でカバーできるとき，モデルは個々の主語に関する知識がなくとも偶然の一致により正答できる可能性が高くなる．

T-REx データセットの 41 種の関係別事例集合を対象に，言語モデル毎の予測

カバレッジおよび正解カバレッジと各種評価指標に基づく評価スコアを計算し，相関をとる．テンプレートの偏りを示す2つの指標と評価指標との間に正の相関（RC-AUCの場合は負の相関）があるとき，その評価方法は予測やデータセットの偏りにより言語モデルの能力を過大評価していると解釈できる．表3.1に結果を示す．ここでは精度評価と一貫させるため，RC-AUCは符号を反転してから相関をとっている．表から，精度評価と比較して`DropoutVar`以外の全ての確信度指標において，選択的予測に基づくRC-AUCの評価は正解カバレッジの相関が小さくなっており，また予測カバレッジでは強い負の相関を示している．このことから，選択的予測に基づく評価はテンプレートに起因する予測の偏りによる言語モデル知識の過大評価を低減できていると考えられる．

3.2.2 選択的予測によるモデル評価

評価結果概要

表3.2に異なる確信度指標でのRC-AUC評価を示す．最も単純な`Token`指標が一貫して高い性能を示す一方，最高性能を示す確信度指標はモデル・データセット毎に異なる．`Gap`と`TemplateDiff`はWikipediaに基づく三つ組の関係を扱うGoogle-REとT-REx で性能が高い傾向にあり，BERT-baseにおける`Gap`，RoBERTa-baseにおける`TemplateDiff`がそれぞれ`Token`指標の性能を上回った．

確信度指標間の比較

`Token`指標の性能を上回ったモデルと指標の組み合わせについて，`Token`指標との指標間比較を行った結果を表3.3に示す．BERT-baseにおいて`Gap`が`Token`より高い性能を示すケースは，予測精度が高い，つまり難易度が低い事例であった．逆にRoBERTa-baseにおいて`TemplateDiff`が`Token`より高い性能を示すケースは，予測精度が低い難易度が高い事例であった．`Gap`の性能が高い事例では予測カバレッジが低い傾向もみられた．これは，`Gap`指標がその定義から過剰に予測確率が高くなるケースを除外することが困難であることに起因すると考えられる．

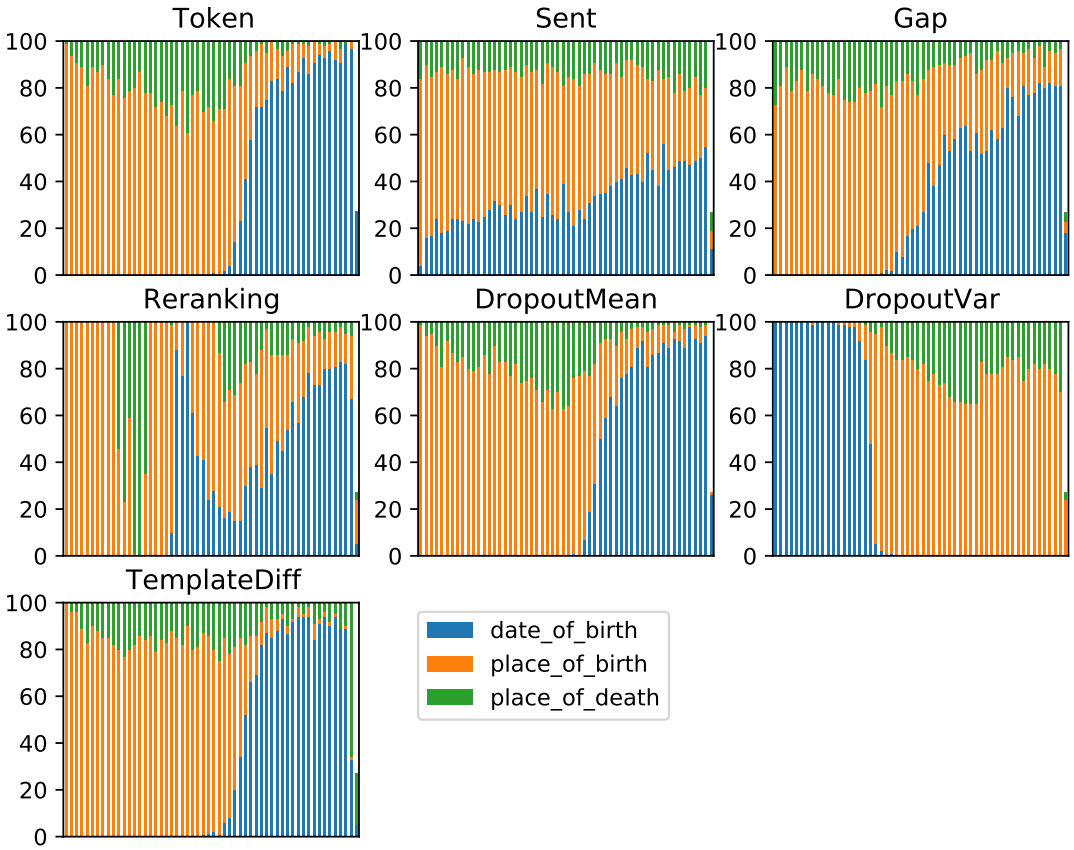


図 3.3: BERT-base による Google-RE データセット予測の関係テンプレートによる内訳．左から確信度スコアの大きい順にソートしている．

関係テンプレートと確信度関数

図 3.3は，Google-RE に含まれる 3 種類の関係 (`date-of-birth`, `place-of-birth`, `place-of-death`) に関する事例をそれぞれの確信度スコア順にソートし可視化したものである．ここでの予測モデルは BERT-base を用いている．`Token` 指標の図から，BERT-base モデルは `place-of-birth` の事例に対しては高い予測確率を，`date-of-birth` の事例に対しては低い確率を付与する傾向があることがわかる．このように，単語予測尤度に基づく確信度指標には関係テンプレート依存の強いバイアスが存在し，テンプレートによってはエンティティによらず高い確信度を付与する可能性があることに注意が必要である．

`Token` のほかに `Gap`, `DropoutMean`, `TemplateDiff` も同様に関係テンプレート

に依存した確信度分布を示している．一方，Sent と Reranking は関係の種類への依存性が比較的小さかった．DropoutVar は Token などと逆の傾向を示している．

関係テンプレートへの依存性が高い Token と依存性が低い Reranking の 2 つの指標について，確信度スコア上位の予測の単語頻度に基づく比較を行った結果を表 3.4に示す．date-of-birth では予測単語に大きな違いはなく，数種類の単語に予測が偏っている．他 2 つの関係タイプにおいては，確信度指標によって傾向が異なる．確信度が高く判定された予測単語は異なっているものの，各関係タイプにおいて特定の少数の予測への偏りはいずれのケースでもみられる．place-of-birth では，いずれの指標でも 5 つの単語が上位の予測の半数以上を占めており，place-of-death では，単一の単語が上位の予測の 4 割を占めている．以上のことから，Reranking 指標は関係テンプレート間のスコアの偏りは小さいが，各関係テンプレート内での特定の予測単語への偏りは解消していないことがわかる．

各指標を予測に用いた場合

ここまでの実験では，常に式 3.2に従い予測尤度最大の単語 w' を予測単語として扱ってきた．しかしながら，本章で導入した確信度指標は一部を除き直接予測単語を決定するために用いることも可能である．そこで，式 3.2で $P_{\mathrm{LM}}(w_t|c_t)$ の代わりに 3.1.1 節で導入した確信度関数を用いることで予測単語を決めることを考え，各指標を予測の決定に用いた場合に予測精度 (P @ 1) の改善に効果があるかを確認する．ただし， Gap 指標は式 3.5の定義では予測尤度最大の単語にしか適用できないので，式を他の候補単語に適用できるよう拡張する．具体的には，$w^{(k)}$ を予測確率が k 番目に大きい候補単語としたとき，以下のようにする:

$$\phi_{\mathrm{G}}(w', c_t) = \frac{1}{k}(\log P_{\mathrm{LM}}(w^{(k)}|c_t) - \log P_{\mathrm{LM}}(w^{(k+1)}|c_t)). \tag{3.12}$$

なお，予測確率が最下位の単語のスコアは 0 とする．また，Sent スコアの計算には各事例ごとに $\mathcal{O}(|W'| \cdot V)$ の順方向計算を要する（V は語彙サイズ）ことから，計算コスト緩和のため，語彙候補を Token スコアに基づく上位 100 件に絞ることとした．

結果を表 3.5に示す．全てのモデルにおいて，DropoutMean が最も高い性能を

達成したが，DropoutVar を除く全ての指標の性能はほとんど変わらなかった．確信度指標として用いた場合と異なり，Gap や TemplateDiff が T-REx データセットで優位になることもなかった．よって，各指標を直接予測に用いる場合において，これらの指標間に大きな性能差はみとめられず，確信度指標として用いた場合の性能と予測に用いた場合の性能に強い関連性は確認できなかった．このことから，予測と確信度推定に用いる指標に求められる性質は異なることが示唆される．

3.3 本章のまとめ

本章では LAMA probe による言語モデルの知識評価に選択的予測を導入し，モデル出力の確信度をもとに誤りの可能性がある予測を除外する能力を含めた知識評価を行うことを提案した．また，モデルの入出力や内部状態に基づく複数の確信度指標を導入し，確信度推定における性能を比較評価した．実験結果からは以下のことが示唆された:

- 選択的予測による評価は，従来の予測精度に基づく評価と比較してモデル予測や評価データの偏りによる過大評価を低減することができる (表 3.1).
- 最適な確信度指標はデータセットにより異なり，各確信度指標が特定の関係の種類や単語に対し偏った評価をする傾向がある (図 3.3).
- 各指標を予測に用いた場合の性能と，確信度指標として用いた場合の性能には強い関連性がない（表 3.5）.

本実験の範囲では，タスクやデータセットによらず最も単純な Token 指標よりも有効といえる指標は確認できなかった．タスクやモデルと確信度指標の関係についてさらに分析することで適切な確信度指標を選択することや，より汎用性の高い確信度指標の設計は今後の課題である．

表 3.2: RC-AUC によるモデル評価結果．値は小さいほど良い．T-REx データセットについては，関係の種類により細分化した結果を併記している．1-1 は一対一関係，N-1 は多対一関係，N-M は多対多関係の結果を示す．“Oracle” はオラクル確信度（2.5 節）を表す．TemplateDiff は事例の subject ラベルを用いるが，ラベルが未定義の SQuAD データでは計算できないため除外する．

(a) BERT-base

確信度	Google-RE	T-REx				ConceptNet	SQuAD	全体
		1-1	N-1	N-M	All			
Token	**.775**	**.118**	.434	.611	.478	**.686**	**.755**	.545
Sent	.834	.163	.549	.776	.594	.797	.815	.652
Gap	.798	.133	**.422**	.604	**.470**	.714	.794	.548
Reranking	.835	.248	.580	.623	.597	.834	.798	.633
DropoutMean	**.775**	.123	.425	.609	.473	.690	.762	**.543**
DropoutVar	.962	.525	.834	.883	.850	.918	.912	.886
TemplateDiff	.778	.119	.427	**.603**	.472	.782	-	-
Oracle	.663	.070	.301	.456	.344	.551	.583	.413

(b) BERT-large

確信度	Google-RE	T-REx				ConceptNet	SQuAD	全体
		1-1	N-1	N-M	All			
Token	.763	.085	.409	.575	.445	**.616**	**.669**	.506
Sent	.815	.119	.520	.740	.560	.738	.768	.614
Gap	.801	.092	.412	.597	.456	.650	.712	.525
Reranking	.826	.170	.552	.610	.576	.792	.785	.609
DropoutMean	**.762**	.086	**.402**	**.572**	**.441**	**.616**	.670	**.504**
DropoutVar	.960	.370	.775	.894	.817	.881	.907	.858
TemplateDiff	.763	**.084**	.406	.574	.444	.730	-	-
Oracle	.648	.048	.277	.459	.327	.489	.522	.388

(c) RoBERTa-base

確信度	Google-RE	T-REx				ConceptNet	SQuAD	全体
		1-1	N-1	N-M	All			
Token	.818	.191	.540	.635	.562	.618	**.741**	**.599**
Sent	.876	.267	.631	.761	.657	.754	.780	.716
Gap	.827	.197	.545	.632	.565	.657	.782	.610
Reranking	.865	.276	.637	.627	.636	.804	.828	.669
DropoutMean	.815	.201	**.536**	.633	.562	**.615**	.744	**.599**
DropoutVar	.979	.643	.924	.920	.920	.896	.907	.923
TemplateDiff	**.813**	**.189**	.537	**.626**	**.558**	.744	-	-
Oracle	.730	.106	.416	.492	.432	.503	.571	.474

表 3.3: T-REx データセットにおける確信度指標の各種スコア平均の比較．上段は BERT-base における `Token` と `Gap` を，下段は RoBERTa-base における `Token` と `TemplateDiff` を比較している．“X-win” は指標 X が他方よりも性能がよかった事例集合を示す．Δ は 2 つの部分集合間のスコア差．

	BERT-base			
	All	`Token`-win	`Gap`-win	Δ
Accuracy	0.311	0.283	0.413	-0.130
RC-AUC `Token`	0.558	0.577	0.466	0.111
RC-AUC `Gap`	0.566	0.597	0.443	0.154
Answer Cov.	0.285	0.276	0.334	-0.058
Prediction Cov.	0.547	0.579	0.464	0.115
	RoBERTa-base			
	All	`Token`-win	TD-win	Δ
Accuracy	0.242	0.315	0.231	0.085
RC-AUC `Token`	0.643	0.545	0.657	-0.112
RC-AUC TD	0.638	0.546	0.650	-0.103
Answer Cov.	0.237	0.285	0.235	0.050
Prediction Cov.	0.562	0.586	0.541	0.045

表 3.4: 異なる確信度指標に基づく上位 100 件の予測に含まれる予測単語の内訳の比較．モデルは BERT-base，データセットは Google-RE．括弧内の数字が各予測の出現頻度を示す．

Relation	Confidence	Top predictions
`date-of-birth`	`Token`	1979 (47), 1944 (33), 1988 (10), 1990 (8)
	`Reranking`	1979 (44), 1944 (32), 1953 (13), 1970 (3), 1949 (2)
`place-of-birth`	`Token`	Budapest (18), Prague (10), Istanbul (8), Athens (8), Paris (7), Moscow (7), Helsinki (6), Bucharest (6), Tehran (5), Stockholm (4)
	`Reranking`	London (30), Dublin (12), Paris (12), Moscow (5), Madrid (4), Philadelphia (4), Chicago (4), Warsaw (3), Tehran (3), Berlin (2)
`place-of-death`	`Token`	Paris (38), Rome (32), Moscow (6), Madrid (4), infancy (4), office (3), Athens (2), Helsinki (2), Warsaw (2), Amsterdam (2)
	`Reranking`	London (46), Paris (14), Rome (7), office (6), Moscow (4), Munich (3), Amsterdam (3), infancy (2), prison (2), Stockholm (2)

表 3.5: 各スコアを予測に直接用いた場合の予測精度（P@1）．Reranking は予測尤度に基づく予測を前提に計算される指標なので，ここでは除外する．

モデル	予測指標	Google-RE	T-REx	ConceptNet	SQuAD	All
BERT-base	Token	10.3	29.6	**15.8**	14.1	24.3
	Sent	**10.5**	29.6	14.6	14.4	24.1
	Gap	9.7	28.6	15.3	**15.1**	23.5
	DropoutMean	10.3	**29.8**	15.4	14.1	**24.4**
	DropoutVar	0.2	0.1	0.1	0.0	0.1
	TemplateDiff	9.6	29.4	14.2	-	-
BERT-large	Token	11.0	31.0	19.3	17.4	26.1
	Sent	**11.2**	31.5	17.6	15.7	26.1
	Gap	10.4	29.6	18.6	17.4	25.0
	DropoutMean	10.9	**31.7**	**19.6**	**17.7**	**26.7**
	DropoutVar	0.2	0.0	0.0	0.0	0.1
	TemplateDiff	10.6	30.5	17.0	-	-
RoBERTa-base	Token	7.5	23.0	**18.5**	14.7	20.2
	Sent	**8.2**	24.3	17.0	12.2	20.7
	Gap	7.6	22.0	17.4	14.7	19.3
	DropoutMean	8.0	**24.4**	18.3	**15.7**	**21.1**
	DropoutVar	0.1	0.1	0.1	0.0	0.1
	TemplateDiff	7.5	23.2	16.4	-	-

第 4 章

訓練データに基づく確信度指標

3 章では，LAMA probe による事前学習済み言語モデルの知識評価に選択的予測を導入し，言語モデルと確信度指標で構成されるシステムが正しい知識を区別可能な形で出力できるかを評価可能にした．一方，3 章で導入した確信度指標のうち，最も単純な予測尤度を用いる指標 (Token) を明確に改善する指標は確認できなかった．本章では，事前学習済み言語モデルに関する追加の情報を利用することで，確信度指標の改善ができるかを検討する．具体的には，言語モデルの事前学習に用いられた訓練データを用いた確信度指標を導入し，選択的予測のもと LAMA probe での効果を検証する．

2.4 節で述べたように，言語モデル出力の確信度推定では，異なる入力に対するモデルの出力や内部状態の振る舞いに着目した指標が多数提案されてきた．一方で，言語モデルの訓練データを確信度推定に用いることはこれまでに十分な検討がなされていない．これには，代表的な商用の大規模言語モデルへのアクセス手段が API を介したものに限られており，利用できる情報が限定的であるという背景があると考えられる．しかしながら，訓練データの著作権，訓練データに起因するバイアス，プライバシー保護といった観点から，大規模言語モデルの訓練データ透明化の要請が高まっており (Zhang et al. 2024, Wei et al. 2024)，訓練データに対する検索ができる仕組みや (Piktus et al. 2023)，公開データのみから構成された訓練データを用いたモデルの公開が進んでいる (Touvron et al. 2023)．また，訓練データの利用はテキスト生成の質改善において有効性が確認されている．例えばKhandelwal et al. (2020) は，テキスト生成時に現在の文脈と類似する訓練データ中の文脈を参照しながら次の単語の出力確率を補正することで検索の

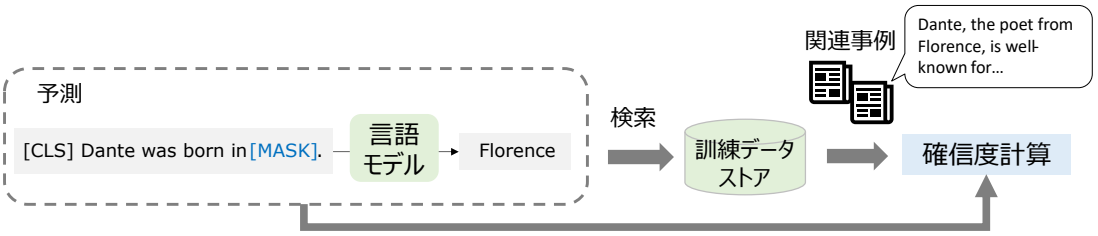


図 4.1: 訓練データを用いた確信度計算の概要図.

質を向上できることを報告している．こうした背景から，言語モデルの訓練データにアクセスできる状況を想定し，訓練データを用いることが確信度推定の性能向上に寄与するかを確かめることが本章の目的である．

4.1 言語モデルとデータストア

図 4.1に訓練データに基づく確信度計算の概要を示す．確信度計算には，言語モデルの訓練データをベクトルまたはテキスト形式で保持するデータストアを用いる．データストアは，入力文とそれに対する言語モデルの出力を参照し，関連する訓練データ中の事例を検索するために用いられる．その後，検索によって得られた関連事例をもとに確信度推定を行う．

4.1.1 言語モデル

訓練データを用いた確信度推定には言語モデルと参照可能な訓練データのセットの組が必要である．ここでは中規模な言語モデルを自前で訓練し，訓練データからデータストアを構築した．訓練する言語モデルとしては BERT-large (Devlin et al. 2019) を採用した．訓練データとしては，2020 年 1 月版の英語 Wikipedia 全文を用い， MLPerf Training Benchmark (Mattson et al. 2020) の実装に基づき事前学習を行った.* ただし，実験環境の違いや評価タスクへの調整のため，学習時の設定にいくつかの変更を加えている．まず，評価タスクとして用いる LAMA probe はアルファベットの大文字と小文字を区別する仕様であるため，使用する言語モデルも大文字小文字の区別がある語彙（cased）に基づき訓練した．これに伴い，MLPerf の提供する大文字小文字の区別がない（uncased）モデルチェックポイン

*https://github.com/mlcommons/training/tree/master/language_model/tensorflow/bert

表 4.1: MLPerf と本研究で構築した BERT モデルの設定比較.

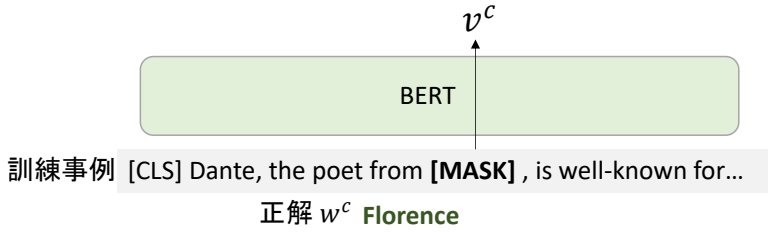
	MLPerf	Ours
語彙	uncased	cased
GPU	V100×8	P100×4
バッチサイズ	24	12
初期チェックポイント	あり	なし

トを使用せず，ランダムに初期化したパラメータから学習を開始した．訓練に用いた GPU は NVIDIA P100 を 4 基，バッチサイズは 12 とした．表 4.1に MLPerf の実装と本研究での実装との変更点をまとめる．

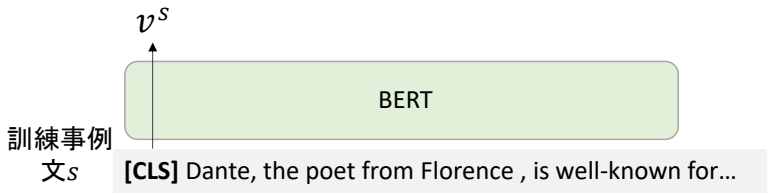
訓練データの前処理方法は Devlin et al. (2019) に準じている．具体的な手順は次のとおりである．各入力文は，次文予測タスクのため 50%の確率でコーパス中の隣り合う文，50%の確率でランダムな別の文と結合され，一件の入力事例を構成する．各入力事例に対し，まず WordPiece (Wu et al. 2016) を用いてテキストのトークン分割を行う．次に，これらのテキストのランダムな位置にマスクをかけ，マスク付きの入力を作成する．具体的には，まず各文に対しトークン全体の 15%をマスク対象としてランダムに選択し，これらをマスクトークン `[MASK]`，ランダムなトークン，元のトークンのいずれかにそれぞれ 80%, 10%, 10% の割合で置き換える．以上の処理をコーパス全体に対し 10 回繰り返し，延べ 1600 万事例からなる訓練データを作成した．予測対象箇所は全部で 61 億となった．モデルはマスク予測と次文予測を用いて計 770 万ステップの学習を行った．学習済みモデルの開発データにおけるマスク単語予測および次文予測の精度はそれぞれ 0.691, 0.986 であった.

4.1.2 データストア

訓練データは言語モデルへの入力およびモデルの生成内容に応じて，関連する文脈情報を取得することに用いる．保存方式による検索の質と確信度推定の影響の違いを確認するため，異なるレベルで情報を保存した複数のデータストアを構築した.



(a) トークンレベル文脈表現.



(b) 文レベル分散表現.

図 4.2: 文表現の概要図.

トークンレベル文脈表現

4.1.1節で述べたように，言語モデルの訓練時には，訓練データをトークン分割し，予測対象の位置をマスクした前処理済みの入力が用いられる．トークンレベル文脈表現としては，この前処理済みの入力の予測対象位置の文脈を学習済みモデルで改めてエンコードした文脈ベクトル v_c を用いる．ここでの文脈ベクトルは，予測位置に対応する BERT モデルの最終層の隠れ状態とする（図 4.2a）.

データストアは，文脈ベクトル v^c をキー，該当文脈の正解トークン w^c を値とする組 (v^c, w^c) を保存する．文脈ベクトルの保存と検索には，密ベクトル向けのベクトル検索ライブラリ FAISS (Johnson et al. 2019) を用いた．訓練時に使用した文脈情報は 61 億箇所あり，これら全てを生ベクトルの状態で保存すると膨大なサイズになる．リソース消費を抑えるため，直積量子化 (product quantization, PQ) に基づくベクトル量子化を施した．また，検索効率向上のため転置インデックス (IVF) を採用し検索空間を削減した．これにより，近傍ベクトルの検索結果は近似値となり，厳密解と一致しない場合が生じる．ベクトル量子化や近似検索のためのハイパーパラメータは，訓練データの一部を用いて十分な再現率が得られるよう調整した．近傍検索を行う際には，構築したインデックスに基づき上位 100 件の近傍を取得した後，検索結果に対し文脈ベクトルを再計算し，厳密な L2

距離を求めてリランキングを行った．

文レベル分散表現

文レベル分散表現は，トークンレベルの文脈ではなく生成文と文単位で類似する訓練事例を検索するために用いる．マスク処理を行う前の訓練事例を文単位で分割し，各文を学習済み言語モデルでエンコードする．最終層の文頭 [CLS] トークンに対応する隠れ状態 v^s を文表現とみなし，文 s との組 (v^s, s) をデータストアに保存する（図 4.2b）．なお，実装上は文の元テキストと前処理済みの特徴量，メタデータを共に保存している．[CLS] トークンの隠れ状態は，事前学習において次文予測タスクにより訓練されている．ベクトルのインデックス化や検索の方式はトークンレベル分散表現と同様である．

4.1.3 テキスト一致検索

言語モデルの学習した事実に関する知識を確認する上では，特定のエンティティに関して記載された文を参照することが有用な場合がある．そこで，分散表現に基づく類似文脈の検索に加え，単純なテキスト一致に基づく検索も実施する．

4.2 確信度指標

4.2.1 トークンレベル文脈表現に基づく確信度

検索

トークンレベル文脈表現の検索では，入力文 c_t のマスク箇所を言語モデルでエンコードした文脈ベクトル v^{q} をクエリとし，4.1.2節の方法で構築した文脈ベクトルデータストアに対して k 近傍検索を行い近傍事例 $\mathcal{N} = \{(v_1^{\mathrm{c}}, w_1^{\mathrm{c}}), \ldots, (v_k^{\mathrm{c}}, w_k^{\mathrm{c}})\}$ を得る．以下の実験では $k = 100$ を用いた．このように事前学習済みモデルの文脈表現を用いて近傍事例を検索する方法は，近傍事例による生成補助において有効性が示されている (Khandelwal et al. 2020)．確信度推定においても，学習で獲得した文脈表現に基づき，入力事例に関連する訓練データを意味的類似度を考慮して検索することで確信度の改善に寄与することを期待し，同様の方法を採用した．

近傍事例に基づく尤度補正 (kNN-LM)

言語モデルによる文生成において，現在の文脈と類似する訓練データ中の文脈情報を活用することで生成の質を向上できることが報告されている (Khandelwal et al. 2020)．この方法に基づき，言語モデルの予測尤度を訓練データ中の近傍の文脈ベクトルにより補正し，確信度として用いる．

まず，クエリベクトル v^{q} に対し，データストアから上位 k 個の近傍事例を検索し，検索結果に基づき補正用の分布 $p_{\mathrm{kNN}}(w|c_t)$ を計算する．補正分布における単語 w の確率は，w を正解とする近傍事例の文脈ベクトル v^{c} とクエリベクトル v^{q} の L2 距離 $d(v^{\mathrm{q}}, v^{\mathrm{c}})$ に基づく:

$$p_{\mathrm{kNN}}(w|c_t) = \left(\sum_{(v^{\mathrm{c}}, w^{\mathrm{c}}) \in \mathcal{N}} \mathbf{1}_{w=w_{\mathrm{c}}} \exp(-d(v^{\mathrm{q}}, v^{\mathrm{c}})^2/\tau) \right) / Z, \tag{4.1}$$

$$Z = \sum_{w'} \sum_{(v^{\mathrm{c}}, w^{\mathrm{c}}) \in \mathcal{N}} \mathbf{1}_{w'=w_{\mathrm{c}}} \exp(-d(v^{\mathrm{q}}, v^{\mathrm{c}})^2/\tau). \tag{4.2}$$

ただし，τ はハイパーパラメータである．これとモデルの元の予測分布との重み和をとり計算された補正分布に基づく尤度を kNN-LM 確信度とする:

$$\phi_{\text{kNN-LM}}(w, c_t) = \log(\lambda p_{\mathrm{kNN}}(w|c_t) + (1-\lambda) p_{\mathrm{LM}}(w|c_t)). \tag{4.3}$$

ただし，λ はハイパーパラメータである．

4.2.2 文レベル分散表現に基づく確信度

検索

文レベル分散表現の検索では，入力文 c_t のマスク箇所を予測単語 $\hat{w}$ で埋め完成させた文をエンコードし，`[CLS]` トークンに対応する隠れ状態 v^{q} をクエリベクトルとする．これを用いて，4.1.2節で構築した文レベル分散表現のデータストアに対して k 近傍検索を行い近傍事例 $\mathcal{N} = \{(v^{\mathrm{s}}_1, s_1), \ldots, (v^{\mathrm{s}}_k, s_k)\}$ を得る．以下の実験では $k = 10$ を用いた．

文脈を付与して再予測 (kNN-sent-context)

大規模コーパス中から入力文と類似する文を検索し，生成時の文脈に追加することで生成の質を向上できることが知られている (Lewis et al. 2020)．これと同様に，訓練データ中の類似文を文脈情報として利用し再度予測を行うことで，出力の確信度予測の改善を試みる．具体的には，検索によって得られた訓練データ中の類似文 s_i を元の入力文 c_t の先頭に文脈情報として付与した文脈つき入力 $s_i \oplus c_t$ を作成し，予測単語 w の出力確率 $P_{\mathrm{LM}}(w|s_i \oplus c_t)$ を得る．これを k 個の類似文全てに対して行い，得られた対数尤度の最大値を確信度として用いる:

$$\phi_{\text{kNN-CTX}}(w, c_t) = \max_{(v_i^{\mathrm{s}}, s_i) \in \mathcal{N}} (\log P_{\mathrm{LM}}(w|s_i \oplus c_t)) \tag{4.4}$$

なお，k 種の予測尤度の平均値をとる方法も検討したが，予備実験において性能に大きな差はみられなかった．

近傍事例中のエンティティ頻度 (kNN-sent-entity)

予測対象の文と意味的に類似する訓練事例中に，問われている知識について直接述べている文がどれだけ多いかを頻度に基づき指標化する．(subject, relation, object) の三つ組からなる関係知識に基づく入力文 c_t について， subject にあたるエンティティを e^{s} とする．検索結果の k 文のうち， e^{s} と object に相当するモデルによる予測単語 $\hat{w}$ の両方が含まれる文の件数 N^{sp} をカウントし，$\phi_{\text{kNN-Ent}}(w, c_t) = N^{\mathrm{sp}}/k$ を確信度とする．

4.2.3 テキスト検索に基づく確信度

テキスト一致件数 (CorpusSearch-count, CorpusSearch-bin)

テキスト検索に基づく確信度では，訓練データ中に予測内容と関連する事例が一定数存在するかどうかをエンティティに基づき判定し，確信度推定に用いる．(subject, relation, object) の三つ組からなる関係知識に基づく入力文 c_t について， subject にあたるエンティティを e^{s} と object に相当するモデルによる予測単語 $\hat{w}$ の両方が含まれている訓練データ中の文を検索し，該当する訓練データ中の文の件数を確信度として用いる: $\phi_{\text{CorpusSearch-count}}(w, c_t) = N(e^{\mathrm{s}}, \hat{w})$．また，訓練データ中の該当文

表 4.2: 評価に用いたモデルの LAMA データセットにおける（全ての予測を用いた場合の）予測精度と，予測尤度に基づく確信度 ϕ_{Token} に基づく RC-AUC ．比較のため，右列に Google BERT モデルによる同条件での評価を示している．

	Ours		（参考）Google BERT-large	
データセット	精度 (↑)	RC-AUC (↓)	精度 (↑)	RC-AUC (↓)
Google-RE	.084	.814	.110	.763
T-REx	.246	.546	.310	.445
ConceptNet	.115	.789	.193	.616
SQuAD	.105	.814	.174	.669

の有無のみに基づく二値の確信度も検討する: $\phi_{\text{CorpusSearch-bin}}(w, c_t) = \mathbf{1}_{N(e^{\text{s}},\hat{w})>0}$. なお，訓練データ内に 100 件以上の関連事例が見つかった場合は十分に根拠ありとみなし，計算効率の観点から検索事例は 100 件を上限とした．

文脈を付与して再予測 (CorpusSearch-context)

4.2.2 節と同様に，テキスト一致による検索結果を生成時の文脈として付与し，再予測時の対数尤度を用いる．手続きは 4.2.2 節に準ずるが，テキスト一致検索では検索結果が 0 件の場合があることを考慮し，文脈を付与しない元の文の対数尤度を含めた予測の中から最大値をとる．

4.3 実験

検証においては，中規模の言語モデルとして BERT-large (Devlin et al. 2019) を選択した．BERT モデルは 3 章の評価にも用いた LAMA probe において高い性能を示している．本実験では訓練データ全てにアクセスできることを前提とするため，英語 Wikipedia を訓練データとして自前で事前学習を行ったものを用意した．

4.4 実験結果

4.4.1 言語モデル性能

表 4.2に，評価対象の BERT モデルの LAMA データセットでの基礎性能を示す．ここでの RC-AUC の計算には，予測尤度に基づく確信度 ϕ_{Token} を用いた．比較

表 4.3: 評価データの内訳.

データセット	#dev	#test
Google-RE	554	4973
T-REx	4054	29963
ConceptNet	2102	9356
SQuAD	31	274

対象として，Google BERT-large モデル[†] による同条件での評価結果を示している．本研究では言語モデルの確信度評価としての訓練データの有効性を確認することを主目的とし，検証用モデルが Google-BERT と同等の性能を達成することは必ずしも重視していない．評価用モデルと Google-BERT とで異なる点としては， Google-BERT では Wikipedia に加えて BookCorpus (Zhu et al. 2015) を訓練データに用いている点が挙げられる． BookCorpus は様々なジャンルの書籍からなるテキストで構成されており，Wikipedia でカバーされない多様な表現を含む．また，4.1.1節で述べたように，モデルの訓練においては MLPerf Training Benchmark の学習設定をタスクと実験環境に合わせて改変しており，この点も性能差に影響した可能性がある．

4.4.2 選択的予測に基づく確信度指標の評価

評価データ

表 4.3に示す通り，LAMA データセット中の Google-RE, T-REx, ConceptNet, SQuAD の 4 つのサブセットをそれぞれ開発データと評価データに分割した．開発データは kNN-LM のハイパーパラメータ探索および複数指標の組み合わせと重み探索に用いた．

ベースライン

訓練データを用いない確信度指標として，3 章で導入した指標と比較を行った．Token は出力の対数尤度を直接用いるもの，Sent はマスク箇所を埋めた文レベルの疑似尤度を用いるもの，DropoutMean は推論時にモデルに dropout を適用し

[†]https://github.com/google-research/bert

表 4.4: LAMA データセットの評価結果 (RC-AUC/E-AURC，低いほど良い)．SQuAD データセットは subject ラベルが未定義なため，これを必要とする指標は除外．複数指標組み合わせの結果は，(A) が (A) 群のみからの組み合わせ，(A)+(B) が (A) 群ならびに (B) 群全ての指標からの組み合わせを表す．それぞれの指標群の重み和のうち開発データで最良の組み合わせの結果を示している．

	RC-AUC (E-AURC)			(↓)
確信度	Google-RE	T-REx	ConceptNet	SQuAD
(A) 訓練データ不使用				
Token	0.8132 (0.1049)	0.5796 (0.1497)	0.7877 (0.1341)	0.8101 (0.1618)
Sent	0.8661 (0.1578)	0.6556 (0.2257)	0.8616 (0.2080)	0.8272 (0.1789)
DropoutMean	0.8152 (0.1069)	0.5790 (0.1491)	0.7917 (0.1381)	0.8228 (0.1745)
TemplateDiff	0.8156 (0.1073)	0.5526 (0.1227)	0.8590 (0.2054)	-
(B) 訓練データ使用				
kNN-LM	0.8130 (0.1047)	0.5796 (0.1497)	0.7954 (0.1418)	0.8101 (0.1618)
kNN-sent-context	0.8433 (0.1350)	0.6596 (0.2297)	0.8528 (0.1992)	0.8324 (0.1841)
kNN-sent-entity	0.9444 (0.2361)	0.6953 (0.2294)	0.9355 (0.2819)	-
CorpusSearch-count	0.8374 (0.1291)	0.6445 (0.2146)	0.9180 (0.2644)	-
CorpusSearch-bin	0.8326 (0.1243)	0.6317 (0.2018)	0.9469 (0.2933)	-
CorpusSearch-context	0.7580 (0.0497)	0.6209 (0.1910)	0.8141 (0.1605)	-
複数指標組み合わせ				
(A)	0.8117 (0.1034)	0.5796 (0.1497)	0.7954 (0.1418)	0.8101 (0.1618)
(A)+(B)	0.7569 (0.0486)	0.5299 (0.1000)	0.7838 (0.1355)	0.8101 (0.1618)

て予測の統計情報を用いるもの，TemplateDiff は予測における subject 情報の有無による予測確率の差分を用いるものである．

単体評価

表4.4に，3 章で導入した選択的予測の設定に基づく評価結果を示す．なお，SQuAD については subject ラベルが定義されておらずエンティティを用いる確信度指標を適用できないため，それらは評価から除外した．訓練データを使用しない確信度指標の中では Token または TemplateDiff の性能が最も良く，どの指標も対数尤度を直接用いる Token 指標を大きく改善することはなかった．これは他のモデルで評価した既存研究の傾向とも適合する．訓練データを使用した確信度指標については，kNN-LM 指標が関係知識を問うデータセットにおいてベースラインをわずかに改善した．一部を除き，文レベル分散表現やテキスト一致に基づく指標は Token 確信度よりも悪い結果となったが，Google-RE に関しては

CorpusSearch-context が Token 指標を改善した．

確信度の組み合わせ効果

次に，これらの確信度指標を組み合わせて用いた場合の効果を検証する．複数の確信度指標の組み合わせとしては，確信度指標の重み付き線形和をとる．各データセットについて，開発データを用いて最良の指標組み合わせと重みを探索し，評価データセットで評価した．指標組み合わせと重みの探索を現実的な試行回数に収めるため，次のような段階的手続きによる探索を行った．まず，ベースラインで最良である Token 確信度とそれ以外の確信度を 1 つずつ組み合わせ， Token 確信度の性能から改善のあった指標に絞り込む．このとき，それぞれについて 0.1, 1.0, 10.0 の 3 種類から最良の重みを決定した．次に，絞り込み後の指標群について全ての組み合わせを検証し，性能が最良となるものを探索した．

結果を表 4.4下部に示す．テキスト検索に基づく指標を適用できない SQuAD を除き，いずれのデータセットにおいても，訓練データを使用する指標と使用しない指標を組み合わせて用いた場合に最良の性能が得られた．特に Google-RE と T-REx については，尤度に基づく指標を含む全指標に対し，単独使用の場合の性能を大幅に改善した．訓練データを使用しない指標のみを組み合わせた場合には同様の性能改善は得られなかった．

表 4.5に，各データセットで用いられた指標の組み合わせとアブレーション分析の結果を示す．Google-RE と T-REx においては，コーパス検索に基づく指標が特に性能に寄与していることが確認できる．ConceptNet についてはいずれの指標も一定の寄与があるものの，性能の改善幅は関係知識を問うデータセットと比較して小さかった．

表 4.6は，T-REx データセット上でコーパス検索に基づく指標 (CorpusSearch-bin) とモデル予測の正誤のクロス集計を行った結果である．CorpusSearch-bin はバイナリ指標であり，予測に関係する事例が訓練データ中に見つかれば 1，見つからなければ 0 の値をとる．表より，正解事例の 94.9%において，訓練データ中に関連文が存在している．すなわち，コーパス検索に基づく指標は単独の確信度指標としては精度が粗いが，訓練データ中に関連文が存在しない事例をフィルタリングすることによる効果が高いため，他の指標と組み合わせることで高い効果を得られたと考えられる．

表 4.5: 複数指標組み合わせ評価のアブレーション分析．評価値は RC-AUC．最上段は各データセットの開発データで最良の組み合わせで，数値は重みを示す．２行目以降は，他の指標の重みは変更しないまま，当該指標のみを除いた場合の結果．SQuAD は Token 指標を単独で用いた場合に最良であったため除外．

(a) Google-RE

指標	RC-AUC	差分
(A) × 1.0 + (B) × 10.0	0.7569	
− (A) kNN-sent-context	0.7580	+0.0011
− (B) CorpusSearch-context	0.8433	+0.0864

(b) T-REx

指標	RC-AUC	差分
(A) × 1.0 + (B) × 0.1 + (C) × 0.1 + (D) × 10.0 + (E) × 1.0	0.5299	
− (A) Token	0.5560	+0.0261
− (B) kNN-sent-context	0.5302	+0.0003
− (C) kNN-sent-entity	0.5305	+0.0006
− (D) CorpusSearch-bin	0.5623	+0.0324
− (E) CorpusSearch-context	0.5286	−0.0013

(c) ConceptNet

指標	RC-AUC	差分
(A) × 1.0 + (B) × 1.0 + (C) × 10.0 + (D) × 0.1 + (E) × 10.0	0.7838	
− (A) Sent	0.7876	+0.0038
− (B) TemplateDiff	0.7831	−0.0007
− (C) kNN-LM	0.8060	+0.0222
− (D) CorpusSearch-count	0.7863	+0.0025
− (E) CorpusSearch-context	0.7870	+0.0032

4.5 分析

4.5.1 事例分析

4.4節では，訓練データを用いない確信度と訓練データに基づく確信度，特にコーパス検索に基づくものとの組み合わせが有効に働くことが示唆された．そこで，ベースラインである尤度に基づく確信度 (Token) とコーパス検索に基づく文脈付

表 4.6: CorpusSearch-bin 指標の値と予測正誤のクロス集計表.

		CorpusSearch-bin 0	1
予測	不正解	15136	10524
	正解	424	7933

与 (CorpusSearch-context) の振る舞いを，効果の大きかった Google-RE と効果が比較的小さかった ConceptNet の事例を用いて比較する．表 4.7に示す Google-RE の事例では，予測が誤りである例について，コーパス検索では関連事例が見つからないため相対的に確信度を下げることができている (1,2)．逆に 3,4 では，予測が正しい事例についてコーパス検索に基づき高い確信度を与えることができている．一方，コーパス検索の効果が低かった ConceptNet においては，検索により関連事例が見つかっているにも関わらず正解事例の相対的な確信度を落としてしまう例がみられた (3,4)．ConceptNet の事例は一般名詞や形容詞といった単語で表される概念間の意味的関係を問うもので，固有名詞と比較して訓練データ中での出現頻度は高く，関連事例が見つかりやすい傾向がある．しかしながら，検索された事例は問われている概念間の関係を直接的に述べていない場合が多く，モデルの確信度の強化に貢献しにくかったと推定される．

4.5.2 データストア検索と出力真偽の関係

実験に用いたデータストアがモデル出力の真偽判定に寄与しうる検索結果を返すことができているかを検証するため，以下の分析を行った．各検索方法による検索結果について，各予測に対する検索結果として用いられた文書のうち，入力文の subject にあたるエンティティ e^s と予測単語 $\hat{w}$ の両方を含む事例が存在する場合に予測を正解，存在しない場合に不正解とみなす単純な分類器を考える．この分類器に基づきモデル予測の真偽判定を行ったときの予測精度が高ければ，データストアがモデル出力の真偽判定の根拠となりうる情報を適切に提示できている可能性が高いと考えられる．表 4.8に結果を示す．いずれのデータセットにおいても，トークン文脈ベクトルおよび文ベクトル検索結果は recall が低く，真偽判定の根拠となりうる情報が検索結果から漏れてしまっているケースが多いと考えられる．このことから，ベクトル検索手法を改良し，訓練データに存在する根拠

表 4.7: 予測事例に対する 2 つの確信度 (Token, CorpusSearch-context) による順位付けの比較例．Δrank は，同じ関係ラベルをもつデータセット（分母は総数）内での Token と CorpusSearch-context に基づく順位の差． 下向き矢印↓は Token と比べて CorpusSearch-context による順位が下がっていることを示す．それぞれ，コーパス検索で該当事例がない不正解事例の順位を下げられた例 (1,2)，コーパス検索により正解事例の順位を上げられた例 (3,4)，コーパス検索で事例があるにもかかわらず正解事例の順位を下げてしまった例 (5,6)．

No.	データセット	入力文	モデル予測	正解 (✓)	Δrank	検索事例
1	Google-RE	Tzipora Laskov was born in [MASK] .	Moscow	×	302↓/2937	なし
2	Google-RE	Ivan Triesault was born in [MASK] .	Paris	×	801↓/2937	なし
3	Google-RE	Shiva Boloorian was born in [MASK] .	Tehran	✓	2767↑/2937	Shiva Boloorian (, born 5 October 1973 Tehran) is an Iranian Playwright, Actress and both Film and Theatre director, as well as a Television presenter.
4	Google-RE	Owen McAuley was born in [MASK] .	Belfast	✓	2757↑/2937	Owen McAuley (born 5 October 1973 in Belfast, Northern Ireland) is a British auto racing driver.
5	ConceptNet	Bathing requires [MASK] and soap.	water	✓	57↓/532	Near the Lycus River, there is evidence that water channels had been cut out of the rock with a complex of pipes and sluice gates to divert water for bathing and for agricultural and industrial purposes.
6	ConceptNet	Talking requires opening your [MASK] and breathing out over a tongue shaped in different ways for different sounds.	mouth	✓	89↓/532	At the end of the sketch, he demands of the smart-mouthed talking parrot on his shoulder, "Do you want to be an ex-parrot?"

表 4.8: 各検索方式で確信度推定に用いた検索結果をもとに，予測対象エンティティの出現有無に基づき真偽判定をした場合の精度評価．SQuAD データセットは subject ラベルが未定義のため除外．

(a) トークンレベル文脈表現

	Recall	Precision	F1
Google-RE	0.002	1.000	0.004
T-REx	0.033	0.215	0.058
ConceptNet	0.236	0.136	0.173

(b) 文レベル分散表現

	Recall	Precision	F1
Google-RE	0.002	0.333	0.004
T-REx	0.090	0.601	0.156
ConceptNet	0.106	0.109	0.107

(c) テキスト一致検索

	Recall	Precision	F1
Google-RE	0.664	0.393	0.494
T-REx	0.949	0.430	0.592
ConceptNet	0.971	0.120	0.213

情報をより高精度に検索することで，ベクトル表現に基づく指標の性能が改善できる可能性がある．テキスト一致検索では，いずれのデータセットでも recall が高く， precision は文ベクトル検索と同程度であった．このことは，モデル出力が正しいケースでは多くの場合に訓練データ中に根拠情報が含まれていることを示唆しており，訓練データがモデル出力の確信度推定に寄与することを裏付けている．

各検索方法による検索結果の例を表 4.9および表 4.10に示す．表中の予測事例は全て同一の関係テンプレート “<subject> is the capital of [MASK] .” に関するものである．トークンレベル文脈表現による検索結果は，周辺文脈のトピックには関連があるものの，予測事例に対しての言及が無かったり無関係の事柄について述べている事例が目立つ．一方，文レベル分散表現はほとんどの場合に問われている関係（ここではある都市を首都とする国や地域）について直接言及してい

る文が検索されており，特に予測が正解の場合 (表 4.9 (1,2)) には出力内容を支持する文を検索することができている．予測が不正解の場合 (表 4.10 (3,4)) にも，文中に出現するエンティティを含む表現を優先的に検索することができているが，同時に別の都市の首都に関する類似の文も多く検索される傾向にあった．このように，類似の文脈だが注目するエンティティとは無関係の事実について述べた文が近傍に多く存在することで，確信度推定に悪影響を与えた可能性がある．

テキスト一致検索はエンティティの表層が一致している文を意味を考慮せずに取得する．したがって，訓練データ中に当該事例について直接述べている文があれば漏れなく検索することができる一方で，無関係の文脈に 2 つのエンティティが同時に出現している場合にノイズとなる検索結果が多くなる傾向にある (表 4.10 (3))．訓練データ中での出現頻度が高いエンティティほどノイズとなりうる事例は多くなる．

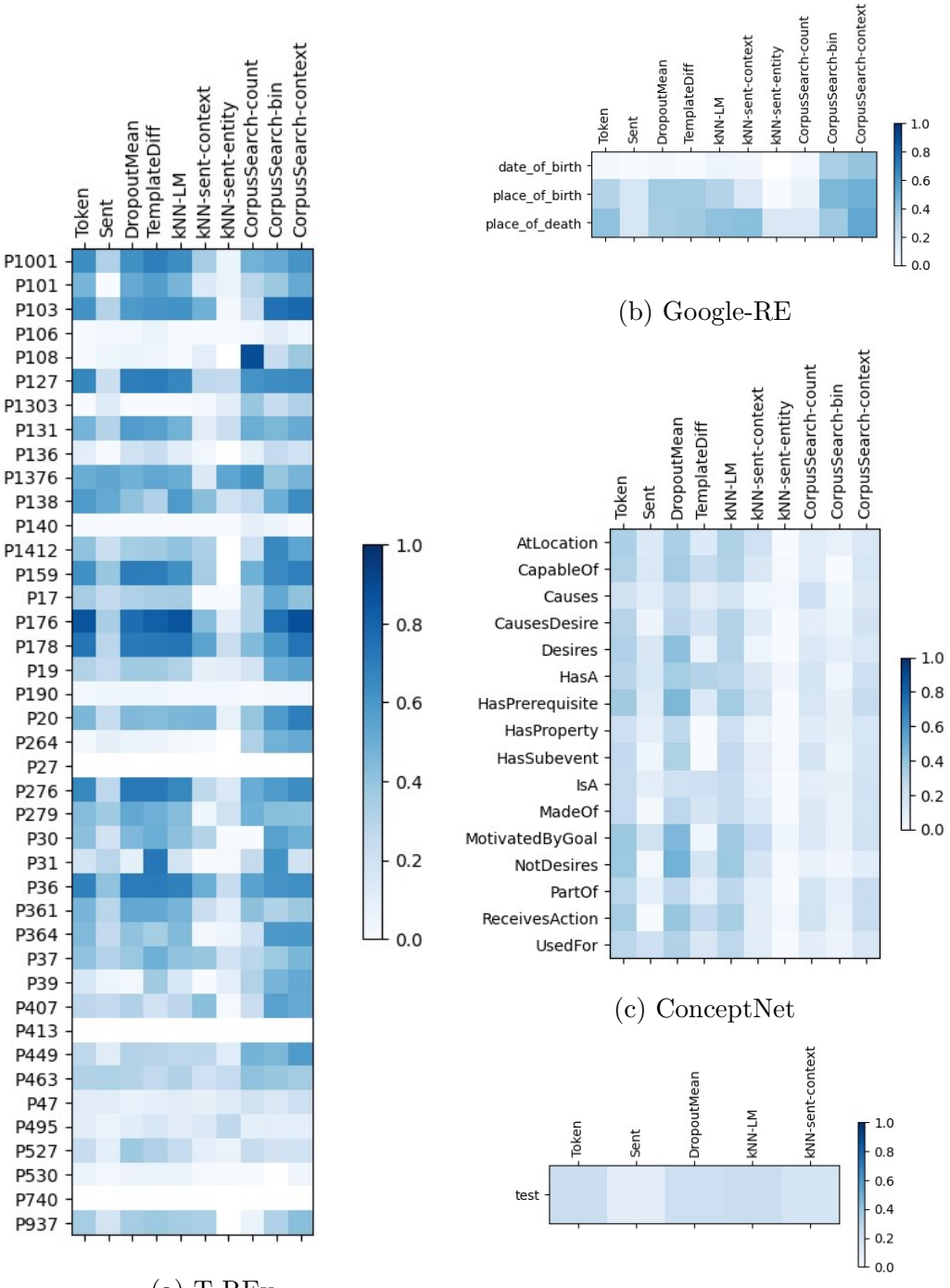
以上より，関連事例検索の主な課題は，問われている事実に直接関係する文のみをより高精度に検索し，無関係な事例がノイズになることを防ぐことといえる．文レベル分散表現は文脈が類似する文を検索するのに適しているが，無関係なエンティティについて述べた文も同時に検索されることは知識を問うタスクでの確信度推定には適さない．これを解決するには，訓練を伴う文ベクトル検索手法 (Khattab and Zaharia 2020, Wang et al. 2024) を導入し，出力内容を支持する文のみが検索されるようにすることが有効であると考えられる．また，確信度の計算時に出力内容との関連度合いに応じて検索結果に重みづけをするなど，検索結果のノイズに頑健な指標への改良も必要である．

4.5.3　データセット・関係タイプと確信度指標

確信度指標の寄与度がデータセットや関係タイプにどの程度依存するのかを調べるため，データセット・関係タイプ毎に確信度指標と正誤の相関を計算した結果を図 4.3に示す．三つ組の関係知識に関するデータセットである Google-RE と T-REx では，エンティティ頻度に基づく指標 (kNN-sent-entity) を除くほぼ全ての指標について，正誤との間に比較的高い相関があり，特に予測尤度 (Token) およびコーパス検索に基づく指標 (CorpusSearch-bin, CorpusSearch-context) が高い相関を示している．一方，語彙間の関係知識を問う ConceptNet においては，訓練データに基づく指標と正誤の相関が一貫して小さくなっている．表 4.8に示

表 4.9: 予測が正解である事例に対する各検索方式による検索結果の例（データセット: T-REx:P1376）．テキストが長いものは “...” で省略し一部を表示している．また，トークンレベル文脈表現の検索結果はマスクを元の単語で復元した状態で表示し，文脈表現に対応するマスク位置を下線で示している．

1	Input: Paris is the capital of [MASK] . 正解: France, 予測: France ✓	
	トークンレベル文脈表現	... The " Universal Dictionary of Ancient and Modern France and New France " by Marin Saugrain stated in 1726 that : " BAALON in Champagne, Diocese of Reims, Parliament of Paris, Intendance of Chalons, Election of Rhetel, has 419 inhabitants. ...
		... " BAALON in Champagne, Diocese of Reims, Parliament of Paris, Intendance of Chalons, Election of Rhetel, has 419 inhabitants. The Curate is worth nine hundred livres. " ...
	文レベル分散表現	Paris is the capital of France.
		Warsaw is the capital of Poland.
	テキスト一致検索	Preliminary peace articles were signed in Paris on 30 November 1782, while preliminaries between Britain, Spain, France, and the Netherlands continued until September 1783.
		Nobel was accused of high treason against France for selling Ballistite to Italy, so he moved from Paris to Sanremo, Italy in 1891.
2	Input: Edmonton is the capital of [MASK] . 正解: Alberta, 予測: Alberta ✓	
	トークンレベル文脈表現	... Thus after 2014 the present Muğla central district will be named " Menteşe " and the name Muğla will be reserved for the metropolitan municipality . (Menteşe was the name of a 14th - century beylik in and around Muğla Province.) ...
		... Its name comes from the largest throne in the community of the Amazigh, before becoming the name of the District . Oued Rechache District is characterized by the steppe climate which prevails in the Aures highlands. ...
	文レベル分散表現	Edmonton is the capital city of the Canadian province Alberta.
		Toronto is the provincial capital of Ontario.
	テキスト一致検索	In 1915, Sidney Ells of the Federal Mines Branch experimented with separation techniques and used the product to pave 600 feet of road in Edmonton, Alberta.
		Alberta’s capital, Edmonton, is near the geographic centre of the province and is the primary supply and service hub for Canada’s crude oil, the Athabasca oil sands and other northern resource industries.



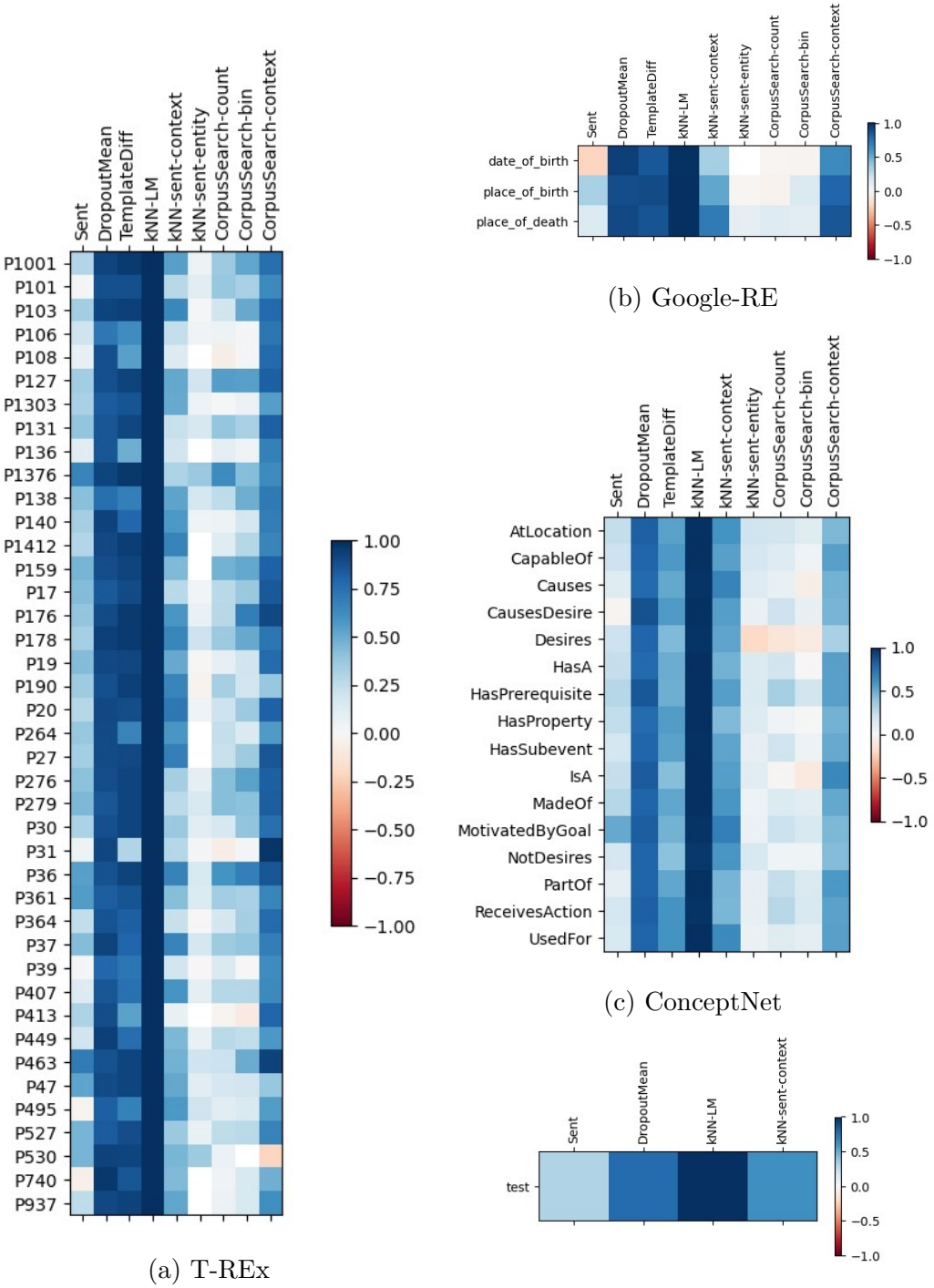
(a) T-REx

(b) Google-RE

(c) ConceptNet

(d) SQuAD

図 4.3: データセット・関係タイプ毎の予測正誤と確信度の相関.



(a) T-REx

(b) Google-RE

(c) ConceptNet

(d) SQuAD

図 4.4: データセット・関係タイプ毎の Token 確信度と他指標の相関.

表 4.10: 予測が不正解である事例に対する各検索方式による検索結果の例（データセット: T-REx:P1376）．テキストが長いものは “...” で省略し一部を表示している．また，トークンレベル文脈表現の検索結果はマスクを元の単語で復元した状態で表示し，文脈表現に対応するマスク位置を下線で示している．

3	Input: Amsterdam is the capital of [MASK] . 正解: Netherlands, 予測: Belgium ×	
	トークンレベル文脈表現	... They successfully blocked the Indian reinforcements and subsequently captured Dras and Kargil as well, cutting off the Indian communications to Leh in Ladakh. [SEP] in the United States (by state then city) [SEP]
		[CLS] Bruce Lee ’ s character in the movie practically is Shang - Chi, and his uneasy alliance with wisecracking pal Roper John Saxon echoes, in many ways, Shang - Chi ’ s give - and - take with agent Clive Reston. [SEP] " (by state then city) " [SEP]
	文レベル分散表現	Amsterdam is the capital of the Netherlands.
		Brussels is the capital city of Belgium.
	テキスト一致検索	In 1920, Amsterdam assisted in hosting some of the sailing events for the Summer Olympics held in neighbouring Antwerp, Belgium by hosting events at Buiten Y.
		Amsterdam distributed grain to the major cities of Belgium, Northern France and England.
4	Input: Porto-Novo is the capital of [MASK] . 正解: Benin, 予測: Portugal ×	
	トークンレベル文脈表現	[CLS]. village jani gabol [SEP]. village attur abro asa [SEP]
		... In Arabic, Am Timan means " mother of twins, " although the reason for the name was back then there a female of Buffalo gave a twins birth in that particular place so the name came from there / As the capital of the prefecture, it has the area’of many towns and villages around it including Zakuma national park. ...
	文レベル分散表現	Lisbon is the capital city of Portugal.
		Brasília is the capital of Brazil.
	テキスト一致検索	（該当なし）

すように，ConceptNet は訓練データ中のエンティティ出現有無による分類精度が低いことを踏まえると，一般名詞を主とする語彙の共起はノイズを多く含むため，訓練データを利用した真偽判定との相性が悪いと考えられる．

図 4.4は Token 確信度と他の確信度指標のスコア相関をデータセット・関係タイプ毎に計算したものである．検索結果を文脈に付与して尤度を更新する CorpusSearch-context を除けば，訓練データを用いる指標は訓練データを用いないものと比較して Token 確信度との相関が一貫して低いことから，両者の組み合わせにより高い効果が得られたと考えられる．

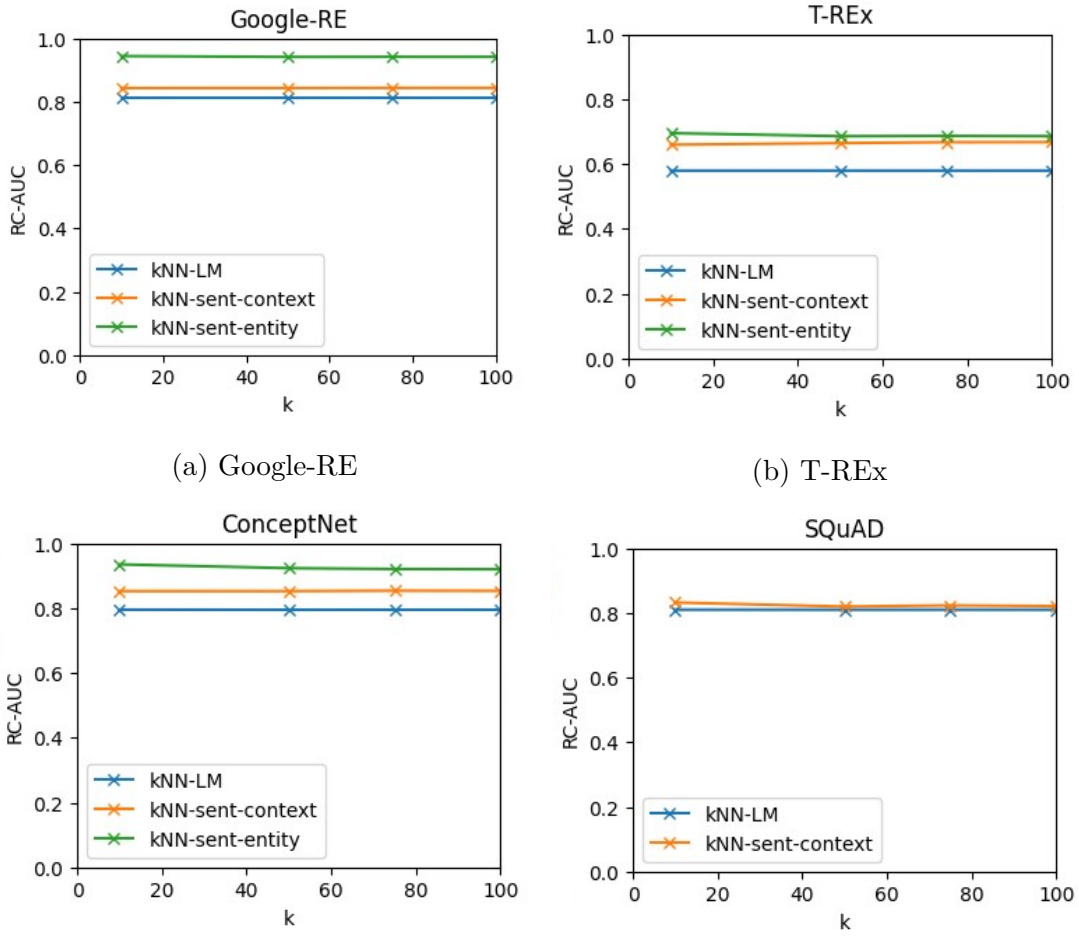
4.5.4　近傍事例数の確信度への影響

4.2.1節および 4.2.2節の検索事例を用いた確信度指標においては，検索する近傍事例数をそれぞれ $k = 100$，$k = 10$ とした．近傍事例数 k を変えたときに確信度推定に与える影響を調べるため，異なる k を用いて単体性能の比較を行った結果を図 4.5に示す．$k = 10, 50, 75, 100$ における確信度指標の性能は，いずれのデータセットにおいてもほぼ横ばいとなった．4.5.2 節で議論したように，モデルの文脈表現や文表現に基づくベクトル検索では出力された知識に対応する関連事例のカバレッジが低いため，検索事例数を増やした際に関連の低い事例も混入しノイズになっていることが推定される．

4.6　本章のまとめ

本章では，言語モデルの知識評価における確信度推定に言語モデルの訓練データを利用することの有効性を検証した．確信度指標は，モデルの入出力内容に関連する事例を訓練データ全文から検索して用いることにより計算される．検証には英語 Wikipedia を用いて訓練した BERT モデルを使用し，文脈ベクトル，文ベクトル，テキスト一致による検索方式を検討した．実験の結果，訓練データに基づく確信度推定は有効であり，さらに従来の尤度や内部状態を用いる確信度指標と組み合わせることで性能が改善することが確認できた．

訓練データの検索方式としては，入出力に含まれるエンティティのテキスト一致に基づく検索が最も効果が高い結果となったが，これは事前学習済みモデルによりエンコードされた文脈ベクトルや文ベクトルを用いた関連事例検索が十分に機能していなかったことが一因と考えられる．ベクトル表現に基づく検索は入出力の形式の制限が少なく，意味的な類似を捉えられるといった利点がある．検索器の学習を伴う方法を含め，ベクトル検索の性能向上により文脈ベクトルや文ベクトルに基づく確信度指標の性能が向上できるかは今後の課題である．



(a) Google-RE

(b) T-REx

(c) ConceptNet

(d) SQuAD

図 4.5: トークン・文レベル分散表現に基づく確信度指標の検索事例数 k と性能の関係.

第 5 章

結論

本論文では，言語モデル出力の知識評価における従来の評価の枠組みを改善するため，以下の 2 点に取り組んだ．まず，既存の言語モデル知識評価タスクである LAMA probe に選択的予測を導入し，確信度を考慮した知識評価を提案した．選択的予測では，言語モデルの予測の出力可否を確信度指標に基づく選択関数に基づき判別するシステムを評価することで，個別の予測の真偽に関する判断の信頼性という実用上重要な観点を取り入れた評価が可能となる．また，LAMA probe は決められた入力に対する言語モデルの出力を予測精度に基づき評価しており，モデルの予測や評価データの偏りに起因するモデル知識の過大評価が指摘されている．これに対し，提案する選択的予測に基づく評価は，予測精度に基づく評価と比べこうした偏りの影響を低減することを実験的に確認した．提案手法はモデルの入出力や評価データの分布を改善する従来のアプローチと異なり，評価データに依存せず適用することが可能であり，こうした従来手法と組み合わせて用いることもできる．

次に，選択的予測で用いる確信度推定の改善に取り組み，言語モデルの訓練データに基づく確信度指標を提案した．3 章で導入した言語モデルの入出力や内部状態を用いる確信度指標間の比較においては，予測単語に対するモデルの予測尤度を用いる最も単純な指標の性能が一貫して高かった．4 章では言語モデルの事前学習に用いられた訓練データにアクセスできる状況を想定し，訓練データから予測内容に関連する事例を検索して用いる方法を複数検討した．訓練データからの関連事例の検索方法としては，単語単位または文単位でベクトル化してベクトル検索を行う方法と，予測に関わるエンティティのテキスト一致検索を行う方

法を検討し，それぞれについて複数の確信度指標を設計した．実験においては，エンティティに基づくテキスト一致検索で予測に関連する事例を検索できる割合が高く，検索結果を用いて予測尤度に基づく確信度を補正することにより高い効果が得られることを確認した．

以下に本研究の課題と今後の展望を述べる．

大規模言語モデルへの拡張　本研究では，言語モデルの知識評価の対象として主に BERT などのマスク言語モデルを用いた．BERT モデルは近年の大規模言語モデルと比較すると小規模なモデルであり，かつ近年主流となっているデコーダを主とする言語モデルとはモデル構造や推論の仕組みといった点において違いがある．モデルの規模やアーキテクチャの違いにより知識評価や確信度指標の効果が異なる可能性があるが，こうした比較は本研究では扱えておらず，今後さらなる検証が必要となる．

また，本研究で評価したモデルと大規模言語モデルは事前学習のための訓練データの規模や質の点でも異なる．4 章では英語 Wikipedia を用いて訓練した BERT モデルを用いて，訓練データに基づく確信度指標を評価した．Wikipedia は文体がある程度均質化されており，記述内容には誤りが含まれる場合があるものの，多くの目に触れ編集される仕組みにより大部分のページで最低限の品質が保たれているといえる．一方，近年の言語モデルの訓練データは，文体がまばらであったり真偽の確認が不十分な Web 上のテキストを含む大量のテキストデータから構成される場合が多い．こうした訓練データの量や質の違いを念頭におくと，本研究で用いた確信度推定方法を大規模言語モデルに適用するにあたり追加で検討が必要な点がいくつかある．まず，訓練データの大規模化により，本研究で構築したものと同様のベクトル表現に基づくデータストアを訓練データ全件を網羅する形で構築し保持することが現実的ではない可能性がある．その場合には，学習を伴うベクトル検索を含めた関連事例検索の高度化，確信度推定に有用な訓練データのフィルタリングによる省データ化などを検討する必要があると考えられる．また，訓練データの質のばらつきが大きくなることで，訓練データ自体の記述内容の真偽にも注意を払う必要が生じる可能性がある．学習データにおける誤りを含む記述の割合が増加したときに訓練データに基づく確信度指標の有効性が損なわれないかは明らかではなく，異なるデータを用いて訓練されたモデルに

よる検証が必要である．

訓練データの範囲外の知識　本研究では言語モデルが訓練データから獲得した知識に忠実な出力を行えるかを評価することを主目的とし，学習範囲外の知識に対する推論やその正誤判定は研究の範囲外とした．評価に用いたLAMAデータセットは英語Wikipediaに基づく一般ドメインの事実に関する知識を主な評価対象としており，より専門性の高いドメイン特化の知識に対する振舞いは評価していない．さらに，言語モデルの訓練時点からの時間経過により情報が更新され言語モデルのもつ知識が古いものになった場合に，出力の確信度を更新する手段は本研究の範囲では考慮していない．しかし，実用においては，上記のような言語モデルの学習範囲外の知識への対処がしばしば求められる．

未知の事例への対応に対する一つのアプローチとしては外部知識源の活用が考えられる．近年は外部の知識源を活用することで言語モデルの生成内容の真偽確認や生成補助を行う手法が発展しており (Rashkin et al. 2023, Gao et al. 2024)，言語モデルが訓練時に獲得した知識と外部の知識源を組み合わせて生成を行うことが主流になりつつある．こうした状況を踏まえ，外部の知識源へのアクセスが可能な状況において，言語モデルの学習範囲外の知識に対し外部知識を活用した予測を行うことを想定し，適切な確信度推定手法を開発することが求められる．

業績リスト

ジャーナル論文

1. 吉川和，岡崎直観．2025．訓練データを用いた言語モデル生成の確信度推定．自然言語処理 32 (1) (single column, 23 pages).

2. Chunpeng Ma, Aili Shen, Hiyori Yoshikawa, Tomoya Iwakura, Daniel Beck, Timothy Baldwin: On the Effectiveness of Images in Multi-modal Text Classification: An Annotation Study. 2023. ACM Transactions on Asian and Low-Resource Language Information Processing, Volume 22, Issue 3: pp. 1—19 (single column, 19 pages).

国際学会発表

1. Hiyori Yoshikawa, Naoaki Okazaki. 2023. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. Findings of the Association for Computational Linguistics: EACL 2023: pp. 2017–2028 (double columns, 12 pages).

2. Yuan Li, Jiayuan He, Hiyori Yoshikawa, Biaoyan Fang, Zenan Zhai, Christian Druckenbrodt, Camilo Thorne, Saber A Akhondi, Karin Verspoor: End-to-End Chemical Reaction Extraction from Patents. 2022. The 3rd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2022): pp. 2—3 (double columns, 2 pages).

3. Yuan Li, Biaoyan Fang, Jiayuan He, Hiyori Yoshikawa, Saber A Akhondi, Christian Druckenbrodt, Camilo Thorne, Zenan Zhai, Zubair Afzal, Trevor Cohn, Timothy Baldwin, Karin Verspoor: The ChEMU 2022 Evaluation Campaign: Information Extraction in Chemical Patents. 2022. Advances in Information Retrieval (ECIR 2022): pp. 400—407 (single column, 8 pages).

4. Yuan Li, Biaoyan Fang, Jiayuan He, Hiyori Yoshikawa, Saber A Akhondi, Christian Druckenbrodt, Camilo Thorne, Zubair Afzal, Zenan Zhai, Timothy Baldwin, Karin Verspoor: Overview of ChEMU 2022 Evaluation Campaign: Information Extraction in Chemical Patents. 2022. The 13th International Conference of the CLEF Association (CLEF 2022): pp. 521—540 (single column, 20 pages).

5. Qian Sun, Aili Shen, Hiyori Yoshikawa, Chunpeng Ma, Daniel Beck, Tomoya Iwakura, Timothy Baldwin: Evaluating Hierarchical Document Categorisation. 2021. The 19th Annual Workshop of the Australasian Language Technology Association (ALTA 2021): pp. 179-184. (double columns, 5 pages)

6. Yuan Li, Biaoyan Fang, Jiayuan He, Hiyori Yoshikawa, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Zubair Afzal, Zenan Zhai, Timothy Baldwin, Karin Verspoor. 2021. Overview of ChEMU 2021: Reaction Reference Resolution and Anaphora Resolution in Chemical Patents. The 12th International Conference of the CLEF Association (CLEF 2021): pp.292–307. (single column, 14 pages)

7. Hiyori Yoshikawa, Tomoya Iwakura, Kimi Kaneko, Hiroaki Yoshida, Yasutaka Kumano, Kazutaka Shimada, Rafal Rzepka, Patrycja Swieczkowska. 2021. Tell Me What You Read: Automatic Expertise-Based Annotator Assignment for Text Annotation in Expert Domains. Recent Advances in Natural Language Processing (RANLP 2021): pp. 1575–1585. (double columns, 8 pages)

8. Hiyori Yoshikawa, Saber A. Akhondi, Camilo Thorne, Christian Druckenbrodt, Ralph Hoessel, Zenan Zhai, Jiayuan He, Timothy Baldwin, Karin Verspoor. 2021. Chemical Reaction Reference Resolution in Patents. The 2nd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2021). (double columns, 7 pages)

9. Chunpeng Ma, Aili Shen, Hiyori Yoshikawa, Tomoya Iwakura, Daniel Beck, Timothy Baldwin. 2021. On the (In) Effectiveness of Images for Text Classification. The 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021): pp. 42–48. (double columns, 5 pages)

国内口頭発表

1. 吉川和, 岡崎直観. 2022. 確信度を考慮した言語モデルの関係知識評価. 言語処理学会第 28 回年次大会 (NLP2022). (double columns, 4 pages) [委員特別賞]

参考文献

Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari (2024) "To Believe or Not to Believe Your LLM: IterativePrompting for Estimating Epistemic Uncertainty," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Murat Seckin Ayhan and Philipp Berens (2018) "Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks," in *Medical Imaging with Deep Learning*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton (2016) "Layer Normalization."

Bernd Bohnet, Vinh Q. Tran, Pat Verga et al. (2022) "Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models," DOI: 10.48550/ARXIV.2212.08037.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi (2019) "COMET: Commonsense Transformers for Automatic Knowledge Graph Construction," in Anna Korhonen, David Traum, and Lluí s Màrquez eds. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4762–4779, Florence, Italy: Association for Computational Linguistics, July, DOI: 10.18653/v1/P19-1470.

Tom Brown, Benjamin Mann, Nick Ryder et al. (2020) "Language Models are Few-Shot Learners," in H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan,

and H. Lin eds. *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901: Curran Associates, Inc.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu (2021) "Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases," in Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli eds. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1860–1874, Online: Association for Computational Linguistics, August, DOI: 10.18653/v1/2021.acl-long.146.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li (2018) "Faithful to the Original: Fact-Aware Neural Abstractive Summarization," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18: AAAI Press.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song (2019) "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," in *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, pp. 267–284, USA: USENIX Association.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov (2019) "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context," in Anna Korhonen, David Traum, and Lluís Màrquez eds. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, Florence, Italy: Association for Computational Linguistics, July, DOI: 10.18653/v1/P19-1285.

Nicola De Cao, Wilker Aziz, and Ivan Titov (2021) "Editing Factual Knowledge in Language Models," in Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih eds. *Proceedings of the 2021 Conference on Em-*

pirical Methods in Natural Language Processing, pp. 6491–6506, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, November, DOI: 10.18653/v1/2021.emnlp-main.522.

Shrey Desai and Greg Durrett (2020) "Calibration of Pre-trained Transformers," in Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu eds. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, Online: Association for Computational Linguistics, November, DOI: 10.18653/v1/2020.emnlp-main.21.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Jill Burstein, Christy Doran, and Thamar Solorio eds. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota: Association for Computational Linguistics, June, DOI: 10.18653/v1/N19-1423.

Ran El-Yaniv and Yair Wiener (2010) "On the Foundations of Noise-Free Selective Classification," *Journal of Machine Learning Research*, Vol. 11, No. 53, pp. 1605–1641.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl (2018) "T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples," in Nicoletta Calzolari, Khalid Choukri, Christopher Cieri et al. eds. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan: European Language Resources Association (ELRA), May.

Di Feng, Lars Rosenbaum, and Klaus Dietmayer (2018) "Towards Safe Autonomous Driving: Capture Uncertainty in the Deep Neural Network For Lidar 3D Vehicle Detection," in *2018 21st International Confer-*

ence on Intelligent Transportation Systems (ITSC), pp. 3266–3273, DOI: 10.1109/ITSC.2018.8569814.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem et al. (2024) "Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities."

Yarin Gal and Zoubin Ghahramani (2016) "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, p. 1050–1059: JMLR.org.

Yunfan Gao, Yun Xiong, Xinyu Gao et al. (2024) "Retrieval-Augmented Generation for Large Language Models: A Survey."

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali et al. (2023) "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, Vol. 56, No. 1, pp. 1513–1589, October, DOI: 10.1007/s10462-023-10562-9.

Yonatan Geifman and Ran El-Yaniv (2017) "Selective Classification for Deep Neural Networks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4885–4894: Curran Associates Inc.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv (2019) "Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers," in *International Conference on Learning Representations*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych (2024) "A Survey of Confidence Estimation and Calibration in Large Language Models," in Kevin Duh, Helena Gomez, and Steven Bethard eds. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, Mexico City, Mexico: Association for Computational Linguistics, June.

Google (2013) “Google Relation Extraction (Google-RE) Corpus,” Accessed: 2024-12-31.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger (2017) “On calibration of modern neural networks,” in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, p. 1321–1330: JMLR.org.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang (2024) “Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey,” *Transactions on Machine Learning Research*.

Guozhi Hao, Jun Wu, Qianqian Pan, and Rosario Morello (2024) “Quantifying the uncertainty of LLM hallucination spreading in complex adaptive social networks,” *Scientific Reports*, Vol. 14, No. 1, p. 16375, July, DOI: 10.1038/s41598-024-66708-4.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016) “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, DOI: 10.1109/CVPR.2016.90.

Lei Huang, Weijiang Yu, Weitao Ma et al. (2024) “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” *ACM Trans. Inf. Syst.*, November, DOI: 10.1145/3703155, Just Accepted.

Eyke Hüllermeier and Willem Waegeman (2021) “Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods,” *Machine Learning*, Vol. 110, No. 3, pp. 457–506, March, DOI: 10.1007/s10994-021-05946-3.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung (2023) “Towards Mitigating LLM Hallucination via Self Reflection,” in Houda Bouamor, Juan Pino, and Kalika Bali eds. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843,

Singapore: Association for Computational Linguistics, December, DOI: 10.18653/v1/2023.findings-emnlp.123.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch et al. (2023) "Mistral 7B."

Hang Jiang, Xiajie Zhang, Robert Mahari et al. (2024) "Leveraging Large Language Models for Learning Complex Legal Concepts through Storytelling," in Lun-Wei Ku, Andre Martins, and Vivek Srikumar eds. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7194–7219, Bangkok, Thailand: Association for Computational Linguistics, August, DOI: 10.18653/v1/2024.acl-long.388.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig (2020) "How Can We Know What Language Models Know?" *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 423–438, DOI: 10.1162/tacl_a_00324.

Qiao Jin, Zifeng Wang, Charalampos S. Floudas et al. (2024) "Matching patients to clinical trials with large language models," *Nature Communications*, Vol. 15, No. 1, p. 9074, November, DOI: 10.1038/s41467-024-53081-z.

Jeff Johnson, Matthijs Douze, and Hervé Jégou (2019) "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, Vol. 7, No. 3, pp. 535–547.

Saurav Kadavath, Tom Conerly, Amanda Askell et al. (2022) "Language Models (Mostly) Know What They Know."

Amita Kamath, Robin Jia, and Percy Liang (2020) "Selective Question Answering under Domain Shift," in Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, Online: Association for Computational Linguistics, July, DOI: 10.18653/v1/2020.acl-main.503.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel (2023) "Large language models struggle to learn long-tail knowledge," in *Proceedings of the 40th International Conference on Machine Learning*, ICML'23: JMLR.org.

Jared Kaplan, Sam McCandlish, Tom Henighan et al. (2020) "Scaling Laws for Neural Language Models."

Nora Kassner and Hinrich Schütze (2020) "Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly," in Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7811–7818, Online: Association for Computational Linguistics, July, DOI: 10.18653/v1/2020.acl-main.698.

Nora Kassner, Philipp Dufter, and Hinrich Schütze (2021) "Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models," in Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty eds. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3250–3258, Online: Association for Computational Linguistics, April, DOI: 10.18653/v1/2021.eacl-main.284.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis (2020) "Generalization through Memorization: Nearest Neighbor Language Models," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*: OpenReview.net.

Omar Khattab and Matei Zaharia (2020) "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, p. 39–48, New York, NY, USA: Association for Computing Machinery, DOI: 10.1145/3397271.3401075.

Alexander Kurz, Katja Hauser, Hendrik Alexander Mehrtens et al. (2022) "Uncertainty Estimation in Medical Image Classification: Systematic Review.," *JMIR medical informatics*, Vol. 10, No. 8, p. e36427, DOI: 10.2196/36427.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield et al. (2019) "Natural Questions: A Benchmark for Question Answering Research," *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 452–466, DOI: 10.1162/tacl_a_00276.

David D. Lewis and William A. Gale (1994) "A Sequential Algorithm for Training Text Classifiers," in Bruce W. Croft and C. J. van Rijsbergen eds. *SIGIR '94*, pp. 3–12, London: Springer London.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (2020) "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online: Association for Computational Linguistics, July, DOI: 10.18653/v1/2020.acl-main.703.

Patrick Lewis, Ethan Perez, Aleksandra Piktus et al. (2020) "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin eds. *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9459–9474: Curran Associates, Inc.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg (2023) "Inference-Time Intervention: Eliciting Truthful Answers from a Language Model," in *Thirty-seventh Conference on Neural Information Processing Systems*.

Shiyu Liang, Yixuan Li, and R. Srikant (2018) "Enhancing The Reliability of

Out-of-distribution Image Detection in Neural Networks," in *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal et al. (2019) "RoBERTa: A Robustly Optimized BERT Pretraining Approach," DOI: 10.48550/ARXIV.1907.11692.

LLM-jp (2024) "LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs."

Aman Madaan, Niket Tandon, Prakhar Gupta et al. (2023) "Self-Refine: Iterative Refinement with Self-Feedback."

Potsawee Manakul, Adian Liusie, and Mark Gales (2023) "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," in Houda Bouamor, Juan Pino, and Kalika Bali eds. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore: Association for Computational Linguistics, December, DOI: 10.18653/v1/2023.emnlp-main.557.

Peter Mattson, Christine Cheng, Gregory Diamos et al. (2020) "MLPerf Training Benchmark," in I. Dhillon, D. Papailiopoulos, and V. Sze eds. *Proceedings of Machine Learning and Systems*, Vol. 2, pp. 336–349.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald (2020) "On Faithfulness and Factuality in Abstractive Summarization," in Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online: Association for Computational Linguistics, July, DOI: 10.18653/v1/2020.acl-main.173.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau (2022) "Reducing Conversational Agents' Overconfidence Through Linguistic Calibration," *Transactions of the Association for Computational Linguistics*, Vol. 10, pp. 857–872, DOI: 10.1162/tacl_a_00494.

OpenAI (2022) "ChatGPT: Optimizing Language Models for Dialogue," Accessed: 2024-12-31.

Nicolas Papernot and Patrick McDaniel (2018) "Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning."

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018) "Deep Contextualized Word Representations," in Marilyn Walker, Heng Ji, and Amanda Stent eds. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana: Association for Computational Linguistics, June, DOI: 10.18653/v1/N18-1202.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller (2019) "Language Models as Knowledge Bases?" in Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan eds. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Hong Kong, China: Association for Computational Linguistics, November, DOI: 10.18653/v1/D19-1250.

Fabio Petroni, Aleksandra Piktus, Angela Fan et al. (2021) "KILT: a Benchmark for Knowledge Intensive Language Tasks," in Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer et al. eds. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, Online: Association for Computational Linguistics, June, DOI: 10.18653/v1/2021.naacl-main.200.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers (2023) "The ROOTS Search Tool: Data Transparency for LLMs," in Danushka Bollegala, Ruihong Huang, and Alan Ritter eds. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*,

pp. 304–314, Toronto, Canada: Association for Computational Linguistics, July, DOI: 10.18653/v1/2023.acl-demo.29.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze (2020) "E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT," in Trevor Cohn, Yulan He, and Yang Liu eds. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 803–818, Online: Association for Computational Linguistics, November, DOI: 10.18653/v1/2020.findings-emnlp.71.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018) "Improving Language Understanding by Generative Pre-Training," technical report, OpenAI.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019) "Language Models are Unsupervised Multitask Learners."

Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert D. Kleinberg, Sendhil Mullainathan, and Jon M. Kleinberg (2018) "Direct Uncertainty Prediction for Medical Second Opinions," in *International Conference on Machine Learning*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang (2016) "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in Jian Su, Kevin Duh, and Xavier Carreras eds. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas: Association for Computational Linguistics, November, DOI: 10.18653/v1/D16-1264.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm et al. (2023) "Measuring Attribution in Natural Language Generation Models," *Computational Linguistics*, Vol. 49, No. 4, pp. 777–840, 12, DOI: 10.1162/coli_a_00486.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu (2023) "Out-of-Distribution Detection and

Selective Generation for Conditional Language Models," in *The Eleventh International Conference on Learning Representations*.

Adam Roberts, Colin Raffel, and Noam Shazeer (2020) "How Much Knowledge Can You Pack Into the Parameters of a Language Model?" in Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu eds. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5418–5426, Online: Association for Computational Linguistics, November, DOI: 10.18653/v1/2020.emnlp-main.437.

Marc Rußwurm, Mohsin Ali, Xiao Xiang Zhu, Yarin Gal, and Marco Körner (2020) "Model and Data Uncertainty for Satellite Time Series Forecasting with Deep Recurrent Models," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7025–7028, DOI: 10.1109/IGARSS39084.2020.9323890.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff (2020) "Masked Language Model Scoring," in Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2699–2712, Online: Association for Computational Linguistics, July, DOI: 10.18653/v1/2020.acl-main.240.

Robin Senge, Stefan Bösner, Krzysztof Dembczyński, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier (2014) "Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty," *Information Sciences*, Vol. 255, pp. 16–29, DOI: https://doi.org/10.1016/j.ins.2013.07.030.

Murat Sensoy, Lance Kaplan, and Melih Kandemir (2018) "Evidential deep learning to quantify classification uncertainty," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, p. 3183–3193, Red Hook, NY, USA: Curran Associates Inc.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan

Li Zhu (2002) “Open Mind Common Sense: Knowledge Acquisition from the General Public,” in Robert Meersman and Zahir Tari eds. *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pp. 1223–1237, Berlin, Heidelberg: Springer Berlin Heidelberg.

Robyn Speer and Catherine Havasi (2012) “Representing General Relational Knowledge in ConceptNet 5,” in Nicoletta Calzolari, Khalid Choukri, Thierry Declerck et al. eds. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC‘12)*, pp. 3679–3686, Istanbul, Turkey: European Language Resources Association (ELRA), May.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014) “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, Vol. 15, No. 56, pp. 1929–1958.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang (2021) “Can Language Models be Biomedical Knowledge Bases?” in Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih eds. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4723–4734, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, November, DOI: 10.18653/v1/2021.emnlp-main.388.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal (2018) “FEVER: a Large-scale Dataset for Fact Extraction and VERification,” in Marilyn Walker, Heng Ji, and Amanda Stent eds. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana: Association for Computational Linguistics, June, DOI: 10.18653/v1/N18-1074.

Hugo Touvron, Thibaut Lavril, Gautier Izacard et al. (2023) “LLaMA: Open and Efficient Foundation Language Models.”

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (2017) "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 6000–6010: Curran Associates Inc.

Denny Vrandečić and Markus Krötzsch (2014) "Wikidata: a free collaborative knowledgebase," *Commun. ACM*, Vol. 57, No. 10, p. 78–85, September, DOI: 10.1145/2629489.

Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren (2019) "Automatic Brain Tumor Segmentation Using Convolutional Neural Networks with Test-Time Augmentation," in Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum eds. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 61–72, Cham: Springer International Publishing.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei (2024) "Text Embeddings by Weakly-Supervised Contrastive Pre-training."

Jason Wei, Xuezhi Wang, Dale Schuurmans et al. (2023) "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models."

Johnny Wei, Ryan Wang, and Robin Jia (2024) "Proving membership in LLM pretraining data via data watermarks," in Lun-Wei Ku, Andre Martins, and Vivek Srikumar eds. *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13306–13320, Bangkok, Thailand: Association for Computational Linguistics, August, DOI: 10.18653/v1/2024.findings-acl.788.

BigScience Workshop (2023) "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model."

Yonghui Wu, Mike Schuster, Zhifeng Chen et al. (2016) "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation."

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi (2024) "Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs," in *The Twelfth International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning (2018) "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering," in Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii eds. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium: Association for Computational Linguistics, October-November, DOI: 10.18653/v1/D18-1259.

Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, and Desmond Ong (2024) "Large Language Models are Capable of Offering Cognitive Reappraisal, if Guided," in *First Conference on Language Modeling*.

Michael Zhang and Eunsol Choi (2021) "SituatedQA: Incorporating Extra-Linguistic Contexts into QA," in Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih eds. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7371–7387, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, November, DOI: 10.18653/v1/2021.emnlp-main.586.

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng (2024) "Pretraining Data Detection for Large Language Models: A Divergence-based Calibration Method," in Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen eds. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5263–5274, Miami,

Florida, USA: Association for Computational Linguistics, November, DOI: 10.18653/v1/2024.emnlp-main.300.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang (2023) "How Do Large Language Models Capture the Ever-changing World Knowledge? A Review of Recent Advances," in Houda Bouamor, Juan Pino, and Kalika Bali eds. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8289–8311, Singapore: Association for Computational Linguistics, December, DOI: 10.18653/v1/2023.emnlp-main.516.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang (2021) "Enhancing Factual Consistency of Abstractive Summarization," in Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer et al. eds. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 718–733, Online: Association for Computational Linguistics, June, DOI: 10.18653/v1/2021.naacl-main.58.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015) "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 19–27, DOI: 10.1109/ICCV.2015.11.