

研究課題別事後評価結果

1. 研究課題名： 説明性の高い自然言語理解ベンチマークの構築

2. 個人研究者名

菅原 朔（情報・システム研究機構 国立情報学研究所 助教）

3. 事後評価結果

人間のように文章を理解する AI システムの開発における評価指標と品質保証に着目し、それらを実現するための説明性の高いベンチマークの構築と、その持続的開発を可能にする構築の方法論（ロードマップまたはチェックリスト）の研究開発を目標とする独創的かつ波及効果の高い取組である。本研究では、以下の3つのテーマに関して研究を行なった。

(1) 持続的開発を可能にするロードマップの作成に関して、心理学における文章理解の議論に基づき、ベンチマーク評価手法が満たすべき要件を整理した。はじめに読解の定義と評価方法について理論的な基礎を整備し、次に心理測定学における妥当性議論という枠組みを参照することで、測定結果の解釈の妥当性を手続き的に向上させるための段階的手順を提案し、自然言語理解ベンチマークにおけるチェックリストの形で具現化して、当該分野のトップ国際会議の論文として提案した。(2) データセット収集の方法論に関しては、ニューヨーク大学との国際共同研究により、クラウドソーシングを用いた高品質なデータ収集の方法論を開発し、4つの収集プロセスに関して各1000件を超える問いを収集・編纂し、作成したデータセットを国際的に公開した。この結果は、2編の論文として、当該分野の最難関国際会議に採択され、発表された。(3) ベンチマークデータセットの提案に関しては、上記で実施した二つの研究項目の知見を生かして、状況に応じて結末が変わるような文脈における常識推論や、賛否のある複雑な議論における論理的推論、埋め込み的に複雑な状況にある語用論的推論などの高度な文章理解力を評価するためのタスクとデータセットを研究開発した。これらは、現在出現しつつある極めて高い性能をもつ大規模言語モデルに対しても評価可能な、高度な言語能力と言語現象を必要とする言語理解タスクを提案しており、当該分野の最難関国際会議で採択・発表されるなど、注目を集めた。

さきがけ研究における一連の研究は、多数の優れた研究論文として当該分野の複数の最難関国際会議に採択・発表され、文章を理解する AI システムの信頼性に関して国際的にも高く評価されている。データセットの公開も積極的に行なっており、将来の信頼される AI の実現のために不可欠である。その一方で、客観的評価が困難とされるシステムとデータの信頼性の評価と品質保証に関して、さきがけでの分野融合的議論にも積極的に参加しながら、言語理解の本質に迫る深い理解の下に着実に研究を遂行し、研究を進展させた。さきがけでの成果は、現在、急速に発展し、社会・経済に影響を与えつつある大規模言語モデルに関しても、今後の開発の基盤となる評価データ・手法を提供するものであり、今後のさらなる飛躍を期待する。