2024 年度年次報告書
信頼される AI システムを支える基盤技術
2020 年度採択研究代表者

乾 健太郎

東北大学 言語 AI 研究センター 教授 /
Professor Mohamed bin Zayed University of Artificial Intelligence /
理化学研究所 革新知能統合研究センター チームディレクター

知識と推論に基づいて言語で説明できる AI システム

主たる共同研究者:
久木田 水生（名古屋大学 大学院情報学研究科 准教授）
黒橋 禎夫（京都大学 大学院情報学研究科 特定教授 / 国立情報学研究所長）
戸次 大介（お茶の水女子大学 基幹研究院 教授）

研究成果の概要

本研究は、自分の判断を言語で説明することができ、対話的な説明コミュニケーションを通して人の判断を支援するAIシステムの設計論の構築を目指している。

乾 G は、言語モデル（LM）が生成する説明の誠実性（モデル内部の推論プロセスを反映した説明になっているか）をどう保証するかという課題に説明生成の内部機序解明の観点から取り組み、実体概念の数値特徴のような世界知識が低次元の線形部分空間にエンコードされており（Heinzerling+ ACL 2024）、そうした解釈性の高い部分空間が実際の推論に活用されている（Oumer et al.+ NAACL 2025）といったLMの知識表象と推論過程の解明を進めた。また、複数ステップからなる推論におけるLMの解探索の機序解明にも取り組み、推論と説明の関係性（すなわち説明の誠実性）に関わる一連の知見を得た（Aoki+ EMNLP 2024）。

戸次 G は、2024 年度は、研究実施項目【DNN と相互作用する高階論理推論】において、Neural DTS の実装とその分析（飯沼+2025）や、定理自動証明器 wani のニューラル実装に向けた研究（宮川+2025）を行ったほか、CCG 統語解析器 lightblue と wani を組み合わせた自然言語推論システムの実装（富田+2025）を行った。また、研究実施項目【説明可能 AI の証明論的意味論】においては、日本語関係節における弱交差現象の分析(R-WCO)（Fukushima+2024, 福島+2025）、DTS を用いた一般化交差現象(GCO)の分析(Matsuoka+2024a), 二重目的語構文における弱交差現象(WCO)の分析（藤田+2025）を行い、LFS の方法論に基づいて形式文法の経験的検証を行うという研究プログラムを確立しつつある。その他、DTS による日本語のテンスの分析(Matsuoka+2024b)、DTS の様相拡張である Modal DTS の研究（飯村+2025）、CCG 統語解析器 lightblue のインターフェースの改良（佐伯+2025）等を行った。

黒橋 G は、話者の内的状態（知識や興味）を考慮した応答生成を目的に、2 万以上の発話からなる内部状態注釈付き映画推薦対話データセット「RecomMind」を構築した。このデータセットを用いて推薦成功に寄与する内部状態を分析し、その知見に基づき内部状態を効果的に考慮する応答生成モデルを提案した。また、言語モデルに内在するバイアスの検出・緩和に向けた予備的検討として、アンカリング効果と呼ばれる認知バイアスに着目し、GPT-4o を含む複数の言語モデルに対して様々なシナリオでの影響を検証した結果、言語モデルが人間同様にアンカリング効果を受けることを確認した。

久木田 G では、前年度に引き続き誤情報に対する人間の心理と行動の研究を行なっている。特に昨年度は、誤情報に対する訂正情報をデバイス上で対話的に提示することが信念の改訂にとってポジティブな効果をもたらす可能性を探究するために実験用の対話システムを構築し、予備的な実験を行った。その予備実験ではこの仮説が肯定される可能性が示されている。またそれらの実証的な研究に並行して、人工知能や、広く情報技術に関連する ELSI（倫理的法的社会的問題）や誤情報等の課題についての研究を継続して行い、多様な専門家を招いた研究会等を開催すると同時に、それらの問題について啓発する活動（講演、論文など）を行った。

【代表的な原著論文情報】

1) Benjamin Heinzerling and Kentaro Inui. Monotonic Representation of Numeric Attributes in Language Models. In Proceedings of the 62nd Annual Meeting of the Association for

Computational Linguistics (ACL 2024), pp.175–195, August 2024.
2) Yoichi Aoki, Keito Kudo, Tatsuki Kuribayashi, Shusaku Sone, Masaya Taniguchi, Keisuke Sakaguchi, Kentaro Inui. First Heuristic Then Rational: Dynamic Use of Heuristics in Language Model Reasoning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), pp.14255-14271, November 2024.
3) Haruka Fukushima, Daniel Plesniak, Daisuke Bekki. Matrix and relative weak crossover effects in Japanese: An experimental investigation. Proceedings of the 2024 SMOG International Conference on Syntax and Semantics, Andong, Korea, August 2024.
4) Daiki Matsuoka, Daisuke Bekki, Hitomi Yanaka. A propositions-as-types approach to the generalized crossover effect. In Proceedings of Sinn und Bedeutung 29, Siracusa, Italy, September 2024.
5) Daiki Matsuoka, Daisuke Bekki, Hitomi Yanaka. Relative Tense in Japanese: A Case of Multiply Embedded Relative Clauses. In Proceedings of the 31st Japanese-Korean Linguistic Conference (JK31), Melbourne, Australia, September 2024.
6) Takashi Kodama, Hirokazu Kiyomaru, Yin Jou Huang, Sadao Kurohashi. RecomMind: Movie Recommendation Dialogue with Seeker's Internal State. In Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024), pp.46–63, November 2024.
7) 武並佳輝, Yin Jou Huang, 村脇有吾, Chenhui Chu. LLM による価格交渉シミュレーションにおけるアンカリング効果の検証, 言語処理学会 第 31 回年次大会 発表論文集, pp.920-925, 2025 年 3 月.
8) 呉羽真, 久木田水生, 藤川直也 「メタバースは解放をもたらすか？――改善論の立場から」, 『科学基礎論研究』, 52 巻 1-2 号,pp. 1-14, 2025 年 3 月.
https://doi.org/10.4288/kisoron.52.1-2_1