往復翻訳を用いたSMTとNMTのハイブリッドシステム

Hybrid System of SMT and NMT using Two-way 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。

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

大量の対訳コーパスから機械翻訳システムを構築する手法に統計的機械翻訳（SMT：Statistical Machine Translation）とニューラル機械翻訳（NMT：Neural Machine Translation）がある。SMT は対訳コーパスからフレーズテーブルと呼ばれる一種の対訳辞書を作成し、それを用いて訳語を決定する。そこで翻訳対象原文に含まれる原語に対する訳語のペアは対訳コーパス中に出現するペアに限られる。一方、NMT では訓練結果が多次元のベクトル空間としてモデル化されているため、対訳辞書が陽に存在せず、原語に対する訳語のペアが対訳コーパスに出現しないものが現れることがある [1]。

しかし平均的には NMT は SMT より翻訳精度も高く、特に流暢性の観点から SMT を大幅に上回っている。このように NMT と SMT はそれぞれに得失がある。

本文では SMT と NMT を組み合わせたハイブリッドシステムについて考察する。翻訳対象原文を NMT と SMT で独立に翻訳し 2 つの翻訳結果を評価して評価値の高い方を出力として選択するものである。評価方法としては往復翻訳を用いる。

2 システム構成

ベトナム語から日本語（越日）への翻訳を例として、システム構成を【図 1】に示す。

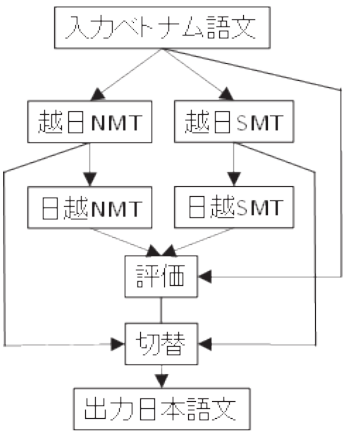

図 1　システム構成

越日対訳コーパスを用いて越日 NMT、日越 NMT、越日 SMT、日越 SMT の 4 つの機械翻訳システムを構築する。翻訳時には入力ベトナム語文を越日 NMT および越日 SMT でそれぞれ日本語文に翻訳する。越日

NMT の出力日本語文は日越 NMT でベトナム語文に戻される。越日 SMT の出力日本語文も日越 SMT でベトナム語文に戻される。両者を入力ベトナム語文と比較して近い方が正しい翻訳結果を得ているとみなし、そちらの出力を選択して最終的な出力日本語文とする。ここで近さの基準は文単位で計測した BLEU を用いる [1]。

このような往復翻訳による評価は以前から提案されており [3]、最近も用いられている [4]。

3 実験設定

特許文書から構築した越日文対応コーパスから以下のデータを抽出して実験を進めた。

訓練データ　1,220,086 文対

試験データ　2,097 文対

開発データ　2,175 文対

NMT システムとしては Marian Transformer を用いた [5]。SMT システムとしては Moses を用いた [6]。ベトナム語の形態素分割には UET segmenter を用いた [7]。日本語の形態素分割には Unidic ベース [8] の Mecab を用いた [9]。サブワード化として越日ともに Latin 特殊文字とアラビア数字を 1 文字ずつに分割した。SMT システムの訓練と試験にはこのサブワード化の結果を用いた。NMT システムの訓練と試験には、さらにベトナム語・日本語別々に 2 万 piece で SentencePiece 化 し た [10]。SentencePiece model の訓練には翻訳システム訓練用のデータを流用した。

【図 1】の切替部分で、NMT を優先する手法を用いた。具体的には、NMT 側の往復翻訳結果の BLEU 値を BLEU（NMT）、SMT 側の往復翻訳結果の BLEU 値を BLEU（SMT）とし、BONUS を正の定数とするとき

BLEU（SMT）＞ BLEU（NMT）＋ BONUS

が成立する場合は SMT による翻訳結果を選択し、成立しない場合は NMT による翻訳結果を選択する。BONUS の値としては 0.2 を設定した。本設定については 5. で述べる。

4 実験結果

各システムの試験データに対する BLEU 値を【表 1】に示す。BLEU を計算するときの形態素分割は 3. で述べたサブワード単位で行った。これは NMT、SMT に共通である。

表 1　実験結果

システム	BLEU
越日NMT単独	0.5865
越日SMT単独	0.5043
オラクル	0.6448
ハイブリッド	0.6132

ハイブリッドシステムは越日 NMT 単独システムと比較して 0.0267 ポイント BLEU が向上している。表中オラクルとは越日 NMT と越日 SMT の結果のうち BLEU が高い方を理想的に選択できた場合であり、ハイブリッドシステムが到達出来うる BLEU の最高値である。

NMT 単独とハイブリッドシステムの文単位の BLEU 値の頻度分布を【図 2】に示す。

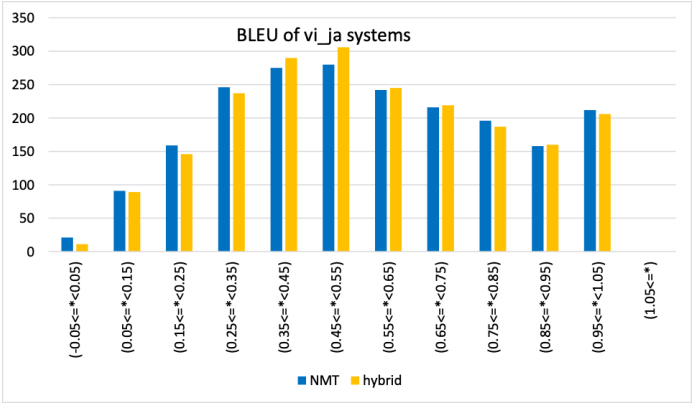

図 2　NMT 単独とハイブリッドシステムの文単位の BLEU 値の頻度分布

1　BLEU の計算には mteval-v13a.pl を用いた [2]。

BLEU が 0.35 以下の部分と 0.75 以上の部分でハイブリッドシステムの頻度が減少しており、今回の実験設定では、ハイブリッドシステムが平均的に BLEU を向上させるとともに、ばらつきも少なくしていることが分かった。

5 BONUS の値の設定

【表 2】に越日 NMT、越日 SMT、越日越 NMT、越日越 SMT の各システムの試験データに対する BLEU 値を示す。前 2 者は日本語で計測しており、後 2 者はベトナム語で計測している。

表 2　各種システムの BLEU 値

システム	BLEU
越日NMT単独	0.5865
越日SMT単独	0.5043
越日越NMT	0.6419
越日越SMT	0.6841

越日では NMT の方が BLEU が高いが、越日越では SMT の方が BLEU が高い。このことから越日越の BLEU を評価基準とするハイブリッドシステムでは SMT を過度に優遇する可能性が高い。そこで NMT を優先させるための BONUS を設定した。BONUS の値を変化させて試験データに対するハイブリッドシステムの BLEU を調べたところ【表 3】を得た。この結果から BONUS=0.2 と設定した。

表 3　BONUS の値の変化に対する BLEU の変化

BONUS	BLEU
0	0.5806
0.1	0.6037
0.2	0.6132
0.3	0.6103
0.4	0.6068

6 まとめ

NMT と SMT の良いところ取りを目的に両者のハイブリッドシステムを構築した。往復翻訳を用いて NMT 出力と SMT 出力を比較し評価値の高い方を選択する。実験の結果、NMT 単独のシステムと比較して BLEU 値を 0.0267 ポイント向上させることができた。

参考文献

[1] 江原暉将，岡俊行．2019. 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.

[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl

[3] Yokoyama, Shoichi et al. 2001. An Automatic Evaluation Method for Machine Translation using Two-way MT, Proceedings of Machine Translation Summit VIII.

[4] 岩田一成．2023．音声翻訳機を使うための日本語，第 14 回産業日本語研究会･シンポジウム資料，14-3-1.pdf.

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

[6] Koehn, Philipp et al. 2007. 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, pp 177-180.

[7] Nguyen, Tuan-Phong et al. 2016. A Hybrid Approach to Vietnamese Word Segmentation, IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF) .

[8] 伝康晴ほか．2007．コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用，日本語科学，Vol.22, pp.101-123.

[9] Kudo, Taku, Kaoru Yamamoto, Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp.230-237.

[10]　Kudo, Taku, John Richardson. 2018.

SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp.66-71.