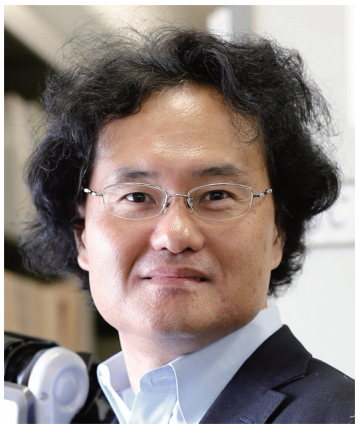
第１回 IGI セミナー（学内）

お茶の水女子大学
Ochanomizu University

事前学習済み言語モデル BERTの注意機構に着目したジェンダーバイアスの要因分析



インターネット上のコーパスで学習された大規模言語モデル（LLM）には社会的バイアスが含まれている可能性がある。本研究では、LLM 構築の基盤技術となった事前学習済み BERT に対して、職業に基づくジェンダーバイアスを対象に、その言語モデルを構築した際のアーキテクチュアとなった Transformer の注意機構を分析した。性別語のみが異なる文を比較し、自己注意の重みと文脈の男女らしさとの関係を可視化した結果、モデルの後半層における注意機構で職業語から性別語への注意（Attention）差とバイアスの相関が確認された。さらに単語間の注意の偏向を矯正する教師パターン行列で注意を誘導しバイアス軽減を試みたところ、既存評価指標の改善は限定的だったが、性別単語である“He”および“She”の出力確率の偏りは緩和された。

講師：小林　一郎　（理学部 情報科学科 教授）

お茶の水女子大学 基幹研究院 自然科学系 教授。
産業技術総合研究所人工知能研究センター 招聘研究員、情報通信研究機構 脳情報通信融合研究センター 特別研究員。
人工知能学会理事（2016–2018）、日本機能言語学会理事（2002–2015）を歴任。
言語知能の計算機的実現を目指し、自然言語処理・人工知能・機能言語学などを融合した研究に従事。「コンピュータが考える—人工知能とはなにか」、「人工知能の基礎」、「ことばは生きている—選択体系機能言語学序説」等の著書がある。

日　時：2026年5月20日（水）
9:30-10:30

対　象：本学の学生・教職員
※お茶大のメールアドレスをお持ちの方

開催方式：オンライン方式（Zoom ミーティング）

要申込

締切 5月19日（火）12:00 まで

【主催】ジェンダード・イノベーション研究所（IGI）【お問合せ先】ocha-igi@cc.ocha.ac.jp