ニューラル機械翻訳における訳語の統一

Translation unification 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。特許翻訳では技術用語の正確な翻訳が求められる。NMTでは大量の対訳コーパスから技術用語の対訳を学習するため、頻度の少ない技術用語や固有名詞などに対して所望の訳とは異なる訳が出力されることがある。

NMTにおいて対訳辞書を用いた訳語統一の手法が様々に提案されている[1]。その中に、出力文を修正すること（後修正方式）[2]、または入力文を修正すること（前修正方式）[3][4]で訳語統一を行う方法がある。これらの方法は、翻訳システム自体には手を入れる必要がないという利点がある。本文ではさらに進んで、入力文と出力文を修正すること（前後修正方式）で訳語統一を行う方法を提案する。

2 後修正方式

英日翻訳を例に説明する。次の入力英文

in certain embodiments , the family B archael DNA polymerase is 9°N polymerase .

に対して正解日本文（参照訳文）が

ある態様において、ファミリーB古細菌DNAポリメラーゼは9°Nポリメラーゼである。

のように与えられているとする。技術用語の英日対応辞書として

certain embodiments ⇒ある態様

family B archael DNA polymerase ⇒ファミリーB古細菌DNAポリメラーゼ

N polymerase ⇒Nポリメラーゼ

の3項目が得られる。

通常のNMT[1]の出力訳文として

ある実施態様においては、前記Bアーチセル DNAポリメラーゼは、9°Nポリメラーゼである。

が得られたとする。次の2点の後修正が必要となる。

ある実施態様⇒ある態様

前記Bアーチセル DNAポリメラーゼ⇒ファミリーB古細菌DNAポリメラーゼ

1　実験に用いたNMTシステムはMarian Transformer[5]であり、800万文対の英日特許対訳コーパスで訓練したシステムである。英日混合の2万トークンでSentencepiece[6]化を行った。

このような修正を行うためには入力英文と出力訳文の次の対応を何らかの方法で求めることが必要である。

certain embodiments ⇒ある実施態様

family B archael DNA polymerase ⇒前記 B アーチセル DNA ポリメラーゼ

その後、技術用語の英日対応辞書を用いて正解訳と入れ替えることで後修正する。

文献［2］では、このような対応を原文と訳文の各語の間のアテンション（source-target attention）の情報から得ている。

後修正方式では、入力英文と出力訳文の間の技術用語の対応が得られなかったり対応が誤っていたりすると、後修正自体も誤ってしまうという欠点がある。

3 前修正方式

後修正方式では、原文と訳文の技術用語の部分に対して、対応情報を必要とした。そのような情報を必要としない方法として前修正方式がある。この方式では、前記の入力英文に対して、技術用語の英日対応辞書を用いて前修正を加える。その結果

in ある態様 ,the ファミリー B 古細菌 DNA ポリメラーゼ is 9゜ N ポリメラーゼ .

のように技術用語の部分だけ日本語化した英文（日本語語彙化英文と呼ぶ）が得られる。このような日本語語彙化英文を元に翻訳することで日本語の正しい技術用語が訳出されることが期待できる。ただし訓練時に用いる英文側も日本語語彙化しておく必要がある。

前修正方式では、日本語と英語が混在した日本語語彙化英文を用いているので、語彙数が増大し翻訳の困難度が増すという欠点がある。

4 前後修正方式

後修正方式と前修正方式の良いところをとる方式として前後修正方式を提案する。前修正方式では原文英文の技術用語の部分を日本語化したが、前後修正方式では日本語化まではせず、技術用語の部分であることを示す特殊なタグで囲むことのみ行う。これによって、後修正時に必要な技術用語の対応の正確性を増せる可能性がある。前記の例で示す。タグとしては全角の山かっこ（＜と＞）を用いた。タグ化英文は

in ＜ certain embodiments ＞ ,the ＜ family B archael DNA polymerase ＞ is 9 ゜ ＜ N polymerase ＞ .

となり、翻訳結果は

＜ある実施態様＞において、＜ B 型アーチセル DNA ポリメラーゼ＞は 9゜＜ N ポリメラーゼ＞である。

となった。もちろん訓練時に用いる英文と日本文には、なんらかの対応手段を用いてタグを付与しておく必要がある[2]。

タグを頼りに技術用語部分の英日対応を得て、後修正を行うことができる。対応を得る手法としては、2 と同様に原文と訳文の各語間のアテンション情報を用いる手法に加えて GIZA++ の force align[3] を用いる手法がある。5 に示す実験では両者を比較している。

5 実験結果

2 で述べた NMT システムを用いて実験を行った。4,000 文の試験文を用いた実験の結果、【表 1】のような BLEU 値が得られた。ただし今回の実験は、訳語統一の精度を比較するのが目的であるため、試験データから構築した理想的な技術用語の英日対応辞書を用いて実験した。訳語統一をしない通常の NMT において 43.65 であった BLEU 値が、最良システムである前後修正方式（force align）では 50.02 と 7 ポイント以上向上した。後修正における attention の利用と force align の利用では後者のほうが BLEU 値が高かった。後修正方式と前修正方式を比較すると前修正方式の方が BLEU 値が高かった。

表 1　実験結果

システム	BLEU
NMT	43.65
後修正方式 force align	49.11
後修正方式 attention	48.25
前修正方式	49.15
前後修正方式 force align	50.02
前後修正方式 attention	48.55

2　今回の実験では、訓練データに対する技術用語の対応を GIZA++ を用いて求めた。

3　https://github.com/moses-smt/mgiza/blob/master/mgizapp/scripts/force-align-moses.sh

これまで用いてきた例文に対して、原文や訳文などを【表 2】に示す。

表 2　翻訳例

入力英文	in certain embodiments , the family B archael DNA polymerase is 9 ° N polymerase .
参照日本語訳文	ある 態様 に おい て 、ファミリー B 古 細菌 DNA ポリメラーゼ は 9 °　N ポリメラーゼ で ある 。
NMT出力訳文	ある 実施 態様 に おい て は 、前記 B アーチ セル DNA ポリメラーゼ は 、9 °　N ポリメラーゼ で ある 。
後修正方式訳文 force align	ある 実施 態様 に おい て は 、ファミリー B 古 細菌 DNA ポリメラーゼ は 、9 °　N ポリメラーゼ で ある 。
後修正方式訳文 attention	ある 態様 に おい て は 、前記 ファミリー B 古 細菌 DNA ポリメラーゼ は 、9 °　N ポリメラーゼ で ある 。
日本語語彙化英文	in ある 態様 , the ファミリー B 古 細菌 DNA ポリメラーゼ is 9 ° N ポリメラーゼ .
前修正方式訳文	ある 態様 で は 、ファミリー B 古 細菌 DNA ポリメラーゼ は 、9 °　N ポリメラーゼ で ある 。
タグ化英文	in ＜ certain embodiments ＞ , the ＜ family B archael DNA polymerase ＞ is 9 °　＜ N polymerase ＞ .
タグ化英文の訳文	＜ ある 実施 態様 ＞ に おい て 、＜ B 型 アーチ セル DNA ポリメラーゼ ＞ は 9 °　＜ N ポリメラーゼ ＞ で ある 。
前後修正方式訳文 force align	ある 態様 に おい て 、ファミリー B 古 細菌 DNA ポリメラーゼ は 9 °　N ポリメラーゼ で ある 。
前後修正方式訳文 attention	ある 態様 施 態様 に おい て 、ファミリー B 古 細菌 DNA ポリメラーゼ は 9 °　N ポリメラーゼ で ある 。

6　まとめ

技術用語の対応辞書を用いた訳語統一の手法として、後修正方式、前修正方式、前後修正方式を報告した。これらの手法は NMT システム自体には手を入れる必要がないため、どのようなシステムに対しても適用できる。英日翻訳での実験の結果 force align を用いた前後修正方式が最も BLEU 値が高かった。

参考文献

[1] 今村 賢治 , 越前谷 博 , 江原 暉将 , 後藤 功雄 , 須藤 克仁 , 園尾 聡 , 綱川 隆司 , 中澤 敏明 , 二宮 崇 , 王 向莉 : 特許機械翻訳の課題解決に向けた機械翻訳技術解説、自然言語処理解説論文 Vol.29, No.3, pp.925-985, 2022.

[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［寄稿集］*, pp.292-295、Nov. 2019.

[3] 江原暉将、岡 俊行 : ニューラル機械翻訳における訳語誤りの改善、*Japio YEAR BOOK 2020［寄稿集］*、pp. 292-295, Nov. 2020.

[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*, pp.3063–3068, July 2019.

[5] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, Andre F.T. Martins, and Alexandra Birch : Marian: Fast Neural Machine Translation in C++, *Proceedings of ACL 2018, System Demonstrations*, pp.116–121, 2018.

[6] Taku Kudo, John Richardson : SentencePiece:A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, *arXiv:1808.06226*, 2018.