ニューラル機械翻訳の進展
―深層学習による系列変換モデルの進化と様々な応用―

Recent Advances in Neural Machine Translation

奈良先端科学技術大学院大学　先端科学技術研究科情報科学領域／
データ駆動型サイエンス創造センター　准教授
須藤　克仁

2002 年京都大学大学院情報学研究科修士課程修了。同年 NTT 入社、音声言語処理・機械翻訳の研究に従事。2015 年京都大学博士（情報学）。2017 年より現職。機械翻訳を中心に自然言語処理・音声言語処理の研究に従事。AAMT/Japio 特許翻訳研究会　副委員長。

sudoh@is.naist.jp

1 はじめに

特許情報処理に機械翻訳が活用されるようになって久しいが、その間に発展を続けてきた機械翻訳技術が近年の深層学習技術により急速に進展し、従前の機械翻訳を大きく上回る翻訳精度を発揮するようになったと言われている。深層学習技術がニューラルネットワーク（NN；Neural Network）によるものであることから、この機械翻訳はニューラル機械翻訳（NMT；Neural Machine Translation）と呼ばれている。これに先だって広く用いられてきた統計的機械翻訳（SMT；Statistical Machine Translation）とは大きく方式が異なるものの、対訳文のデータ（対訳コーパス）を利用し、機械学習に基づいて機械翻訳を行うという面では NMT も統計的手法と言える。その他の深層学習関連技術と同様に、その発展を支えたのは（1）膨大なデータの蓄積、（2）GPU 等の高速並列計算デバイスの進化とコモディティ化、（3）機械学習モデルの発展、（4）プログラム実装用のソフトウェアフレームワークの普及、である。本稿では主に（3）の面に着目し、機械翻訳のために考案された種々の機械学習モデルを紹介するとともに、NMT の補助技術についても紹介する。

2 ニューラル機械翻訳の諸モデル

深層学習技術の機械翻訳への応用は統計的機械翻訳のモデルの一部を NN に置き換えることから始まった。2014 年に SMT と全く異なる考え方に基づく手法が提案され、急速な発展を経て NMT という呼称が一般化するに至った。本節ではその発展の過程で提案された代表的な手法をいくつか紹介する。

2.1　系列変換モデル

NMT の基本となったモデルは 2014 年に発表された系列変換（sequence-to-sequence；seq2seq とも）モデルである[1]。従来の SMT がまず原文中の語句等の翻訳候補を列挙しその並べ替えによって訳文を構成する方式であったのに対し、系列変換モデルはまず原文を NN に順次入力して NN の内部状態ベクトルとして「記憶」させ、その後訳文をその記憶から順次「合成」するという方式である。これにより、SMT で行われていた翻訳候補の辞書を複数の処理を経て学習し多数のモデルを重み付きで混合するという複雑な手順は、NMT では「記憶」と「合成」のための二種類のモデルを一括で学習する比較的単純な手順に置き換えられた。「記憶」する箇所は入力文をベクトルに符号化する役割を担うことから「エンコーダ」、「合成」する箇所はベクトルから出力文へ復号する役割を担うことから「デコーダ」と呼ばれ、この二つを有することからこの系列変換モデルはエンコーダ・デコーダモデルと呼ばれることもある。

このモデルで用いられたのは回帰型 NN（RNN；Recurrent Neural Network）と呼ばれる種類の NN であり、その中で特に LSTM（Long Short-Term Memory）と呼ばれるものを利用することで学習の安定化を図っている。LSTM 自体は 1997 年に提案されたものであるが、データ量と計算能力の恩恵によって脚

光を浴びることになった。

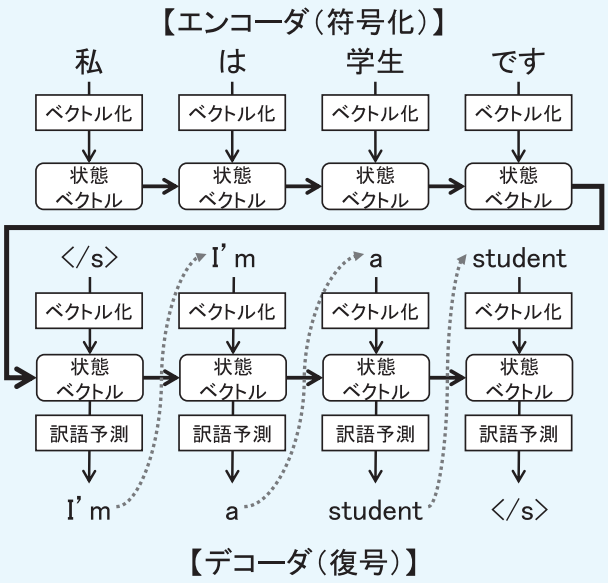

図１　系列変換モデル [1] の略図

このモデルの略図を図１に示す。各入力単語はモデルにより実数値ベクトルに変換され、エンコーダのRNNに入力される。入力が完了するとデコーダに状態ベクトルが渡され、出力を開始するための記号（図では系列末尾を表す特殊記号の </s>）を入力すると、訳文の先頭の単語が予測され、出力として得られる。RNNは入力に応じて状態ベクトルが更新され次の出力が得られるので、次の訳語を得るためにその直前の出力単語を入力する。この仕組みを自己回帰（auto-regression）と呼ぶ。この手順は系列末尾が予測されるまで続けられ、最終的に単語列としての翻訳結果が得られる。

しかし、この初期の系列変換モデルは単独では当時のSMTの性能には及ばず、初期値の異なる複数のモデルの多数決でSMTと同等程度の精度を達成するに留まった。一つの大きな理由として、入力文の「記憶」を一つのベクトルに圧縮して持たざるを得ないために、文が長くなると翻訳精度が大きく低下することが挙げられる。NMTでは500次元、1000次元、という高次元のベクトルが用いられるが、訳文を合成するための情報としては十分とは言いがたい。

2.2　注視機構つき系列変換モデル

系列変換モデルの提案からほどなくして、上記の問題の解決を図る方法である「注視機構」（Attention Mechanism）が提案された[2]。基本的な考え方としては、エンコーダが入力文をすべて記憶した最終状態のみを用いるのではなく、入力の各段階の履歴を残しておき、それをデコーダが適宜参照できるようにするのが注視機構である。

注視機構を持つ系列変換モデルでは、訳文の合成における訳語の選択時に、エンコーダの途中状態を重み付きで参照する。具体的には、ベクトルで表現されたエンコーダの途中状態とデコーダの状態との関係（複数の定義あり. 代表的なものの一つは内積≒ベクトルの類似度）に基づいて、和が１になるような重み値を算出し、エンコーダの途中状態ベクトルの重み付け和として計算される「文脈ベクトル」をその時点での訳語選択に利用する。

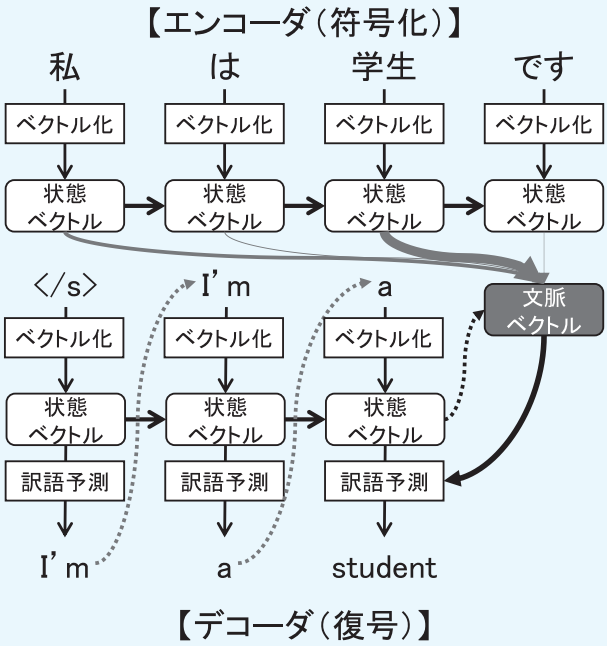

図２　系注視機構付き系列変換モデル [2] の略図（エンコーダ最終状態のデコーダへの受け渡しは省略）

このモデルの略図を図２に示す。図では３番目の訳語である“student”の予測時点の状態を示しており、文脈ベクトルに至る灰色の矢印の太さが注視重みを表しているとする。訳語選択の観点では日本語側の「学生」の情報の寄与分が大きいと予想されるが、訳語が単数形であることの判断材料として「私」の情報も寄与しうることをこの図は表している。なお、注視機構付き系列変換モデルでは、文前方の状態ベクトルに文後方の単語の情報が反映されるように、文先頭から末尾に向かっての符号化と文末尾から先頭に向かっての符号化を行いその結果を結合して状態ベクトルとする双方向符号化（bi-directional encoding）がしばしば行われるが、簡単のためこの図では省略している。

このようにして文脈ベクトルが各訳語の選択に必要な

情報を柔軟に提供することにより、訳語選択の正確性が向上するとともに、前述の長い文における翻訳精度の低下が抑制された。この技術により NMT は SMT を上回る性能を発揮し始め、機械翻訳の技術が NMT に傾倒していくこととなった[3]。

2.3 Transformer

RNN に基づく系列変換モデルは精度の面で大きな進展を遂げたが、入力を一つずつ読み込んで状態ベクトルを更新する RNN の仕組みゆえに入力文長に比例する計算ステップ数が必要になり、特に文が長くなると計算の並列化・効率化が難しいという問題がある。

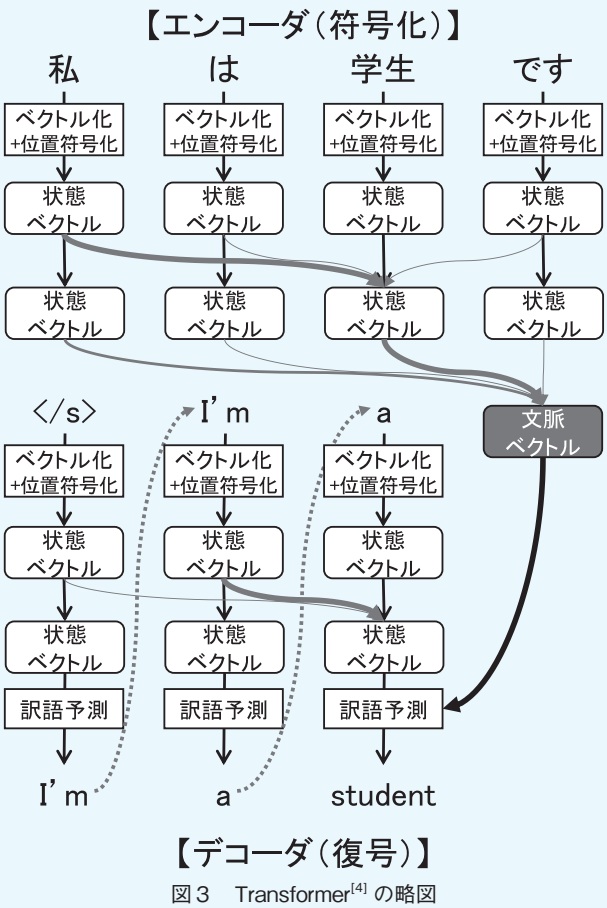

図３　Transformer[4] の略図

そうした流れを受けて 2017 年に提案されたのが Transformer と呼ばれるモデルである[4]。このモデルの略図を図３に示す。Transformer は RNN を排し、多層化したフィードフォワード型 NN と注視機構を活用して計算の並列化を可能にした。Transformer の特徴は自己注視（self-attention）と呼ばれる注視機構であり、前述のデコーダによるエンコーダの情報の注視（交差注視またはエンコーダ・デコーダ注視）と異なり、エンコーダにおいては原文中の、デコーダにおいては訳文中の他の要素に対して注視を行うものである。自己注視は原文・訳文の情報の抽象化・構造化を担っており、交差注視はエンコーダからデコーダへの情報伝達を担っている。図ではエンコーダとデコーダの自己注視およびエンコーダ・デコーダ間の交差注視を灰色の矢印で示している。

自己注視によって RNN のように原文や訳文を先頭から順に処理する必要がなくなり[1] 計算の並列化が可能になる一方で、そのままでは文における語順の情報が失われてしまうため、Transformer では位置符号化（positional encoding）という仕組みが導入された。詳細は割愛するが、それぞれの要素が文中の何番目の位置にあるか表すベクトルを加算することで、位置の情報をモデルに与えるようになっている。

さらにもう一つの特徴が注視機構を多重化する複数ヘッド注視（multi-head attention）である。交差注視も自己注視も他の要素の情報を重み付きで参照するが、自然言語には例えば時制や性・数による活用の違いのように複数の要因が存在し、一つの重み付き和ですべてを集約することは容易でない。複数ヘッド注視は複数の独立した注視を行いその結果を統合することで、そのような複数の要因を考慮することを狙ったものである。

こうした様々な工夫により、Transformer は大きな翻訳精度向上を達成し、現在に至るまで NMT のデファクトスタンダードの位置を占めていると言ってよい。また、Transformer を利用した大規模な自然言語符号化モデルである BERT は様々な自然言語処理タスクでの有効性が示されており、機械翻訳の研究が自然言語処理全体に波及していった好例と言える。

2.4 非自己回帰型ニューラル機械翻訳

Transformer は RNN による系列の逐次処理を排したモデルだが、翻訳実行時には先頭から順に訳語選択を行い、その際には自己回帰によりその直前までの訳語選択結果を利用する。自己回帰は機械翻訳のような出力系列の予測問題においては自然な方式であると考えられる一方、ある時点での予測を誤るとその後の予測に悪影響を与える可能性が高いこと、そして予測時の計算が並列

1　デコーダは訳文を先頭から順に生成する関係上、後方の要素への注視は行わないように設計される。

化できないことが問題となる。

このような問題に対して考案されたのが非自己回帰型ニューラル機械翻訳（non-autoregressive NMT）である[5]。非自己回帰型 NMT では、自己回帰により訳文を先頭から順に生成するという仕組みを排し、訳語をすべて並列に生成することを基本的な考え方とする。単純に並列に訳語予測をするだけでは訳語間の制約が効きづらく、同じ単語が連続して出てきてしまうといった問題が生じるため、様々な工夫が行われている。

非自己回帰型 NMT は精度面ではまだ Transformer に及ばないが、従来と考え方を大きく変えたモデルであり、機械翻訳の高速化の面でも注目を集めている。特に直近 1 - 2 年で多数の研究報告が行われており[6, 7, 8]、今後も注目すべき方式と言ってよい。

3 ニューラル機械翻訳の補助技術

NMT の進展は前節に述べたようなモデルの進化による部分が大きいことは間違いないが、NMT を使いやすく高精度なものにするために、SMT とは異なる様々な工夫が行われてきた。

3.1 サブワードの利用

SMT の訳語選択が辞書的な情報に基づいて数十～数百の候補から行われていたのと異なり、NMT の訳語選択は目的言語の語彙に含まれるすべての語（数万～数十万）を考慮する必要があり、計算の効率だけでなく精度の面でも大きな問題であった。初期の研究では語彙サイズを最大でも５万語程度に制限し、頻出する語のみが翻訳可能な形であったため、語彙への未登録語の翻訳に問題が生じていた。

サブワードとは単語より短い部分文字列を指す。例えば 単語 internationalization を inter/national/iza/tion のような短い単位に分割して扱えるとすると、接尾辞や接頭辞等の共通した要素が含まれており、再利用性が高まることが予想できる。ただこのような情報を言語ごとに整備するのは容易でないため、自動的にサブワードの単位を定める方法が提案され、昨今の NMT において広く用いられている。代表的なものは 2016 年に提案されたバイトペア符号化（BPE; Byte-Pair Encoding）に基づく方法[9] であり、情報圧縮の技術である BPE を利用して、テキストデータから決められたサイズのサブワード語彙を作成可能である。もう一つの著名な方法として SentencePiece がある[10]。SentencePiece はテキストデータを事前に単語分割する必要がないことから、日本語データで特に広く用いられている。

3.2 逆翻訳による学習データ拡張

SMT や NMT 等大量の対訳データを利用する機械翻訳においては十分なデータ量が得られるかどうかが性能に直結する。しかし一般に質の良い対訳データの収集は困難であり、訳語選択の精度や翻訳結果の流暢性に問題が生じることが多い。SMT では目的言語のみのテキストデータを「言語モデル」として翻訳結果の流暢性向上に役立てていたが、NMT はそうした構成は一般的ではない。

「逆翻訳」（back-translation）は文字通り目的言語から原言語への逆方向の翻訳を表す用語である。逆翻訳自体は目的言語を解さない原言語話者が機械翻訳の正確さを推し量るために目的言語への翻訳結果を再度原言語に翻訳させその内容を見る、という用法で知られてきたが、NMT ではこの逆翻訳を目的言語のみのテキストデータとそのテキストデータを入力として NMT で作成した原言語への翻訳結果を擬似的な対訳データとして利用することが広く試みられている[11]。これは逆翻訳を行う NMT がある程度の精度を有する場合には比較的有効であることが知られている。その理由として、目的言語のテキストデータは通常自然で流暢なデータであり、原言語側に少々誤りがあっても流暢な目的言語への翻訳の学習には有効であること、また原言語側にある程度の「揺らぎ」を生じさせることでモデルの頑健性が向上すること、等が挙げられている[12]。

4 おわりに

本稿では、近年急速な進展を遂げた NMT の技術について簡単に紹介した。機械翻訳サービス等での利用も急速に進んでおり、特許翻訳においても Google の機械翻訳技術を利用している欧州特許庁（EPO）や独自の NMT を開発している世界知的財産権機関（WIPO）に加え、本邦特許庁においても情報通信研究機構（NICT）

の技術に基づく NMT が稼働しており、すでに実用されていると言ってよい。

しかし実際の技術として機械翻訳で 100% の翻訳精度を得ることはまだ現実的ではない。特に特許翻訳では翻訳対象がしばしば専門用語を多く含むこと、用語の選択が非常に重要であり分野ごとの訳し分けを考慮する必要があること、また請求項に代表されるように構造が複雑で長い文がしばしば翻訳対象になること等から、パテントファミリに基づく大規模な対訳データの入手性を加味しても、機械翻訳としては依然として難しい課題と言えよう。NMT において顕著な問題として学習な不十分な語句の訳抜けの問題が知られており、特許文書の役割を考えると非常に重大なタイプの誤りと言える。用途を絞らない一般的な NMT の更なる発展が今後期待される一方で、特許翻訳のような実応用に即した NMT の更なる進化が求められる。

参考文献

[1] I. Sutskever et al., Sequence to Sequence Learning with Neural Networks, Proc. NIPS, pp.3104-3112 (2014)

[2] D. Bahdanau et al., Neural Machine Translation by Jointly Learning to Align and Translate, arXiv preprint 1409.0473 [presented at ICLR 2016] (2014)

[3] T. Luong et al., Effective Approaches to Attention-based Neural Machine Translation, Proc. EMNLP, pp.1412-1421 (2015)

[4] A. Vaswani et al., Attention Is All You Need, Proc. NIPS, pp.5998-6008 (2017)

[5] J. Gu et al., Non-Autoregressive Neural Machine Translation, arXiv preprint 1711.02281 [presented at ICLR 2018] (2017)

[6] J. Gu et al., Levenshtein Transformer, Proc. NeurIPS, pp.11181-11191 (2019)

[7] M. Ghazvininejad et al., Mask-Predict: Parallel Decoding of Conditional Masked Language Models, Proc. EMNLP, pp. 6112-6121 (2019)

[8] R. Shu et al., Latent-Variable Non-Autoregressive Neural Machine Translation with Deterministic Inference using a Delta Posterior, Proc. AAAI, pp. 8846-8853 (2020)

[9] R. Sennrich et al., Neural Machine Translation of Rare Words with Subword Units, Proc. ACL, pp.1715-1725 (2016)

[10] T. Kudo et al., SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, Proc. EMNLP (System Demonstrations), pp.66-71 (2018)

[11] R. Sennrich et al., Improving Neural Machine Translation Models with Monolingual Data, Proc. ACL, pp.86-96 (2016)

[12] S. Edunov et al., Understanding Back-Translation at Scale, Proc. EMNLP, pp.489-500 (2018)