AI 活用で挑む学問の革新と創成　　2022 年度
2021 年度採択研究代表者　　年次報告書

平岡 達也

富士通（株） 富士通研究所
研究員

人間と AI の双方に扱いやすいことばの単位の創出

研究成果の概要

2022 年度は、「人間と AI の双方に扱いやすいことばの単位の創出」の小課題として、①AI に扱いやすいことばの単位の収集と②人間に扱いやすいことばの単位の収集の両方面に取り組んだ.

①については, AI に扱いやすいことばの単位が不明であることから, 様々なことばの単位を用いて AI モデルを学習して性能向上を得る方法（サブワード正則化と呼ばれる）に取り組んだ. 特に, テキストをことばの単位に分割する処理（トークナイゼーション）を高速に行うことができる最長一致法にサブワード正則化を適用するための手法を提案し, 自然言語処理の有名国際会議にて発表した（論文 1）. また, AI に扱いやすいことばの単位そのものを見つける手法として, すでに学習済みの機械学習モデルを活用する手法を新たに提案し, 国内会議で発表した（論文 2）.

②については, 様々なトークナイゼーション手法によることばの区切り方に対して, 人間による扱いやすさのアノテーションを収集し, 分析のためのデータを構築した. 具体的には, 日本語の常識に関するクイズについて, 問題文を様々な区切り方で提示し, アノテーターによる正解率や回答時間にどのような影響が出るかを調査した. 現時点での本調査の結果は, 国内学会にて発表済みである（論文 3）. なお, アノテーションによる調査は現時点で課題が山積しており, 2023 年度も引き続きアノテーションを進め, 質の高いデータ収集を目指す. 収集したデータを用いて, 人間と AI それぞれに扱いやすいことばの単位の性質を解明し, 双方に扱いやすいことばの単位を折衷できるかに取り組むことが, 本研究の大目標である.

論文（2, 3）については, さらなる実験の拡充を行った上で, 英語版を用意している. 本研究の国際的な認知を得るためにも, 2023 年度は国際学会への投稿を行う予定である.

【代表的な原著論文情報】

1) Tatsuya Hiraoka. MaxMatch-Dropout: Subword Regularization for WordPiece. In Proceedings of the 29th International Conference on Computational Linguistics (COLING), pages 4864-4872, Gyeongju, Republic of Korea, October 2022.

2) 平岡達也, 岩倉友哉. 人間と機械学習のモデルそれぞれに扱いやすいトークン分割に関する実験と考察. 言語処理学会第 29 回年次大会 (NLP2023), pp. 727-731, 2023 年 3 月.

3) 平岡達也, 岩倉友哉. 語彙制約付きニューラル単語分割器を用いた後処理としての単語分割の後段タスクへの最適化. 言語処理学会第 29 回年次大会 (NLP2023), pp. 1503-1507, 2023 年 3 月.