大規模言語モデルにおけるハルシネーションとその対策

2024年7月17日

国立情報学研究所 大規模言語モデル研究開発センター
特任研究員 蔦 侑磨

大規模言語モデル（LLM）

- 翻訳・要約・対話など様々な用途に利用される

翻訳

要約

対話型サービス（ChatGPT）

- LLMの課題
  - 正しい出力を行えているか（**関連課題：ハルシネーション**）
  - プライバシーとセキュリティ
  - 膨大な計算リソース

LLMの課題：ハルシネーション

- ハルシネーションとは？
  - 生成された文が現実の事実やユーザーの入力と一致しない現象
  - 入力例１：月に最初に着陸した人は誰？
    - 誤回答例：人類で初めて月に降り立ったのはチャールズ・リンドバーグです
  - 入力例２：唾液アミラーゼの働きは？
    - 誤回答例：唾液アミラーゼは、食べ物に含まれるでんぷんを分解し、胃で消化されやすい状態にする
    - 引用https://www.yomiuri.co.jp/kyoiku/kyoiku/news/20240306-OYT1T50080/
- **情報の正確性の重要な分野**では重大な問題
  - 医療、法務、ニュースなど

以降では次の論文を参考に説明

- A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions

ハルシネーションの種類

- 事実性ハルシネーションと忠実性ハルシネーション
  - 事実性ハルシネーション（**factuality hallucination**）
    - **事実と異なる情報を生成**
  - 忠実性ハルシネーション（**faithfulness hallucination**）
    - **ユーザーの指示に反する生成**
- 内因性・外因性ハルシネーションという分類もある
  - 詳細は割愛

事実性ハルシネーションの具体例

事実性ハルシネーション：**事実と異なる情報を生成**

- **事実的不一致（ factual inconsistency ）**
  - 入力: 月に最初に着陸した人は誰ですか？
  - 誤回答例：チャールズ・リンドバーグ
  - 正回答例：ニール・アームストロング

- **事実の捏造（ factual fabrication ）**
  - 入力：ユニコーンの起源について教えてください
  - 誤回答例：
    - ユニコーンは、紀元前10,000年頃にアトランティスの平原を駆け抜けたとされ、王族と神聖視されていました

忠実性ハルシネーションの具体例

- 忠実性ハルシネーション: **ユーザーの指示に反する生成**
  - **指示の不一致 (instruction inconsistency)**
    - 次の英文を翻訳して “What is the capital of Japan?”
    - 誤回答例：The capital of Japan is Tokyo.

  - **コンテキストの不一致(context inconsistency)**
    - 入力：次の文を要約して「信濃川は日本で最も長い川で、**新潟県と長野県**を流れています。川の**源流は長野県の山間部にあり**、日本海に注ぎます。」

    - 誤回答例：信濃川の**源流は新潟県にあり**、日本海に注ぎます。

  - **論理的不一致（ logical inconsistency ）**
    - 入力：次の方程式をステップバイステップで解いてください: 2x + 3 = 11
    - 誤回答例：ステップ1：両辺から3を引いて、$2x = 8$ とします。
      ステップ2：両辺を2で割って、**$x = 3$ とします**

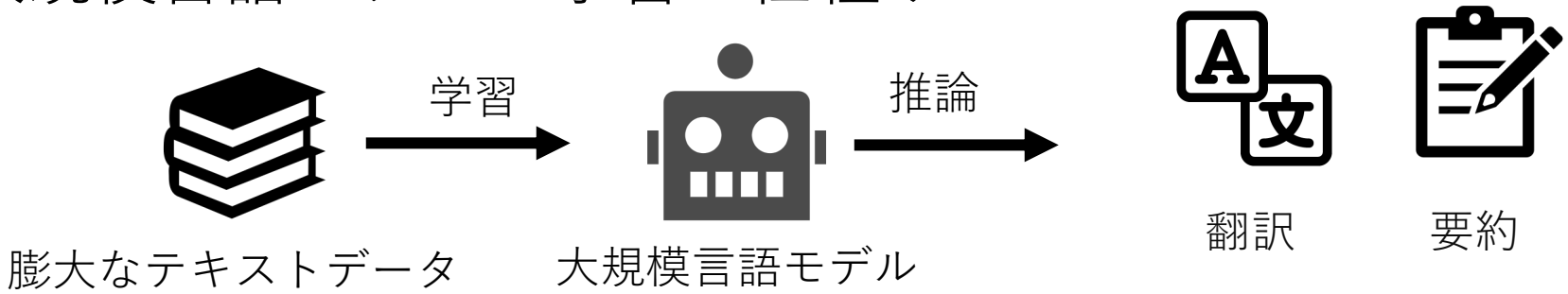
話の流れ

- イントロダクション
- ハルシネーションの概要
- ハルシネーションの種類
- **ハルシネーションの原因**
- ハルシネーションの対策
  - 検出方法
  - 緩和策
- 結論と今後の課題

予備知識：大規模言語モデルの学習について

予備知識：大規模言語モデルの学習

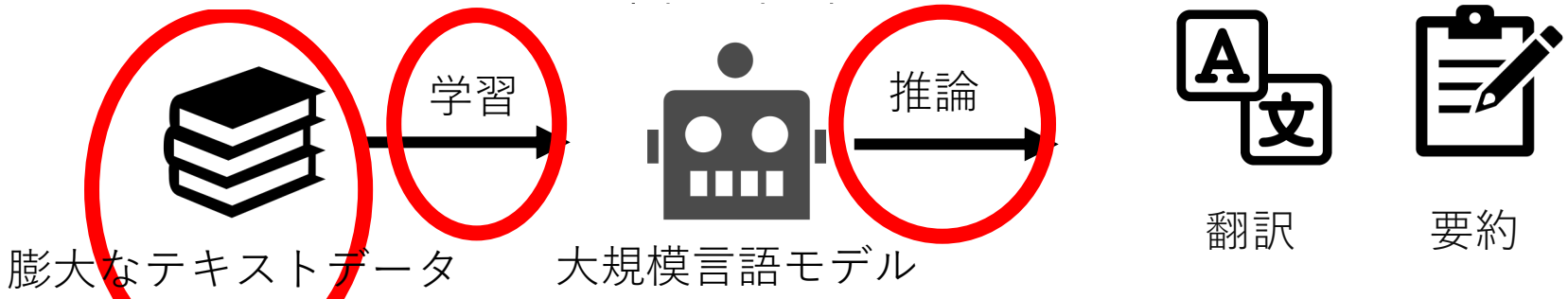
- 大規模言語モデルの学習の仕組み



- 主な学習ステップ
  1. **事前学習**：入力文の次のトークン（≒単語）を予測
     - 例：（入力）私は朝にパンを -> （予測）食べた
  2. **ファイチューニング**：タスク（翻訳や要約など）の入出力を学習
  3. **RLHF**：人にとって魅力的な文章を学習
     - 正解文が必ずしも人にとって魅力的でない
     - **R**einforcement **L**earning from **H**uman **F**eedback の略

ハルシネーションの原因

データの問題　学習方法の問題　推論時の問題



- 主な学習ステップ
  1. **事前学習**：入力文の次のトークン（≒単語）を予測
     - 例：（入力）私は朝にパンを -> （予測）食べた
  2. **ファイチューニング**：タスク（翻訳や要約など）の入出力を学習
  3. **RLHF**：人にとって魅力的な文章を学習
     - 正解文が必ずしも人にとって魅力的でない
     - **R**einforcement **L**earning from **H**uman **F**eedback の略

ハルシネーションの原因（一部）

- **データの質の問題**
  - データ中の**誤った情報**
  - データに含まれる**バイアス**
    - 社会的バイアス：医者は男性が多い
  - **専門知識や最新の知識の不足**

- **学習方法の問題**
  - 人に好ましい応答の学習（RLHF）により**人に合わせた回答を行う**
    - 入力：地球は平面ですか？ -> 回答：はい、地球は平らです [Sharma et al., 2023]

- **推論時の問題**
  - 出力に過度に注意することにより**入力のコンテキストを無視**する

ハルシネーションへの対策

- **ハルシネーション検出**
  - モデルのハルシネーションを検知することで以下が可能に
    - 性能評価
    - 生成後の事後編集・文の再出力
    - 生成テキストの正確性への注意喚起

- **ハルシネーション緩和**
  - **問題を軽減**できる方法を導入
    - データの質の問題
    - トレーニング方法の問題
    - 推論アルゴリズムの問題

ハルシネーションの検出方法

- **外部の事実を利用した検証**
  - 生成文を信頼できる外部の情報源と比較して検証

- **出力の不確実性(uncertainty)を推定**
  - 不確実性：**どの程度同じ出力を一貫して行いにくいか**
  - モデルの内部状態や挙動から不確実性を推定し、ハルシネーションの可能性を評価
    - **想定：ハルシネーションする=情報源が曖昧なため、出力が不確実**

ハルシネーションの検出方法：外部の事実を利用した検証

- **方法**:
  1. ウェブやデータベースなどの**外部情報源から情報を収集**
  2. 収集した情報と生成文で**矛盾がないか確認**

- **具体例**:
  1. **質問：**ヒマラヤ山脈の最高峰は何ですか？
  2. **出力：**ヒマラヤ山脈の最高峰は**K2**です。
  3. **検証：**「エベレスト」が正しい答え（**外部情報源を参照**）

- **利点と欠点**
  - **利点**: 精度が高く、信頼性の高い結果を得られる
  - **欠点**: リアルタイムでの外部情報源へのアクセスが必要であり、時間がかかる

ハルシネーションの検出方法：出力の不確実性を推定

- **想定：ハルシネーションする=情報源が曖昧なため出力が不確実**

- **方法１：モデルの内部状態の利用**
  - トークン出力確率などの内部状態を分析し不確実性を評価
    - トークンの出力確率が低い場合、モデルは他の出力を考慮するため不確実性が高い

- **方法２：モデルの挙動の観察**
  - モデルの出力のパターンを観察し不確実性を推定
    - 同じ質問に対して異なる回答を生成する場合、不確実性が高い

- **利点と欠点**
  - 利点: 外部情報源に依存せず、リアルタイムでの検出が可能。
  - 欠点: 不確実性推定の精度が外部情報源を利用する方法よりも低い

ハルシネーションの緩和策（一部）

- データの質の向上
  - 誤情報の削減： データクリーニング
  - バイアスの軽減：多様なデータソースからのデータ収集
  - データソースを検索可能な仕組みを導入：Retrieval-Augmented Generation model [Lewis et al., 2020]

- トレーニング目標の改善
  - 人間のフィードバックを吟味する

- 推論アルゴリズムの改善
  - アンサンブルモデル：
    - 複数のモデルの出力を組み合わせて、最も信頼性の高い出力を選択

まとめ

- ハルシネーションについて包括的に紹介
  - 種類・原因・対策方法など

- 今後の課題と方向性
  - **ハルシネーションの定義について統一的な基準がない**
    - タスク依存性が高く汎用的な評価方法の確立が困難
  - **ハルシネーションの完全な防止は可能か？**
    - ハルシネーションを完全に防ぐか、最小限に抑えるか
  - **ユーザーインターフェースの構築**
    - ハルシネーションを簡単に識別し、フィードバックを提供できるインターフェースの設計も考慮されたい（情報源の提供など）