審 査 の 結 果 の 要 旨

氏　　名　　菅原　朔

自然言語処理分野では、自然言語の文章を人間のように理解することができる計算機システムを実現することが重要な目標とされている。しかし、何をもって「自然言語を理解した」と言えるかは自明ではない。これまでに自然言語理解能力を評価するための枠組みが多数提案されてきたが、本論文は、その中でも機械読解に着目し、その問題点と解決法を議論している。機械読解とは、自然言語の文章が入力として与えられ、自然言語の質問に答えるという自然言語処理タスクである。自然言語においては情報がさまざまな形で表現されるため、入力文章と質問文、およびその答えをどのように関係づけるかは自明でない。実際、読解問題を正しく解くためには、共参照認識、同義表現認識、時間関係認識といった様々な自然言語理解能力が必要とされる。よって、機械読解タスクにおいて高い性能を示すシステムは、自然言語理解能力を持つことが期待される。近年、大規模かつ多様な機械読解データセットが開発されており、またデータセット上でのベンチマークにおいて人間に匹敵する性能を示すシステムも報告されている。

ただし、ここには2つの問題がある。1つは、評価尺度の単純さの問題である。既存の機械読解データセットにおいては、評価尺度として精度（正答率）が用いられている。しかし、これは様々な自然言語理解の最終的な出力のみを評価するものであり、システムがどのような自然言語理解能力を用いているかを評価することができない。これではシステムが持つ自然言語理解の各能力を評価・説明することができず、自然言語理解システムの着実な研究開発に資することができない。もう1つの問題は、データセットの品質の問題である。データセット中の問い（文章と質問の組）は、人間が持つ自然言語理解能力を問うことができるように設計されると期待されるが、実際には、文字列のパターンマッチなどの単純な手法で解けてしまう問いが多数含まれている可能性がある。データセット中の多くの問いがそのような傾向があれば、特に機械学習に基づくシステムはその傾向を学習することで高精度を達成することができてしまう。これは、自然言語理解能力を評価するという本来の目的から外れており、またそのようなデータセットで学習された自然言語理解システムは実応用に利用することもできない。

本論文は、上記のような問題意識に基づき、自然言語理解能力を評価するための機械読解タスクを設計する際の問題点と解決法を議論している。本論文の貢献は以下のようにまとめられる。評価尺度の単純さの問題については、問いに正答するために必要な自然言語理解スキルを評価する手法、および文章の読みやすさ（readability）との関連

性を分析している（第3章）。データセットの品質の問題については、自然言語理解を必要としない簡単な問いとそうではない難しい問いを自動分類する手法を提案し、データセットが自然言語理解能力の必要性を評価できるかどうかを分析している（第4章）。さらにこの手法を発展させ、各問いを解くのに必要とされる自然言語理解スキルを自動評価する手法を提案している（第5章）。

本論文は以下の7章から構成されている。

第1章では、自然言語理解システムの評価方法の難しさとこれまで行われてきた試みを概観し、機械読解タスクの有用性と問題点、そして本論文の貢献をまとめている。

第2章では、機械読解の既存データセット、既存手法、およびその他の自然言語理解評価手法について詳述している。

第3章では、データセット中の各問いに正答するために必要な自然言語理解スキルのアノテーションデータを構築し、自然言語理解システムがどのスキルを持っているかを評価する手法を提案している。また、文章の読みやすさの評価尺度との関連性を分析し、文章の読みやすさとスキルの必要性は必ずしも関係しないことを示している。

第4章では、文字列のパターンマッチなどの単純な手法で答えられる簡単な問いと、自然言語理解能力を必要とする難しい問いとを自動分類する手法を提案している。既存のデータセットにおいて、データセット全体あるいは簡単な問いに対する精度に対し、難しい問いに対する精度は著しく低いことを示し、これまで報告されているベンチマークの精度は過大評価であることを示している。

第5章では、データセットにおける各問いを解くために必要とされる自然言語理解スキルを自動評価するためのアブレーションテストを設計し、既存のデータセットのベンチマークでは複雑な自然言語理解能力を評価することができないことを示している。

第6章では、読解という能力およびその検証方法に関する知見や理論を説明し、機械読解タスクや既存データセットの説明性のために必要な要件を議論している。

第7章では、本論文の結論をまとめ、将来課題について議論を行なっている。

このように、本論文は、自然言語理解能力を評価するという目的に対して本質的な課題に取り組み、既存データセットの問題点を実証・分析し、自然言語理解能力を適切に評価する方法を提案している。これらの成果は、将来の自然言語処理システムの評価手法の研究のみならず、自然言語処理システムの本質的な性能向上を目指した研究開発や、自然言語理解に対する学術的研究に対して多大な貢献を行っていると評価できる。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。