

2023 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 大学院情報科学研究科 教授 / Visiting Professor
Mohamed bin Zayed University of Artificial Intelligence

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 では、大規模言語モデル (LM) の説明を納得・信頼に繋げるための内部機序解明の研究に注力し、Transformer の数学的分解に基づく個々の構成要素の機能解明 (Kobayashi+ 2023)、LM における記号的概念や関係知識・数量知識の内部表象の性質解明、推論過程の説明を使った学習による頑健性の改善、推論課題における構成性の汎化能力等の解明 (Kudo+ 2023) などを進めた。また、説明付き日本語マルチホップ QA データセット JEMHopQA を構築し、既存 LM の説明能力の問題点を明らかにした。

戸次 G では、DNN と相互作用する高階論理推論の研究を進め、Neural DTS のプロトタイプ実装、組合せ範疇文法と再帰的 NN 文法を融合する枠組み RNN-CCG の開発等を行うとともに、DTS を用いた弱交差現象の分析 (Bekki 2023)、LFS という方法論に基づいて経験的検証を行った。また、Coq による DTS 推論の実装、言語学的に妥当な CCG ツリーバンクの自動構築手法の開発を行った。

黒橋 G では、情報推薦型の説明対話における魅力的な応答の生成に関する研究を行った。本研究で構築した外部知識に基づく人間間対話データセット JMRD に対して個々の発話の情報源のタイプを注釈付けし、外部知識に加えて話者由来の情報が魅力的な応答の鍵になること、現在主流の対話システムの訓練方法では人間ほどは話者由来情報を生成できないことを示した。成果論文は ACL 2023 SRW で Best Paper Award を受賞した。

久木田 G では、誤情報に対する人間の心理と行動の研究として一連のオンライン実験を行ない、自分が信じている誤情報を打ち消すファクトチェック記事を選択的に回避する人が予想以上に多いことを明らかにするとともに、記事の提示順序や提示方法などの介入が誤情報の受容や信念更新に好影響を与える可能性を示した (Tanaka+ 2023)。また、生成 AI の ELSI や誤情報等の課題について多様な専門家を招いた研究会等を開催した。

【代表的な原著論文情報】

- 1) Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi and Kentaro Inui. Transformer Language Models Handle Word Frequency in Prediction Head. In Findings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023 Findings), pp.4523-4535, July 2023.
- 2) Keito Kudo, Yoichi Aoki, Tatsuki Kuribayashi, Ana Brassard, Masashi Yoshikawa, Keisuke Sakaguchi and Kentaro Inui. Do Deep Neural Networks Capture Compositionality in Arithmetic Reasoning? In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023), pp.1343-1354, May 2023.
- 3) Daisuke Bekki, Ribeka Tanaka, Yuta Takahashi. 2023. Integrating Deep Neural Network with Dependent Type Semantics, In Logic and Algorithms in Computational Linguistics 2021 (LACompLing2021), Roussanka Loukanova, Peter LeFanu Lumsdaine, Reinhard Muskens (eds.), Studies in Computational Intelligence, vol.1081. pp.261-284, Springer.

- 4) Takashi Kodama, Hirokazu Kiyomaru, Yin Jou Huang, Taro Okahisa, Sadao Kurohashi . Is a knowledge-based response engaging?: An analysis on knowledge-grounded dialogue with information source annotation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pp. 237-243, July 2023.
- 5) Yuko Tanaka, Miwa Inuzuka, Hiromi Arai, Yoichi Takahashi, Minao Kukita and Kentaro Inui. Who Does Not Benefit from Fact-Checking Websites? A Psychological Characteristic Predicts the Selective Avoidance of Clicking Uncongenial Facts. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI 2023), pp.1–17, April 2023.