信頼される AI の基盤技術
2020 年度採択研究者

2020 年度
年次報告書

西野　正彬

日本電信電話(株)NTT コミュニケーション科学基礎研究所
協創情報研究部言語知能研究グループ
特別研究員

誤りがないことを保証する検証器つき機械学習の研究

§1. 研究成果の概要

今年度は検証器つき機械学習モデルの研究を進めるための土台づくりとして、検証器つき機械学習モデルの検討を進めるための具体的な検証タスクの選定と、そのタスクにおける課題の洗い出しに取り組んだ。外部から検証可能な誤りは多岐にわたり、また機械学習タスクにも様々なものが存在する。そのため、検討を効率的に進めるためには具体的な検証タスクをいくつか選定したうえで、検証器つきモデルの基本的な性質を明らかにする必要があった。そこで、検討の第一ステップとして、誤りの有無が外部から決定的に判定可能であるとし、また機械学習モデルとしては入力 x，出力 y ともに離散値のベクトルであるような識別関数を学習するタスクを想定することとした。この問題設定においてどのような項目を重点的に検討する必要があるか検討し、検証の効率性および予測精度の低下に対処することが必要であるとの知見を得た。次年度は既存の機械学習手法で用いられている技術を参考に、それぞれの課題を解決する方法について検討を進める予定である。

そのほかに関連する研究課題として、機械学習モデルの事後的な検証方法についての調査・検討をすすめた。具体的には、自然言語処理の典型的な問題を解くための機械学習モデルの入力ノイズに対する頑健性を、数理計画ソルバを用いて厳密に検証する方法について検討を進めた。これまで画像を入力とした判別問題を解くための深層学習モデルにおいて、入力へのノイズ付与に対する頑健性を調べる方法は知られていたが、モデルの構造が複雑になりがちな自然言語処理向けの深層学習モデルに適用することはできなかった。本研究では自然言語処理向けの機械学習モデルの中間層にノイズを付与したときの出力の変化を、数理計画ソルバを用いて検証することで、モデルのノイズに関する頑健性を調査できるという知見を得た。