共同企画
# 共同企画7
# 医学医療における AI応用
2021年11月20日(土) 09:10 〜 11:10 B会場 (3号館3階国際会議場)

[3-B-1-02] 自然言語生成の現状と展望

*岡崎 直観[1] （1. 東京工業大学 情報理工学院 情報工学系）
*Naoaki Okazaki[1] （1. School of Computing, Department of Computer Science, Tokyo Institute of Technology）
キーワード：Natural Language Generation, Language Models, Deep Neural Networks

自然言語処理とは、言葉を理解・生成するコンピュータの実現を目指す研究分野であり、自然言語生成はその後段を担うものである。前段の自然言語理解では、言葉から理解した情報を人間に提示することで、文書分類や知識獲得など、様々なアプリケーションが実現されていた。ところが、自然言語生成の伝統的なアプローチはルールに基づくものが多く、株価やスポーツの結果を表す定型的な文章の生成か、機械学習に基づく手法が発展を遂げた機械翻訳などにアプリケーションが限定されてきた。ところが、2010年代に深層学習が自然言語処理に導入されると、その様相が一変する。2014年頃にニューラル機械翻訳モデルが出現し、2017年頃のTransformerモデルにより、40年以上の研究の蓄積があった従来手法を超える性能を達成した。また、2015年頃から抽象型（生成型）自動要約や画像からの文章生成、表などのデータからの生成の研究が盛んに行われるようになり、自然言語生成のアプリケーションが広がりを見せた。さらに、2018年頃から発展が続いている GPTなどの大規模言語モデルの出現により、一見すると自然に見える架空の新聞記事の自動生成や、プログラムの自動生成など、驚くようなアプリケーションが生まれつつある。
ところが、現状の言語生成モデルは大量の学習データに依存しており、学習データがあまり手に入らない自然言語生成タスクを実行させることは容易ではない。従って、対象とするタスクの学習データが少なくても、自然言語生成モデルの出力を柔軟に制御できる手法の研究を進めていく必要がある。本講演では、自然言語生成モデルの発展を概観したのち、東京工業大学情報理工学院・岡崎研究室で進められている研究として、生成文の長さの制御、忠実性の制御、内容の直接的な制御等の研究を紹介し、今後の研究の展望を述べる。

自然言語生成の現状と展望

岡崎　直観[*1]
*1 東京工業大学情報理工学院

Recent Advances and Prospects of Natural Language Generation

Naoaki Okazaki[*1]
*1 School of Computing, Tokyo Institute of Technology

Natural Language Processing (NLP) is a research field aiming to realize computers that can understand and generate text. Natural language understanding has established various applications such document classification and knowledge acquisition. Natural language generation (NLG) is mostly based on either rules or supervised learning. The rule-based approach uses templates for converting data into text, which limits their applications to some specific domains. Supervised learning had successfully been applied to machine translation, where large supervision data have been developed. In the 2010s, deep neural networks made breakthrough in various NLP tasks. Benchmark scores of machine translation were nearly doubled from the former approach (statistical machine translation) with the advances of sequence-to-sequence models, attention mechanism, and the Transformer architecture. This also broadened NLG applications into, for example, abstractive summarization and image caption generation. Furthermore, large-scale language models such as GPT demonstrate new applications such as story generation and program generation. However, the current language generation models rely on a large amount of supervision data. Applying these models to NLG tasks is not straightforward when sufficient training data is not readily available for a target task and domain. Therefore, it is necessary to research methods that can utilize supervision data for different tasks and flexibly control the output of NLG models. In this manuscript, I will review the recent advances of NLP and the foundations of NLG models. I will then describe the research on controlling NLG models and future prospects of NLG.

Keywords: natural language generation, language models, deep neural networks.

1. 自然言語生成の現状

自然言語処理とは、言葉を理解・生成するコンピュータの実現を目指す研究分野であり、自然言語生成はその後段を担うものである。前段の自然言語理解では、言葉から理解した情報を人間に提示することで、文書分類や知識獲得など、様々なアプリケーションが実現されていた。ところが、自然言語生成の伝統的なアプローチはルールに基づくものが多く、天気概況やスポーツの試合結果などの定型的な文章の生成か、訓練データに基づく手法が発展を遂げた機械翻訳などにアプリケーションが限定されてきた。

ここで、テンプレートに基づいてデータから収益報告書を生成する例を紹介する。[1] テンプレートは、データの特定の値を参照する変数、生成内容を制御する条件式（例えば、収支が黒字と赤字の場合で生成内容を変更できる）、および条件と変数から生成される記事内容が記述される。テンプレートの例を以下に示す。

```
The {{ fund_name }} { posted grains | reported a loss } of {{ cm_ret }}% in {{ month }}, compared to last month's return of {{ pm_ret }}%.
```

ここで、{{ }}はデータへの（変数による）参照、{ | }は条件により生成内容を変える部分である。この例では、収益が黒字ならば「posted gains」が、赤字ならば「reported a loss」が生成される。テンプレートに基づく生成は、生成の挙動が透明であるため生成される文章の質を担保しやすいが、テンプレートで定義されている内容しか生成できないため、異なる種類のデータに適用することができない、生成される文章が単調になる、等の問題がある。

一方、機械翻訳の研究は自然言語処理の黎明期から行われ、1990 年代初めに考案された統計的機械翻訳により、訓練データに基づくアプローチが主流となっていた。日英翻訳では、「私は東京に行った」と「I went to Tokyo」のように、2言語間で翻訳の関係にある文の組を訓練データとして、単語間の翻訳関係を表す翻訳モデルと、翻訳先言語における文の自然さを測定する言語モデルを自動的に学習する。翻訳したい文が与えられると、翻訳モデルと言語モデルの両方を考慮した翻訳文が生成される。

2010 年代に深層ニューラルネットワークが自然言語処理に導入されると、機械翻訳の手法にいくつかのブレークスルーが起こった。まず、2013 年に word2vec[2] が出現し、単語を固定長のベクトルで表現するという考え方が急速に支持を得た。その後、系列変換モデル（encoder-decoder models）[3] により機械翻訳を実現するアプローチが盛んに研究されるようになり、その性能向上のために注意機構 [4] が提案された。さらに、2017 年に提案された Transformer モデル [5] は、機械翻訳の性能を大きく改善しただけでなく、その後 GPT [6] や BERT [7] などの大規模言語モデルの基盤アーキテクチャとなった。

機械翻訳モデルの発展に伴い、自然言語生成の研究も活発化した。それまでは実現が難しかった抽象型（生成型）の自動要約は実用レベルに到達した。また、画像やデータから説明文を生成する研究も大きな進展を見せ、自然言語生成のアプリケーションが広がりを見せた。さらに、2018 年頃に出現したGPTやBERTなどの大規模言語モデルにより、一見すると自然に思える架空の新聞記事を生成したり、処理内容からプログラムを自動生成するなど、自然言語生成の驚くようなアプリケーションが登場している。

図1　長さを指定した見出し生成のデモ

ところが、現状の言語生成モデルは大量の学習データ、すなわち入力に対する出力の「お手本」に依存している。学習データから逸脱したタスクを解いたり、学習データがあまり手に入らないタスクで自然言語生成を行うのは容易ではない。したがって、対象とするタスクの学習データが少ない場合でも、別のタスクにおける大量の学習データから構築した言語生成モデルを適用したり、その出力を柔軟に制御できる手法の研究が注目を集めている。

2. 自然言語生成の制御

自然言語生成において、生成されるテキストの内容や長さをコントロールしたい。簡単なことのように思えるかもしれないが、一つの大きなニューラルネットワークとして構成される言語生成モデルはブラックボックスであり、挙動を解析できないこと、言語生成モデルは訓練データに与えた入出力を再現するようにしか学習していないことから、言語生成モデルを制御する方法は自明ではない。

高瀬ら[8]は、Transformerの位置エンコーディングを改良し、言語生成モデルが出力するテキストの長さを自由に制御する手法を提案した。提案手法は出力長を連続的に表現するため、訓練データには現れていない長さの出力も可能である。要約生成タスクの一つである、見出し文生成において長さ制御の実験を行い、提案手法が出力の長さを制御できるだけでなく、品質（ROUGE スコア）も向上させることを示した。また、新聞社の記事制作現場で実際に利用してもらったところ、人手による修正の必要がない見出しが生成されるか、生成された見出しを活用しながら見出しを制作できる事例の割合が80%近くと、高評価を得た。本手法をシステムとして実装したものを、ウェブサイト上で公開している（図1）。

松丸ら[9]は、見出し生成の内容の制御として、忠実性（生成された見出しが伝えるすべての事柄が元記事に基づいているか）を改善する研究を進めた。見出し生成において忠実性が損なわれていると、誤報という深刻な問題を引き起こしてしまう。記事内容から逸脱した見出しが生成される原因は、見出し生成のタスク設定の不備や訓練データにあるという仮説を検証・確認した。続いて、訓練データをクリーニングしたうえで、自己学習に基づく訓練データ拡張を行うことで、見出し生成の品質と忠実性が改善されることを実証した。

Yangら[10]は、新聞記事中の写真の見出し生成において、記事に基づく内容の制御が有効であることを示した。画像からのキャプション生成の研究では、画像を入力、テキストを出力とした機械翻訳モデルを構築することが多い。ところが、画像だけから写っている物体の詳細（人物名や場所名など）を抽出することは難しい。そこで、画像とテキストの両方を入力とする新しい Transformer アーキテクチャを提案し、画像から得られた特徴量に加えて、新聞記事中から得られた固有表現（名前）を用い、見出しの内容を制御した。

丹羽ら[11]は、キャッチコピーの自動生成において、対句による修辞技法を考慮できるモデルを提案した。見出し生成とは異なり、対句生成のための訓練データを大量に確保することは困難である。そこで、丹羽らは事前学習済みの大規模言語モデル BERT を土台として、文脈付きの対義語関係を追加学習する手法を提案した。自動および人手による評価において、提案手法が対義語を高精度に予測できることを示した。

3. おわりに

深層ニューラルネットワークの発展により、自然言語生成の精度が向上しただけでなく、新しい応用が次々と開拓された。ところが、現状の言語生成モデルの成功の鍵は、大量の訓練データにかかっている。このため、言語生成モデルを制御して所望の出力を得たり、大量の訓練データで学習した知識を少量の学習データしか入手できないタスクやドメインへ転用する研究が進められている。これらのアプローチを大局的に見ると、大規模な言語生成モデルを構築し、そのモデルにタスクやドメインに関する汎用性を持たせる試みである。これまでのアプローチでは人間が一生に読む文書の量を遙かに上回る量の訓練データを使っているが、このアプローチの延長線上でうまくいくのか、それとも全く異なるアプローチが必要となるのか、研究者の間でも見通しが立っていないのが現状である。ただ、世の中のありとあらゆるタスクやドメインにおいて、大量の訓練データが得られると仮定するのは現実的ではない。言語生成に汎用性を持たせることは、今後も重要なテーマとして研究が進められていくであろう。

参考文献

1) 岡崎 直観. ロボットジャーナリズムの現状と課題. 映像情報メディア学会誌, 2018; 72(2):70-75.
2) Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In NIPS 2013, 3111-3119.
3) Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In NIPS 2014, 3104-3112.
4) Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In ICLR 2015.
5) Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I. Attention is all you need. In NIPS 2017, 5998-6008.
6) Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Technical report, 2018.
7) Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In NAACL 2018, 4171-4186.
8) Takase S, Okazaki N. Positional encoding to control output sequence length. In NAACL-HLT 2019, 3999-4004.
9) Matsumaru M, Takase S, Okazaki N. Improving truthfulness of headline generation. In ACL 2020, 1335-1346.
10) Yang Z, Okazaki N. Image caption generation for news articles. In Coling 2020, 1941-1951.
11) Niwa A, Nishiguchi K, Okazaki N. Predicting antonyms in context using BERT. In INLG 2021, (to appear).