深層学習による言語生成モデルの制御を目指して

研究ノート

荒　瀬　由　紀*

Towards Control of Neural Language Generation

Key Words：machine translation, dialogue system, text simplification, language generation, decoder constraint

はじめに

自動運転やロボット制御など、深層学習によって人工知能技術は目覚ましい発展を遂げた。人間が書いた文章をコンピュータによって知的かつ高度に処理する技術の実現に取り組む自然言語処理においても、深層学習は大きなブレークスルーをもたらした。コンピュータによる自動翻訳を行う機械翻訳では、特定言語対における機械翻訳の性能が人間による翻訳と遜色ない品質であるという結果が発表され[1]、所与の文章に基づき質問応答を行う機械読解では、深層学習を用いたモデルが人間の正答率を超える事例[2]も多数報告されている。

自然言語処理における大きなゴールの一つに、人間と同様に流暢な文章をコンピュータで生成することを目指す言語生成がある。深層学習による言語生成は、エンコーダ・デコーダモデル[3]を用いるのが一般的である（図1）。エンコーダは「文」をコンピュータで演算できるよう、入力文に含まれる各単語を再帰的に読み込み、高次元ベクトルに変換する役割を担う。デコーダではエンコーダから受け取った入力文のベクトルと一つ前に自身が出力した単語のベクトルに基づき、次の単語を出力する。この処理を繰り返すことで文を生成する。エンコーダ、デコーダそれぞれがニューラルネットワークで構成されており、大規模な学習データを用いて訓練することで、複雑なルールを記述することなく流暢な言語生成が可能となる。

またエンコーダ・デコーダモデルは非常に汎用性の高い機構となっており、英語・日本語の対訳文対を用いて訓練すれば機械翻訳器に、人間の発話・応答対を用いて訓練すれば対話システムに、難しい文・平易な文の対を用いて訓練すれば文の難易度変換器となる。

エンコーダ・デコーダモデルは訓練データから言語生成に必要なパターンを学習するのが最大の利点である一方、これまで蓄積されてきた知識、例えば対訳辞書や単語の難易度辞書、を適用できない。この問題を解決するため、我々の研究グループでは、エンコーダ・デコーダモデルに対し知識を活用したデコーダの制御手法の研究を行っている。

知識に基づくデコーダの制御

デコーダでは各単語を出力する適切さを測るスコアを計算し、最もスコアの高い単語を出力する処理を繰り返している。そして、訓練データとして提供される正解の文（参照文）と比較し、参照文に一致した単語の出力スコアが大きいほど高く評価する損失関数により、モデル全体の訓練が進んでいく（図1）。そこで我々は、デコーダ制御の鍵となるのはこれら出力スコアの計算と損失関数と仮定し、単語報酬モデルと損失関数による出力制御モデルの二つの手法を開発した。

単語報酬モデルは図2に示す通り、デコーダで計算する単語の出力スコア（オレンジ色の棒グラフ）に対しバイアス（黄色の棒グラフ）を加えることで、知識に合致する単語の出力を促進する。図2の例では、対訳辞書を用いて入力文にある「favorite」の翻訳である「好き」や「大好物」のスコアを底上げし、出力しやすくしている。このように単語報酬モデル

* Yuki ARASE
大阪大学大学院情報科学研究科　博士後期課程（2010年）
現在、大阪大学大学院情報科学研究科
准教授　博士（情報科学）
TEL：06-6879-7752
E-mail：arase@ist.osaka-u.ac.jp

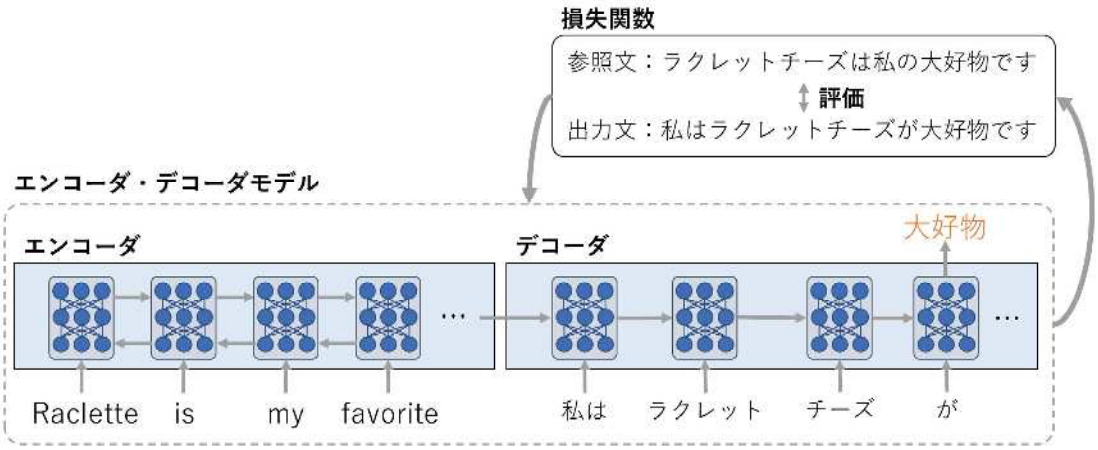

図１　エンコーダ・デコーダモデル

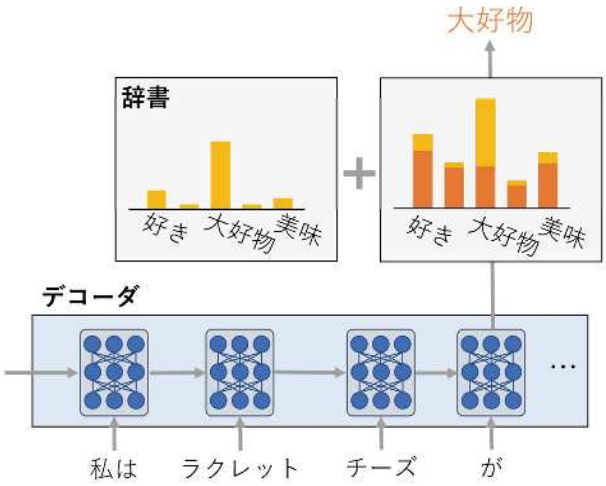

図２　単語報酬モデル

は、ある単語の出力スコアに直接作用するため、モデルの出力を変化させやすい。対訳辞書のように、ある出力で用いるべき知識を規定した辞書が利用できる場合に大きな効果を発揮する手法である。

損失関数による出力制御モデルは、モデル全体の訓練を統括する損失関数に着目した手法である。前述の通り、損失関数は参照文に合致した単語の出力スコアが大きいほど高い評価をモデルに与え、参照文により近い出力を行うようモデルを訓練する。この損失関数を、参照文に加え知識に合致するものの出力スコアが大きいほどより高く評価するよう改良する。例えば文の難易度変換では、大学生レベルの英文を中学２年生レベルの英文に変換する際に、中学２年生レベルの語彙の出力スコアが大きいほど高い評価を与える。これにより、難易度変換器は中学２年生レベルの語彙を使った出力を優先する能力を獲得する。この出力制御モデルは、文の難易度変換のようにある学年の文で使用されるべき語彙が決まっているなど、同じ属性をもつグループが満たすべき制約に関する知識が利用可能な場合に有効な手法である。

デコーダ制御手法の適用

単語報酬モデルを英日・日英機械翻訳に適用した[4]。エンコーダ・デコーダモデルに対し単語報酬モデルにより対訳辞書を適用することで、翻訳性能が大きく改善することが示された。表１に示す翻訳例では、エンコーダ・デコーダモデルが翻訳できていない「癌」「先天免疫」を正しく翻訳できている。

表１単語報酬モデルによる日英翻訳の例

原文	IL-12の癌に対する抵抗性（先天免疫）の生物反応についても考察した
ベースライン	The biological response of the resistance to IL-12 is also discussed.
単語報酬	The biological response of the resistance (congenital immunity) to the cancer of IL-12 is also discussed.

次に単語の難易度辞書を用いて、損失関数による出力制御モデルを英文の難易度変換器に適用した[5]。難易度変換問題は機械翻訳と異なり、入力文を省略する変換が頻繁に起こる一方、入力文を書き換える箇所は文全体の一部分である。そのためエンコーダ・デコーダモデルでは、訓練データで頻繁に起こる省略は学習しやすいが、必要な書き換えを行わないと

いう問題が指摘されている。実験の結果、損失関数による出力制御モデルを用いることで、エンコーダ・デコーダモデルに比較して難易度変換の品質が向上するのに加え、難易度を変換するための英文の書き換えを促進できることが確認された。表2に高校3年生レベルの英文を小学4年生レベルに変換した例を示す。ソフトな出力制御モデルでは、文全体を簡略化するための副詞節・並列構造の省略に加え、難単語である「motivate」をより簡単な「inspire」に書き換えている。

表2　損失関数による出力制御モデルを用いた英文の難易度変換例

高3	入力文	In its original incarnation during the 60s, African-American "freedom songs" aimed to motivate protesters to march into harm's way and , on a broader scale, spread news of the struggle to a mainstream audience.
小4	ベースライン	In the 1960s, African-American "freedom songs are aimed to motivate protesters to march into harm's way.
	ソフトな出力制御	African-American "freedom songs are aimed to inspire protesters to march into harm's way.

最後に、単語報酬モデルと損失関数による出力制御モデルの両方をデコーダ制御として雑談対話システムに適用する[6]。雑談応答のように入力文に対して可能な応答が多様な問題にエンコーダ・デコーダモデルを用いると、「そうですね」や「わかる」のように、可能な中で最も無難で高頻度な応答を生成してしまう問題が起こることが広く知られている。そこでデコーダ制御を用いることで、発話中の単語に関連した語彙の出力を促進し、話題にあった応答を生成する。訓練データ中の発話・応答に現れる単語対の共起頻度に基づいて共起語辞書を作成し、デコーダ制御に適用した。出力例を表3に示す。ベースラインであるエンコーダ・デコーダモデルでは無難であるが面白みのない応答を生成している。一方、デコーダ制御を導入したモデルは「江ノ島」に対し「八景島」、「Android」に対して「iPhone」のように、ユーザ発話中の話題に適した応答を生成できることが分かる。

表3　デコーダ制御を用いた雑談対話システムの出力例

ユーザ発話	江ノ島におるん？
ベースライン	おるで
デコーダ制御	八景島です
ユーザ発話	Android に変えよう
ベースライン	え、まじで？
デコーダ制御	iPhone の方がいい

おわりに

深層学習により言語生成技術は大きく進展し、一見しただけでは人間が書いた文と区別がつかないような文の生成も可能となりつつある。しかし自然言語処理技術の社会実装を進めていくにはまだまだ課題が多い。特に深層学習モデルがブラックボックスであるため、ある出力がなされるメカニズムが未解明であること、また出力の制御が困難であることが大きな問題意識となっている。出力の制御は知識の活用という利点に加え、開発者が意図しない害のある出力を防ぐ上でも重要である。

本稿では単語に基づくデコーダの制御に関する研究を紹介したが、今後は句や文、文章など、より広い文脈を考慮した出力制御について研究を進めていきたい。

参考文献

[1] Hany Hassan et al. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. arXiv:1803.05567. https://arxiv.org/abs/1803.05567

[2] SQuAD2.0. https://rajpurkar.github.io/SQuAD-explorer/

[3] Ilya Sutskever, Oriol Vinyals and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. in Proc. of Neural Information Processing Systems, pp. 3104-3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[4] 竹林 佑斗，Chenhui Chu，荒瀬 由紀，永田 昌明．2019．ニューラル機械翻訳における単語報酬モデルに基づく対訳辞書の利用．自然言語処理 Vol. 26, No. 4, pp. 711-731. https://doi.org/10.5715/jnlp.26.711

[5] 西原 大貴, 梶原 智之, 荒瀬 由紀. テキスト平易化における語彙制約に基づく難易度制御. 自然言語処理 Vol. 27, No. 2 (2020年6月 掲載予定).

[6] Junya Takayama and Yuki Arase. 2019. Relevant and Informative Response Generation using Pointwise Mutual Information. in Proc. of Workshop on NLP for Conversational AI. pp. 133-138. https://www.aclweb.org/anthology/W19-4115/