低資源言語におけるニューラル機械翻訳の精度向上のための文ベクトルの変換

Transformation of sentence vector to improve translation quality of neural machine translation in low-resource language

北海学園大学　大学院工学研究科　教授
越前谷　博

1996 年北海学園大学大学院工学研究科修士課程修了。博士（工学）。2013 年～現在北海学園大学大学院工学研究科教授。機械翻訳の研究に従事。アジア太平洋機械翻訳協会（AAMT）／ Japio 特許翻訳研究会委員。

echi@hgu.jp　　TEL 011-841-1161（内線：7863）

① はじめに

ニューラル機械翻訳や大規模言語モデルの進展により、機械翻訳の精度が向上している。特に日本語英語間等の資源の豊富な言語間の翻訳精度の向上は著しい。日本語英語間のような言語間では対訳コーパス等の言語資源を比較的容易に収集できるため、大規模な学習データを用いた精度の高い翻訳モデルの構築が可能である。

それに対して、言語資源の乏しい低資源言語においては十分な学習データの収集が困難なため、構築するモデルの精度も不十分となる。このような低資源言語に対するニューラル機械翻訳の精度向上のためのアプローチとしては、データ拡張のアプローチと複数の言語モデルを利用するアプローチに大きく分類される。データ拡張のアプローチでは、例えば Zhang ら [1] は機械翻訳システムにより得られた訳文を原言語文に対する目的言語文とすることで疑似対訳コーパスを生成した。また、Sennrich ら [2] は元の対訳コーパスに逆翻訳により生成した疑似対訳コーパスを加えることで翻訳精度の向上を試みた。さらに、疑似対訳コーパスを大規模言語モデルにより生成する手法 [3] も提案されている。しかし、このような疑似対訳コーパスを用いるアプローチではその精度が問題となり、翻訳精度に必ずしも良い影響を与えるとは限らない。特に低資源言語においてはニューラル機械翻訳や大規模言語モデルより得られる訳文の精度は低くなる可能性が高く、その結果、翻訳精度の向上の妨げとなる。

また、複数の言語モデルを利用するアプローチとして、Zoph ら [4] は低資源言語を用いて構築した子モデルの性能向上のために豊富な言語資源より構築された親モデルのパラメータに基づき子モデルの初期化を行った。しかし、この場合、子モデルと親モデルの構築に用いるデータの言語は類似していることが前提となり、類似性が低い言語間においては親モデルのパラメータは子モデルにおいて有効に機能しない可能性がある。

そこで本報告では、低資源言語においては逆翻訳により得られる疑似対訳コーパスの精度が著しく低下する可能性を考慮し、複数の言語モデルを用いるアプローチによる、新たなニューラル機械翻訳を提案する。提案手法では、Transformer[5] に基づくニューラル機械翻訳として提案された BERT-fused NMT[6] において、事前学習したモデルが生成する文ベクトルを Encoder の単語埋め込みベクトルとして用いる。このモデルは原言語の文ベクトルに対応する目的言語の文ベクトルの情報を有した文ベクトルを生成するため、翻訳精度の向上が期待される。本報告では、提案手法とその有効性をアイヌ語日本語間、ベトナム語日本語間、そして、アッサム語英語間による翻訳実験に基づき述べる。

② 提案手法の概要

提案手法に基づくニューラル機械翻訳の概要図を図 1 に示す。提案手法によるニューラル機械翻訳では原言語文の文ベクトルを目的言語文の文ベクトルとの間で対応関係にある文ベクトルに変換し、それを単語埋め込みベクトルとして用いる。まず、原言語文を多言語の

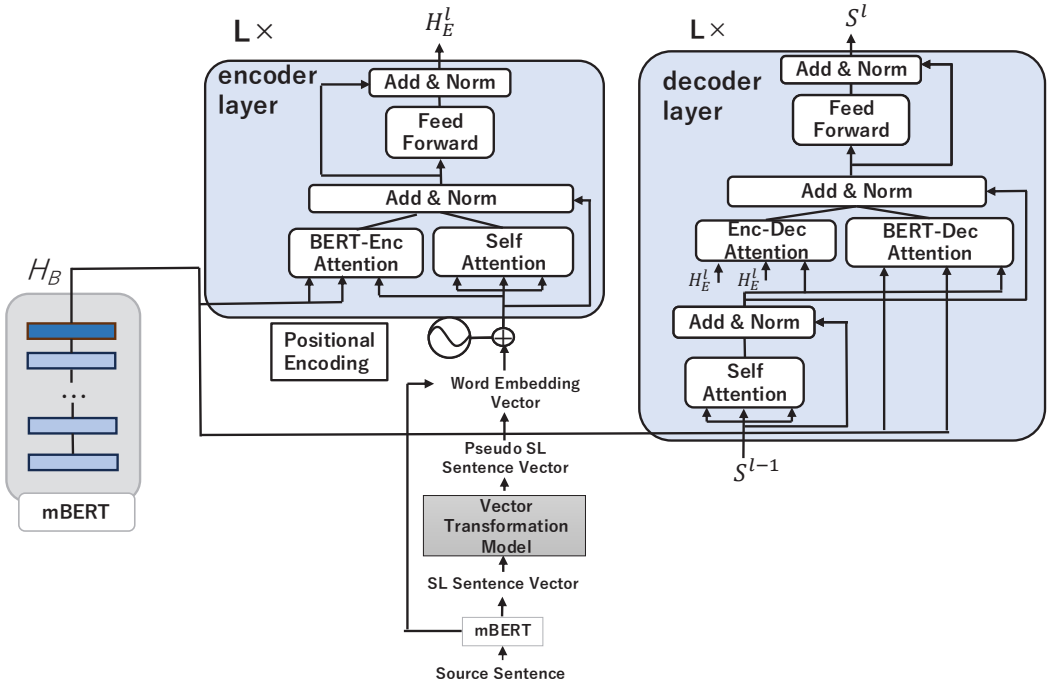

図 1　提案手法に基づくニューラル機械翻訳 Vec-trans NMT の概要図

Encoder モデルである mBERT[7] を用いて文ベクトル（以降、原言語文ベクトルと呼ぶ）を得る。原言語文ベクトルは mBERT により得られる単語ベクトルの平均を求めることで得られる。次いで、事前学習されたニューラルネットワークモデルを用いて目的言語の文ベクトル（以降、目的言語文ベクトルと呼ぶ）に対応する疑似的な原言語文ベクトル（以降、疑似原言語文ベクトルと呼ぶ）を生成する。そして、その疑似原言語文ベクトルを単語埋め込みベクトルとして BERT-fused NMT の Encoder に用いる。文ベクトルから単語ベクトルの変換処理は単語ベクトルの要素ごとの平均が文ベクトルの要素と一致するように単語ベクトルの各要素に任意の固定値を付与することで行う。本報告では疑似原言語文ベクトルを得るための事前学習済みのニューラルネットワークモデルをベクトル変換モデルと呼ぶ。また、このベクトル変換モデルを用いたニューラル機械翻訳を Vec-trans NMT と記す。

③ ベクトル変換モデル

3.1　ベクトル変換モデルの構築の概要

ベクトル変換モデルは原言語文ベクトルとその目的言語文ベクトル、そして、目的言語文に対する評価スコアを用いて学習を行う。原言語文ベクトルと目的言語文ベクトルは共に mBERT より得られる単語ベクトルの平均ベクトルを求めることで得る。評価スコアは mSBERT[8] より得られる原言語文ベクトルと目的言語文ベクトルとの間のコサイン類似度を用いる。mSBERT は文ベクトル間の類似度を求めるためにシャムネットワークにより学習された SBERT[9] に対して、知識蒸留により多言語に拡張した Encoder モデルである。ベクトル変換モデルの目的は入力された原言語文ベクトルとその目的言語文ベクトルの間のコサイン類似度が評価スコアと等価になるように原言語文ベクトルを変換することにある。

ここで対応関係にある原言語文ベクトルとその目的言語文ベクトル間のコサイン類似度は基本的には最大値の 1.0 に近しい値になることが理想ではあるが、mBERT は文ベクトル間の類似度を得ることを目的とはしていないため、文ベクトル間の意味的な類似度を得ることは困難と考えられる。そこで、ベクトル変換モデルにおいて目的言語文ベクトルとの間で意味的な類似度が得られるように原言語文ベクトルを新たな文ベクトル、すなわち、疑似原言語文ベクトルに変換する。そして、その文ベクトルを BERT-fused NMT の単語埋め込みベクトルとして利用することで翻訳精度の向上を図る。

また、学習時に必要な正解データには目的言語文ベクトルとの間のコサイン類似度が評価スコアと等価になるように変換された原言語文ベクトル（以降、正解文ベク

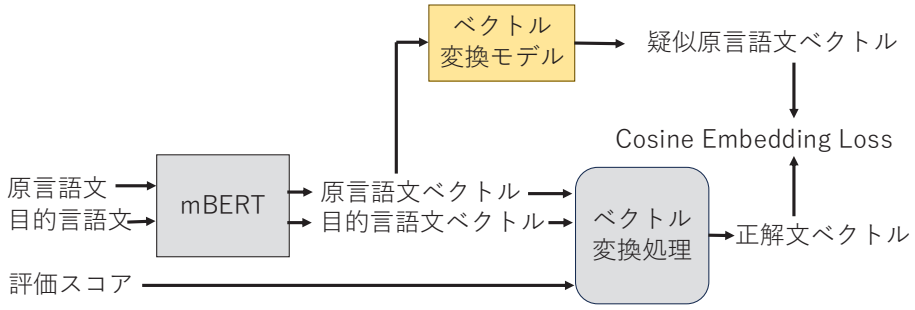

図２　ベクトル変換モデルの構築の概要

トルと呼ぶ）を用いる。この正解文ベクトルの生成方法については次節の 3.2 で詳細を述べる。正解文ベクトルとベクトル変換モデルより得られる疑似原言語文ベクトルとの間の Loss の計算は Cosine Embedding Loss を用いる。図２にベクトル変換モデルの構築の概要を示す。

ベクトル変換モデルは原言語文ベクトルを目的言語文ベクトルと対応関係にある疑似原言語文ベクトルに変換する。変換された疑似原言語文ベクトルに対する正解文ベクトルはベクトル変換処理より原言語文ベクトルと目的言語文ベクトルとの間のコサイン類似度が評価スコアと等価になるように原言語文ベクトルを変換した文ベクトルである。この正解文ベクトルと疑似原言語文ベクトルとの間で Cosine Embedding Loss に基づきベクトル変換モデルの学習を行う。

3.2　正解文ベクトルの生成

ベクトル変換処理における正解文ベクトルの生成について述べる。ベクトル変換モデルの学習に用いる正解文ベクトルは原言語文ベクトルの任意の 1 つの要素に変数 x を加えた際、目的言語文ベクトルと変数 x を付与した原言語文ベクトルの間のコサイン類似度が評価スコアとなるような x を求め、その値を加えることにより得る。変数 x を付与した n 次元の原言語文ベクトルを $S=(s_1,\cdots,s_i+x,\cdots,s_n)$、目的言語文ベクトルを $T=(t_1,\cdots,t_n)$ とした場合のコサイン類似度の計算式を以下の式 (1) に示す。

$$\frac{\sum_{j=1}^{i-1}(s_j \times t_j) + (s_i + x) \times t_i + \sum_{j=i+1}^{n}(s_j \times t_j)}{\sqrt{\sum_{j=1}^{i-1} s_j^2 + (s_i + x)^2 + \sum_{j=i+1}^{n} s_j^2} \times \sqrt{\sum_{j=1}^{n} t_j^2}} = \text{評価スコア} \quad \text{式（1）}$$

式 (1) では s_i に変数 x が付与されている。この計算式は x の 2 次方程式になるため、x の解を求めることで値が得られ、x の値を s_i に加算した原言語文ベクトルが正解文ベクトルとなる。

また、式 (1) の評価スコアは先に述べたように Encoder モデルである mSBERT より原言語文ベクトルと目的言語文ベクトルの間のコサイン類似度を求めることで得られるが、mSBERT に含まれない低資源言語が対象となった場合には代替言語を利用する。代替言語には低資源言語と類似した言語を用い、資源が豊富な言語文を Google 翻訳より代替言語文に翻訳した文を mSBERT に入力することで代替言語文に対応する文ベクトルを生成する。その概要を図３に示す。

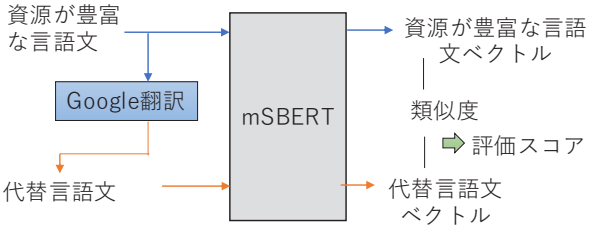

図３　代替言語を用いた評価スコアの生成

低資源言語が mSBERT の学習データに含まれていない場合、その言語の文ベクトルを得ることは困難となるため、低資源言語と文法的に近く、mSBERT の学習データに含まれている言語を用いる。具体的には、Google 翻訳を用いて資源が豊富な言語文を翻訳し、代替言語文を得る。そして、それを mSBERT より代替言語文ベクトルに変換し、資源が豊富な言語文ベクトルとの間のコサイン類似度を得る。ただし、得られたコサイン類似度が閾値以上の場合にのみベクトル変換モデルの学習データである評価スコアに用いる。その結果、より精度の高いベクトル変換モデルが得られると考えられる。

④ 性能評価実験

本報告では提案手法の有効性を確認するためにアイヌ語日本語間、ベトナム語日本語間、そして、アッサム語英語間の翻訳実験を行った。

4.1 実験データ

実験データはベクトル変換モデル及び Vec-trans NMT を構築するためのデータに大きく分かれる。さらに Vec-trans NMT においてはモデル構築のための学習データと翻訳に用いる評価データの 2 種類に分類される。

ベクトル変換モデルの構築においては mSBERT より得られる原言語文ベクトルと目的言語文ベクトルとの間のコサイン類似度の値が閾値を超えている対訳文のみを学習データに用いた。具体的には、アイヌ語日本語間の対訳文においては 72,936 対 [10] から閾値 0.95 を超えていた対訳文 5,787 対を学習データに用いた。その際の訓練データ数は 4,787、検証データ数は 1,000 である。また、アイヌ語は mSBERT の学習データに含まれていないため、アイヌ語に対する代替言語としてモンゴル語、トルコ語、そして、韓国語を用いた。評価スコアはこれら 3 つの言語それぞれの文ベクトルと日本語の文ベクトル間のコサイン類似度の平均値を求め、その平均値が閾値を超えている対訳文をベクトル変換モデルの学習データに使用した。ベトナム語日本語においては対訳文 16,081 対 [11] から文ベクトル間のコサイン類似度が閾値 0.9 を超えている対訳文 6,543 対を学習データに用いた。その際の訓練データ数は 5,543、検証データ数は 1,000 である。そして、アッサム語英語においては対訳文 54,000 対 [12] から文ベクトル間のコサイン類似度が閾値 0.9 を超えている対訳文 7,988 対を学習データとしてベクトル変換モデルを構築した。その際の訓練データ数は 6,988、検証データ数は 1,000 である。また、アッサム語は mSBERT に存在しないため、アッサム語に対する代替言語としてはヒンディー語、フランス語、そして、ギリシャ語を用いて評価スコアを得た。

ニューラル機械翻訳 Vec-trans NMT の構築においてはアイヌ語日本語では上述した対訳文を含む対訳文 91,168 対、ベトナム語日本語では対訳文 20,202 対をそれぞれ 8:1:1 の割合で訓練データ、検証データ、そして、評価データに分割して、学習と翻訳を行った。アッサム語英語においては訓練データ数 50,000、検証データ数 2,000、そして、評価データ数 2,000 を用いて学習と翻訳を行った。

4.2 実験方法

実験は BERT-fused NMT をベースラインとして、提案手法の Vec-trans NMT との比較により行った。BERT-fused NMT、Vec-trans NMT 共に言語ペア間の言語方向毎にモデルを構築した。したがって、提案手法においては 3 つの言語ペアそれぞれにおいて 2 つのベクトル変換モデルと 2 つの Vec-trans NMT のモデルを構築し、翻訳を行った。評価方法は全て自動評価尺度の SacreBLEU を用いた。COMET[13] を始めとしたニューラルネットワークベースの自動評価尺度については学習が行われていない低資源言語の翻訳文に対する評価が不十分となることが予想されるため使用しなかった。

また、ベクトル変換モデルのアーキテクチャは多層パーセプトロンとなっており、768 ニューロンの入力層、1,536 ニューロンの中間層、768 ニューロンの中間層、そして、768 ニューロンの出力層で構成されている。入力層と出力層のニューロンが 768 となっているのは mSBERT による文ベクトルの次元数である 768 と一致させるためである。

4.3 実験結果

表 1 に提案手法による Vec-trans NMT とベースラインである BERT-fused NMT におけるアイヌ語日本語間の BLEU スコアを示す。また、表 2 にベトナム語日本語間の BLEU スコア、そして、表 3 にアッサム語英語間の BLEU スコアを示す。

表 1　アイヌ語日本語の BLEU スコア

	アイヌ語→日本語	日本語→アイヌ語
ベースライン	24.89	32.20
Vec-trans NMT	24.52	33.36

表 2　ベトナム語日本語の BLEU スコア

	ベトナム語→日本語	日本語→ベトナム語
ベースライン	8.69	8.60
Vec-trans NMT	9.42	7.79

表３　アッサム語英語の BLEU スコア

	アッサム語→英語	英語→アッサム語
ベースライン	5.56	14.34
Vec-trans NMT	5.09	15.50

また、ベースラインと提案手法におけるアイヌ語日本語間の翻訳文の例を表４に示す。

表４　翻訳文の具体例

アイヌ語→日本語	
原言語文	situ turasi e = oman wa
正翻訳	尾根に沿って川の上流に行って
ベースライン	尾根の尾根に沿って行って
Vec-trans NMT	尾根に沿って登って行き
日本語→アイヌ語	
原言語文	もう山の上に日が傾いた様子で
正翻訳	tane nupuri ka cup rari kane siran ora
ベースライン	tane nupuri ka cup rari kane sirki
Vec-trans NMT	tane nupuri ka cup rari kane sir ’ an hine

4.4　考察

表１のアイヌ語日本語においては日本語からアイヌ語の翻訳において BLEU スコアが 32.20 から 33.36 へと比較的大きく向上している。表２のベトナム語日本語においてはベトナム語から日本語の BLEU スコアは向上しているが、日本語からベトナム語の BLEU スコアは逆に低下している。しかし、ベトナム語日本語においては、いずれの BLEU スコアも非常に低く、比較するには不十分な結果であったと考えられる。その原因としては学習データが非常に少なく、十分な学習が行われていないためと考えられる。表３のアッサム語英語においては、アッサム語から英語の BLEU スコアは低下しているが、英語からアッサム語の BLEU スコアについては１ポイント以上向上している。

これらの結果より、アイヌ語日本語とアッサム語英語における低資源言語方向の翻訳においては提案手法により BLEU スコアが大きく向上したことが確認できた。これは提案手法によるベクトル変換モデルは mBERT より得られるベクトルを用いて学習していることから資源の豊富な言語を変換した方が精度の高い文ベクトルを生成できるためと考えられる。そこで、ベクトル変換モデルより得られる疑似原言語文ベクトルが目的言語文ベクトルとどの程度類似したものになっているのか調査を行った。調査方法の概要を図４に示す。

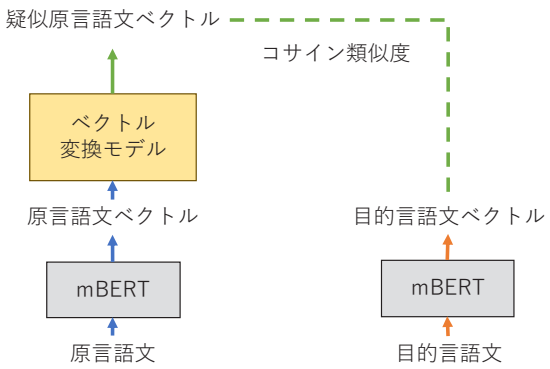

図４　ベクトル変換モデルの精度調査の概要

図４において mBERT より得られる原言語文ベクトルと目的言語文ベクトル間においては先述したようにコサイン類似度を求めても意味的な類似性を得ることは困難であり、コサイン類似度の精度は低いと考えられる。それに対して、ベクトル変換モデルより得られる疑似原言語文ベクトルは目的言語文ベクトルとの間のコサイン類似度が評価スコアと等価になるように学習されているため、目的言語文ベクトルとの間のコサイン類似度は比較的精度の高い値が得られると考えられる。そこで、そのことを確認するためにアイヌ語日本語間、ベトナム語日本語間、そして、アッサム語英語間においてベクトル変換モデルより得られた疑似原言語文ベクトルと目的言語文ベクトルとの間のコサイン類似度を求め、比較を行った。なお、調査に用いた対訳データは３つの言語ペアのいずれも評価データよりランダムに抽出した 100 データであり、それらより得られた 100 のコサイン類似度の平均値を求めた。表５に疑似原言語文ベクトルと目的言語文ベクトルとの間のコサイン類似度の平均値を示す。

表５　文ベクトル間のコサイン類似度の平均値

原言語	目的言語	コサイン類似度の平均値
アイヌ語	日本語	0.7660
日本語	アイヌ語	0.7811
ベトナム語	日本語	0.8563
日本語	ベトナム語	0.8232
アッサム語	英語	0.6085
英語	アッサム語	0.8595

表５において言語ペア毎に双方向でコサイン類似度の平均値を比較すると学習が不十分なベトナム語日本語間を除き、アイヌ語日本語間、アッサム語英語間いずれも資源の豊富な言語から低資源言語方向の翻訳のコサイン類似度の平均値が高いことを確認できる。アイヌ語日本語間では日本語から低資源言語のアイヌ語方向の翻訳時のコサイン類似度の平均値は 0.7811 であり、アイヌ語から日本語方向の翻訳時の 0.7660 に比べて高い。また、アッサム語英語間では英語から低資源言語のアッサム語方向の翻訳時のコサイン類似度の平均値は 0.8595 であり、アッサム語から英語方向の翻訳時の 0.6085 に比べて高い。このような調査結果は表１と表３の資源の豊富な言語から低資源言語方向の BLEU スコアが低資源言語から資源の豊富な言語方向の BLEU スコアよりも高くなっている結果と同様である。したがって、提案手法におけるベクトル変換モデルの精度向上が翻訳精度、すなわち、BLEU スコアの向上に直結していることが確認できた。低資源言語から資源の豊富な言語方向の翻訳精度を向上させるためには、ベクトル変換モデルによる疑似原言語文ベクトルの生成の精度向上が重要と考えられる。

５ まとめ

本報告では複数の言語モデルを用いるアプローチの観点より、ニューラル機械翻訳の精度向上を目的とした新たな手法を提案した。提案手法では Encoder モデルより得られる原言語文ベクトルを目的言語文ベクトルの情報を含む疑似原言語文ベクトルに変換するためにベクトル変換モデルを構築した。そして、生成された疑似原言語文ベクトルをニューラル機械翻訳の単語埋め込みベクトルとして利用した。性能評価実験の結果、疑似原言語文ベクトルを利用することで資源の豊富な言語から低資源言語方向の翻訳において、BLEU スコアがベースラインの BLEU スコアよりも向上し、提案手法の有効性を確認した。

今後は、低資源言語から資源の豊富な言語方向の翻訳精度の向上のためにベクトル変換モデルの改良を行う。具体的には学習時に用いる Encoder モデルを低資源言語に対しても有効利用できるようにファインチューニングを行い、ベクトル変換モデルの精度向上を図る予定である。

参考文献

[1] Jiajun Zhang and Chengqing Zong, “Exploiting Source-side Monolingual Data in Neural Machine Translation,” Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1535–1545 (2016)

[2] Rico Sennrich, Barry Haddow and Alexandra Birch, “Improving Neural Machine Translation Models with Monolingual Data,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 86–96 (2016)

[3] Seokjin Oh, Su Ah Lee, and Woohwan Jung, “Data Augmentation for Neural Machine Translation using Generative Language Model,” arXiv preprint arXiv:2307.16833 (2023)

[4] Barret Zoph, Deniz Yuret, Jonathan May and Kevin Knight, “Transfer Learning for Low-Resource Neural Machine Translation,” Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp.1568-1575. (2016)

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention Is All You Need” , Proceedings of the 31st Conference on Neural Information Processing Systems, pp. 6000–6010, (2017).

[6] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu, “Incorporating BERT into Neural Machine Translation,” Proceedings of 8th ICLR, pp. 1–18, (2020).

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of NAACL-HLT 2019, pp. 4171-4186, (2019).

[8] Nils Reimers and Iryna Gurevych, “Making Monolingual Sentence Embeddings Multilingual using. Knowledge Distillation.” Proceedings of EMNLP, pp. 4512–4525, (2020).

[9] Nils Reimers and Iryna Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Proceedings of EMNLP-IJCNLP 2019, pp.3982–3992, (2019).

[10] 田中蒼大郎 , 越前谷博 , 荒木健治 , “Transformer を用いたアイヌ語 - 日本語間機械翻訳における精度と対訳コーパスの関係 ,” 令和 4 年度電気・情報関係学会北海道支部連合大会講演論文集 , pp. 227-228, (2022).

[11] Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita, “Introduction of the Asian Language Treebank(ALT),” Proceedings of the Tenth International Conference on Language Resources and Evaluation, pp.1574–1578, (2016).

[12] Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure and Sandeep Kumar Dash, “Findings of the WMT 2023 Shared Task on Low-Resource Indic Language Translation,” Proceedings of the Eighth Conference on Machine Translation (WMT), pp. 682–694, (2023).

[13] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie, “COMET: A Neural Framework for MT Evaluation,” Proceedings of EMNLP, pp. 2685–2702, (2020).