

論文内容の要旨

博士論文題目

Teaching Machines to Write from Data

(データからテキストの生成に関する研究)

氏名 磯 颯

(論文内容の要旨)

人工知能分野において、自然言語生成と呼ばれる問題は重要な研究課題の一つであり、人工知能の黎明期から盛んに研究されてきた。高い精度を持つ自然言語生成システムを構築することができれば、日常的な文書の作成業務の負担を大きく減らすことができると期待される。しかし、これまでの言語生成システムでは、流暢かつ多様なテキストを生成することは難しい。さらに、多くの場合、テンプレートを元にシステムが構築されており、そのデータベースを作成するために、多くの人手が必要となっている。

このような背景の中、この数年、ニューラルネットワークを用いたテキスト生成システムにより、人の書いたテキストと見分けのつかないほど流暢なテキストを生成することが可能になってきた。一方で、このニューラルネットワークを用いたモデルは、事実と異なるテキストを生成してしまうなど、長文生成が困難であるといった新たな問題を提示している。本博士論文では、Data-to-Text 生成とテキスト編集の観点から、この問題に対するアプローチを提案した。

まず、Data-to-Text 生成システムとして、スポーツの試合のスコアボードのような入力データに対して、事実に忠実なテキストを文書レベルで生成するための手法について検討した。具体的には、入力データ中の実体ごとにモデル化し、生成の過程でその状態を更新しながら生成をしていくモデルを研究開発した。

次に、生成に関連する新たな問題として、事実を元にしたテキスト編集タスク Fact-based Text Editing を提案し、この方式の精度を検証した。Fact-based Text Editing は、既存のテキスト編集タスクとは異なり、外部の知識に基づき既存のテキストから情報の追加、削除、更新を行う。本研究では、Fact-based Text Editing を達成するためのデータセットを自動構築する手法を導入し、さらに Fact-based Text Editing を正確かつ効率的で、結果の可解釈性が高いモデルを研究開発した。

氏名	磯 颯
----	-----

(論文審査結果の要旨)

本論文は、自然言語生成に関する新手法についての研究である。自然言語生成を高い精度で実現するシステムを構築することができれば、日常的な文書の作成業務の負担を大きく減らすことができる。しかし、文書を生成するということは人間の高度な知性が必要な活動であり、自然言語処理最後の大きな課題とされている。

このような背景の中、人間の書いたテキストと見分けのつかないほど流暢なテキストを生成することができるとして、ニューラルネットワークを用いた自然言語生成システムが大きな注目を集めている。一方で、事実と異なる内容を含んだ文を生成してしまうなど、長文生成が困難であるなど新たな問題を提示している。そこで本論文では、与えられたデータをもとにテキストを生成する課題である Data-to-Text 生成に対して、テキスト編集のアプローチを用いて改善を行う。

まず、Data-to-Text 生成システムとして、入力データに忠実なテキストを文書

レベルで生成するための手法について検討する。本手法では、入力データ中の実体ごとにモデル化し、生成の過程でその状態を更新していきながら生成をしていく。対象としている NBA のバスケットボールの試合のテキスト生成では、選手を実体として、選手毎にテキストを生成している。これまでのニューラル・テキスト生成がコントロール不可能でブラックボックス化していたことに対する大きな改善であり、高い有用性を持つ。

次に、事実を参照しながらテキスト編集し生成結果を改善する方式である Fact-based Text Editing を提案し、この方式の精度の検証を行った。Fact-based Text Editing では、既存のテキスト編集タスクとは異なり、外部の知識に基づき既存のテキストから情報の追加、削除、更新を行うものである。

両手法は、これまで、ニューラル・テキスト生成が流暢であるものの内容の正確さに関しては問題があることに対する解答であり、広くはブラックボックス化した深層学習を現状の技術で制御する試みであると言える。緻密ながら、現在の学習理論全般に通じる有用性の高い内容である。よって、博士(工学)の学位論文としての価値があるものと認める