ニューラル機械翻訳における訳語誤りの改善

Improvement of translation error correction in neural machine translation

元山梨英和大学教授
江原　暉将

1967 年早稲田大学理工学部卒。同年 NHK 入局。2003 年 諏訪東京理科大学教授。2009 年 山梨英和大学教授。2015 年 退職。アジア太平洋機械翻訳協会 (AAMT) ／ Japio 特許翻訳研究会委員。

有限会社アジア産業　研究開発部部長
岡　俊行

1983 年東京工業大学数学科卒。株式会社クロスランゲージなどを経て、現在アジア産業に拠点を置きつつ、主にプログラマとして活動中。

1 はじめに

ニューラル方式機械翻訳（NMT：Neural Machine Translation）によって機械翻訳の性能が格段に向上し、特許翻訳においても実用が進んでいる。しかしNMT も完全ではなく、従来の統計的機械翻訳（SMT：Statistical Machine Translation）には見られなかった問題として、不足翻訳（訳抜け）や過剰翻訳（湧き出し）が指摘されている[1]。さらに NMT においては、「奇妙な」誤訳が存在する[2]。「奇妙な」という意味は、「訓練データに出現しない訳語が翻訳結果に現れる」という意味である。

文献［2］では、中日翻訳の中で「嵌石」が「スラグ」と訳される例が指摘されている。訓練データ中では、「嵌石」は「石噛み」と訳されており「スラグ」と訳された例はなかった。にもかかわらず試験データでは「嵌石」が「スラグ」と訳されていた。

文献［2］では、さらに、このような誤訳を後修正に

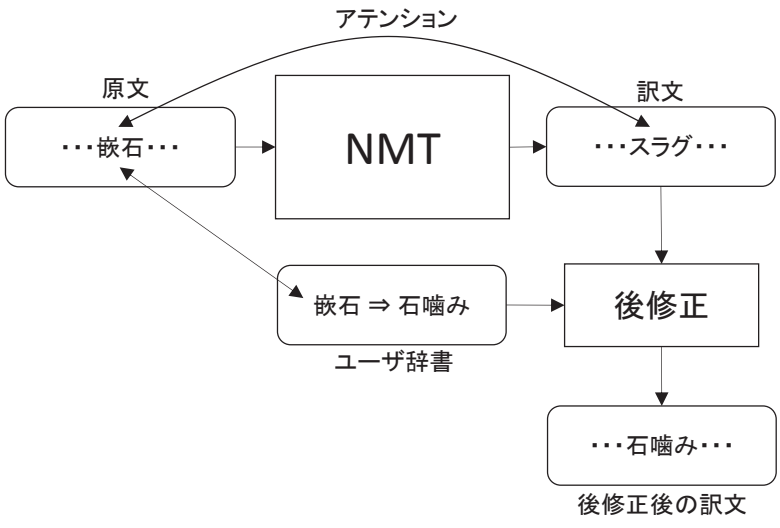

図 1　後修正による訳語の変更

よって正訳に修正する手法が提案されている。その方法を【図１】に示す。「嵌石」を含む原文をNMTによって翻訳したとき、訳文に「スラグ」が含まれるとする。この「スラグ」をユーザ辞書を用いて「石噛み」に後修正する。この時、原文の「嵌石」が訳文の「スラグ」に対応しているという情報が必要であり、このような情報をNMTで用いられるアテンション機構から得ている。

本文で提案する方法は、このような後修正ではなく、前修正で誤訳を回避するものである。前修正による方法はアテンション機構を必要としないという利点がある。

2 前修正による訳語の変更

前修正による訳語の変更方法を【図2】に示す。ここでは英日翻訳を対象としている。英語原文に含まれる”sixth negative strain”を「6 番目の陰性菌株」と訳したい場合、原文の当該部分を日本語に前修正してからNMTで翻訳することで、訳文に「6 番目の陰性菌株」を訳出す手法である。

【図 2】の例の全文を示すと次のようになる。

原文：

the sequence for the sixth negative strain（ MC58 ）was already available from the complete genome sequence .

前修正後の原文：

the sequence for the 6 番目の陰性菌株（ MC58 ）was already available from the complete genome sequence .

前修正後の原文のNMTによる訳文：

6 番目の陰性菌株（ MC58 ）の配列は、完全ゲノム配列から既に入手可能である。

ちなみに前修正を行う前の原文をNMTした場合、以下のような訳文が得られた。

第 6 負株（ MC58 ）の配列は、完全ゲノム配列から既に入手可能である。

つまり"sixth negative strain”が「第6負株」と訳されている。また日本語参照訳文は以下のようである。

6 番目の陰性菌株（ MC58 ）の配列は、全ゲノム配列から既に入手可能であった。

前修正による手法はアテンション機構を必要としないが、前修正後の原文には、英語中に一部日本語が含まれることになり、そのような原文へ対応できる翻訳システムが必要である。

3 日本語混じり英語に対応したNMT

日本語混じり英語に対応したNMTを構築するには、訓練データの英語側も日本語混じり英語にしておく必要がある。筆者らの実験では、約800万文対の英日特許コーパスから、頻度100未満の低頻度の名詞あるいは複合名詞を抽出して、日本語化した。頻度100未満とした理由は、低頻度語が誤る可能性が高いことと、高頻度語は文脈によって訳語が変わる可能性が高いことによる。訓練データの前修正の例を以下に示す。

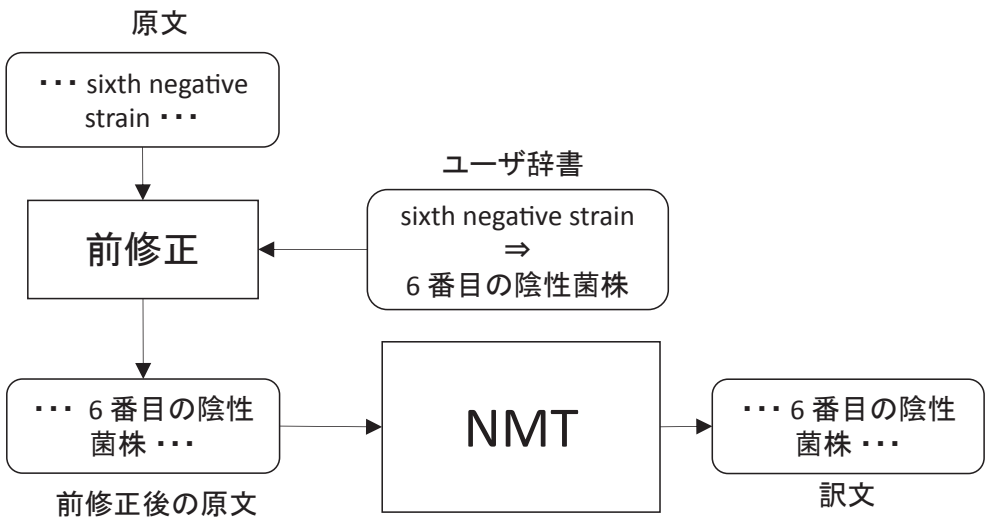

図２　前修正による訳語の変更

原文：

particularly , a laminated structure obtained by a method for producing an orally administrable edible agent of laminate film form using the pressure bonding technique is characterized in that each of laminated edible layers is definitely divided .

前修正後の原文：

particularly , a laminated structure obtained by a method for producing an 可食 性 口腔 内 投与 剤 of 積層 フィルム 状 using the 圧着 法 is characterized in that each of 積層 さ れ た 各 可食 性 層 is definitely divided .

参照訳文：

特に 、この 圧着 法 を 用い た 積層 フィルム 状 の 可食 性 口腔 内 投与 剤 の 製造 方法 に より 得 られる 積層 構造 は 、積層 さ れ た 各 可食 性 層 が 個 々に 明確 に 区分 さ れ て いる こと が 特徴 で あ る 。

このように日本語化された名詞または複合名詞は1,302万個所あった。こうして得られた前修正後の原文と日本語参照訳文を対にしてNMTを訓練した。用いたNMTツールはMarian Transformerである[3]。

4 試験結果の評価

4,000文を用いて試験した。今回は正確な訳語が得られるかどうかを試験するのが目的であるので、試験データの日本語参照訳文から作られた理想的なユーザ辞書を用いて前修正を行った。その結果、試験データ中の9,537個所の名詞または複合名詞が日本語に書き換えられた。

前修正前の原文に対するNMT訳文と前修正後の原文に対するNMT訳文に対するBLEU値を【表1】に示す。前修正によってBLEU値が8ポイント以上向上している。翻訳例としては、[2]に示したものがある。

表1　試験データに対する翻訳評価値

原文	BLEU
前修正前	39.85
前修正後	48.12

5 関連研究

本実験では英語から日本語への翻訳を対象にした。この場合、前修正後の原文の中で、どの部分が英語でありどの部分が日本語であるかは、文字種別によって判別可能である。しかし、中国語と日本語など文字種別による判別ができない言語対もある。そのような場合にも対処できる方法として、各単語に言語名を付加情報として付与する方法[4]や、複数のエンコーダを利用して訳語を指定する方法[5]が提案されている。これらの方法に比較して筆者らの方法は単純であり、英日など、文字種別によって言語が判別可能な言語対に対しては適用が容易である。

6 おわりに

NMTで特徴的な訳語の誤りに対して前修正を行うことで正訳とする手法を提案した。本手法は後修正による手法と比較してアテンション機構を用いる必要がないという利点がある。今後は、中日など文字種別によっては言語判別ができない言語対にも対応できる手法に発展させたい。

参考文献

[1] 江原暉将：統計方式機械翻訳とニューラル方式機械翻訳のハイブリッドシステム、*Japio YEAR BOOK 2018*、pages 300-303、2018年11月。

[2] 江原暉将、岡俊行：ニューラル機械翻訳における訳語誤りについての分析、*Japio YEAR BOOK 2019*、pages 292-294、2019年11月。

[3] Marcin Junczys-Dowmunt et al., : Marian: Fast Neural Machine Translation in C++, *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, July 2018.

[4] Georgiana Dinu et al., : Training Neural Machine Translation To Apply Terminology Constraints, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–

3068, July 2019.
[5] 石川雄太郎、江原暉将：マルチソーストランスフォーマと専門用語辞書を用いた訳語の制御方法、*言語処理学会第 26 回年次大会発表論文集*、pages 553-556、2020 年 3 月。