# ニューラル機械翻訳における訳語誤りについての分析

Analysis of mistranslations 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

深層学習の技術を機械翻訳に応用したニューラル方式機械翻訳（NMT：Neural Machine Translation）は、期待以上の翻訳性能を持ち[1]、特許翻訳においても注目されている[2]。しかしNMTも万能ではない。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[3]。さらにNMTにおいても、訳語の誤りに起因する誤訳が存在する。本文では、この誤訳の問題に着目して考察する。

NMTの翻訳出力の特徴の一つとして、流暢性が高いという点がある。しかし、流暢な訳文の中に訳語の誤りが含まれる場合がある。例えば【表1】に示す例がある。原文の中国語に含まれる「嵌石」が参照訳文[1]では「石噛み」と表現されているが、NMT訳文では「スラグ」と訳されている。この部分以外は、参照訳文とNMT訳文とは、ほとんど一致しており、NMT訳文は流暢である。このような誤訳は気が付きにくく、不足翻訳や過剰翻訳と比較して、むしろ問題が大きいとも言える。

表1　NMT訳文に含まれる流暢な訳語の誤り

| | |
|---|---|
| 原文 | 并且，在本**实**施方式中，**通过**在胎冠主槽６上**设**置突起２３来可靠地抑制作用有**较**大的接地**压**的胎冠主槽６中的**嵌石**，并且**通过**在胎肩主槽５上不**设**置突起而**确**保其槽容**积**并提高湿路性能。 |
| 参照訳文 | また、本実施形態では、クラウン主溝６に突起２３を設けることで、大きな接地圧が作用するクラウン主溝６での**石噛み**を確実に抑制しつつ、ショルダー主溝５には突起を設けないことにより、その溝容積を確保し、ウェット性能を高めている。 |
| NMT訳文 | また、本実施形態では、クラウン主溝６に突起２３を設けることにより、大きな接地圧が作用するクラウン主溝６における**スラグ**が確実に抑制されるとともに、ショルダー主溝５に突起を設けることなくその溝容積を確保してウェット性能を向上させる。 |

## 2 NMTにおける訳語の誤り

SMTでは大量の対訳文対から成る訓練データから作成したフレーズテーブルと呼ばれる句単位[2]の対訳辞書

1　人手による訳文であり、正訳であるとみなす。

2　言語学での「句」とは異なり、単に語の連続という意味である。

を用いて訳語を決めている。そのため、原語と訳語のペアは、訓練データの対訳文に少なくとも 1 回は現れている。一方、NMT でも大量の対訳文対からモデルを学習させる点は SMT と同様であるが、NMT では単語や構文を多次元のベクトルで表現しており、フレーズテーブルに相当するものが陽に存在するわけではない。そこで、原語と訳語のペアが訓練データに存在するという保証はない。実は、【表 1】の例は筆者らの実験[3]において見られたそのような例である。訓練データの中国語側に「嵌石」を含むデータは 1 件しかなく、その日本語側は「石噛み」であり、「スラグ」ではなかった。

【表 2】に訳語誤りの他の例を示す。中国語の「谷醇溶蛋白」が参照訳文では「プロラミン」と表現されている一方、NMT 訳文では「グルテン」と訳されている。

表 2　NMT 訳文に含まれる専門用語の誤訳

| | |
|---|---|
| 原文 | 玉米醇溶蛋白是玉米**种**子胚乳中的**谷醇溶蛋白**贮存蛋白。 |
| 参照訳文 | ゼインは、トウモロコシ種子の内乳における**プロラミン**貯蔵タンパク質である。 |
| NMT 訳文 | ゼインは、トウモロコシ種子胚乳中の**グルテン**貯蔵タンパク質である。 |

訓練データの中国語部分に「谷醇溶蛋白」を含むデータは 0 件であった。一方、訓練データの日本語部分に「グルテン」を含むデータは 164 件であり、それらのうち、中国語側に「谷蛋白」を含むデータは 117 件、「面筋」を含むデータは 21 件、「谷朊」を含むデータは 11 件、「麸质」を含むデータは 8 件、「麦麸」を含むデータは 6 件であった。また訓練データの日本語部分に「プロラミン」を含むデータは 158 件であり、それらのうち、中国語側に「醇溶谷蛋白」を含むデータが全件であった。

## 3 NMT で訳語を誤る理由

なぜ NMT では訓練データに出現しない訳語が訳出されるのであろうか。筆者は NMT 訳文の流暢性の高さにヒントがあると考える。つまり、NMT においては、訳文（ここでは日本文）の流暢性の高さを重視するあまり、訳語としての正確性が失われることがあるのではないだろうか。【表 1】の例で考察する。訓練データの日本語部分に「スラグ」を含むデータは 1314 件である。一方、訓練データの日本語部分に「石噛」を含むデータは 6 件しかない。NMT は、多数回出現する表現を流暢であると認識するため、「石噛」より「スラグ」を選択しやすくなる。実際には、NMT はもっと広範囲な文脈を利用しているので、このような単純なものではないが、訳語を誤る理由の一つとして考えられよう。

SMT においても、流暢性を確保するためのモデル（言語モデル）が使われているが、訳語はあくまでもフレーズテーブルに含まれるものから選ばれるので、NMT でのこのような誤訳は発生しない。

## 4 NMT での訳語誤りの回避

では、NMT 独特のこのような訳語誤りを回避するにはどうしたらよいであろうか。NMT における訳語を決める部分に、フレーズテーブルあるいは他の対訳辞書による制約を加えることが考えられる。実際、例えば文献[7]においては、そのような制約を加えることで、翻訳の自動評価値（BLEU）を向上させている。筆者らが行ったフレーズテーブルを用いた NMT 訳文の自動後修正実験においても【表 1】および【表 2】の NMT 訳文が【表 3】のように後修正された。

表 3　NMT 訳文の自動後修正結果

| | |
|---|---|
| 表 1 の例 | また、本実施形態では、クラウン主溝６に突起２３を設けることにより、大きな接地圧が作用するクラウン主溝６における**石噛**が確実に抑制されるとともに、ショルダー主溝５に突起を設けることなくその溝容積を確保してウェット性能を向上させる。 |
| 表 2 の例 | ゼインは、トウモロコシ種子内乳中の**谷アルコールロッコ**貯蔵タンパク質である。 |

【表 1】の例では「スラグ」が正しく「石噛」と修正された。【表 2】の例では「グルテン」が「谷アルコールロッコ」と修正され、正訳は得られなかったが、流暢な誤訳ではなくなった。

3　中日の特許文、約 745 万文対から成る訓練データを用いて実験した。NMT には OpenNMT[4] の RNN（recurrent neural network）モデル[5]を用い、SMT（フレーズテーブルの作成）には Moses[6] のフレーズベース SMT を用いた。

## 5 おわりに

NMT によって機械翻訳の精度は格段に向上したが、不足翻訳（訳抜け）、過剰翻訳（湧き出し）、流暢な誤訳の問題が残されている。本文では、流暢な誤訳に着目して、その性質を調査するとともに、流暢な誤訳を回避する方法について考察した。回避方法のアイデアは、SMT で用いられていたフレーズテーブルを利用して訳語の範囲を制限するというものであり、実験の結果、ある程度の有効性が見られた。今後は手法を洗練するとともに不足翻訳、過剰翻訳への適用も検討したい。

### 参考文献


[1] 鶴岡慶雅：ニューラル機械翻訳の衝撃、情報処理、Vol. 58、No. 2、pages 96-97、2017 年 2 月。

[2] 本間奨ほか：特許翻訳の新潮流、JTF Journal、No. 298、pages 8-19、2018 年 11-12 月。

[3] 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。

[4] Guillaume Klein et al., : OpenNMT: Open-Source Toolkit for Neural Machine Translation, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 67–72, July 2017.

[5] Dzmitry Bahdanau, KyungHyun Cho and Yoshua Bengio : Neural Machine Translation by Jointly Learning to Align and Translate, arXiv:1409.0473, Sep. 2014.

[6] Philipp Koehn et al., : Moses: Open Source Toolkit for Statistical Machine Translation, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177-180, June 2007.

[7] Eva Hasler et al., : Neural Machine Translation Decoding with Terminology Constraints, arXiv:1805.03750, May 2018.


123456789
Dddddd
Ccccc