AI 活用で挑む学問の革新と創成
2021 年度採択研究者

2021 年度
年次報告書

平岡 達也

東京工業大学 情報理工学院
大学院生(博士後期課程)

人間と AI の双方に扱いやすいことばの単位の創出

§1. 研究成果の概要

2021 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、AI に扱いやすいことばの単位を獲得する機械学習手法の開発を行った。ことばを扱う AI について研究する自然言語処理分野では、一般的にテキストを単語などの小さい単位に分割してから AI へと入力する。この時、人間に扱いやすいことばの単位にテキストを分割することが一般的である。しかしながら、人間に扱いやすいことばの単位が必ずしも AI の性能向上に繋がるとは限らないということが分かってきている。そこで本研究では、AI が扱いやすいことばの単位を、AI 自ら発見するような手法の開発に取り組んだ。

本手法は、自然言語処理におけるトークナイゼーション（Tokenization）という技術をベースとして開発した。トークナイゼーションとは、テキストを単語などの小さい単位（Token）に分割する技術である。一般的なトークナイゼーション手法では、テキストに出現する文字列の頻度の情報を用いて、テキストを細かい単位に分割する。本研究で開発した手法では、文字列の頻度の情報に加えて、AI が学習する後段のタスク（文書分類タスクや機械翻訳タスクなど）に関する情報も活用してトークナイゼーション処理を行う。これにより、後段のタスクを解くうえで AI が扱いやすいことばの単位に応じて、トークナイゼーション処理を行うことができる。

実験を通して、文書分類タスクと機械翻訳タスクにおいて、本研究で開発した手法が性能の向上に寄与することが示された。すなわち、本手法で発見したことばの単位を用いることで、後段のタスクにおける AI の性能が向上することが分かった。ここから、本手法によって AI に扱いやすいことばの単位が獲得されたことが示唆される。実験結果の分析から、AI が扱いやすいことばの単位は、入力するテキストの言語や、テキストのドメイン、後段タスクの性質によって異なることが示された。