多言語BERTを単語埋め込みに用いるMany-to-Many翻訳による低資源言語翻訳*

磯部 僚也† Yizhen Wei† 田村 拓也‡ 宇津呂 武仁† 永田 昌明§
†筑波大学大学院 システム情報工学研究科・群 ‡筑波大学 理工学群 工学システム学類
§NTTコミュニケーション科学基礎研究所

1 はじめに

近年，ニューラル機械翻訳 (Neural Machine Translation; NMT) の研究が盛んに行われている．一般的な NMT モデルは，原言語を一言語，目的言語を一言語設定し，原言語文と目的言語文からなる対訳コーパスを訓練に用い，NMT モデルの品質は対訳コーパスの質や量に依存することが知られている．そのため，日本語とベトナム語 (越語) のような対訳コーパスが少ない低資源言語の翻訳においては，翻訳品質が低くなってしまう問題がある．そこで，本論文では，日本語から越語 (日-to-越)，越語から日本語 (越-to-日) 等 の低資源言語の翻訳に焦点を当て，他の言語資源が豊富な言語間の対訳コーパスを活用することにより翻訳品質を向上させる手法を提案する．提案手法においては，低資源言語を含む小規模な対訳文集合に加えて，言語資源が豊富な言語間の大規模対訳文集合を多種類同時に用いて NMT モデルを訓練する Many-to-One 翻訳 [1]，および，Many-to-Many 翻訳 [1] のアプローチを採用する．本論文では特に，NMT における Many-to-One 翻訳・Many-to-Many 翻訳アプローチにおいて，NMT モデルの訓練において同時に用いられる訓練事例中の複数の言語間で翻訳パラメータの共有が行われることにより，低資源言語文の翻訳精度を大幅に改善できることを示す．さらに，本論文では，事前訓練型多言語言語モデルである Multilingual-BERT (mBERT) [5] をあわせて適用することにより [15]，低資源言語文の翻訳精度を有意に改善することを示す．

具体的には，本論文では，まず，Many-to-One 翻訳アプローチを適用する場合においては，例えば，日-to-越翻訳の場合には，Asian Language Treebank Parallel Corpus (ALT) [10] の日-to-越対訳文集合 1.8 万文に対して，ドメインの異なる TED Talks Parallel Corpus (TED Talks) [2] の英-to-越対訳文集合 19.9 万文を追加した対訳文集合を訓練事例として，単一の NMT モデルを訓練した．これにより，ALT の日-to-越翻訳の翻訳精度が大幅に向上することを示す．次に，Many-to-Many 翻訳アプローチを適用する場合においては，ALT の日-to-越対訳文集合 1.8 万文に加えて，TED Talks の日-to 英対訳文集合 19.9 万文を訓練事例として単一の NMT モデルを訓練した．これにより，ALT の日-to-越翻訳の翻訳精度が大幅に向上することを示す．最後に，NMT モデルのエンコーダ・デコーダにおける単語埋め込みとして，事前訓練型多言語言語モデルである mBERT [5] の出力を利用する手法 [15] を，Many-to-One 翻訳 [1] および Many-to-Many 翻訳アプローチ [1] に適用した．mBERT は越語，日本語を含む豊富な多言語の非対訳単言語コーパスで事前に訓練されたモデルであり，この多言語言語モデルの出力を単語埋め込みとして利用することにより，Many-to-One 翻訳および Many-to-Many 翻訳の翻訳精度がさらに改善することを示す．

2 関連研究

低資源言語の翻訳において，当該言語の対訳資源以外を用いる手法としては，異なる翻訳方向の 1-to-1 対訳文対を混合して単一モデルを訓練することにより，異なる言語・翻訳方向の間でパラメータを共有する Many-to-One/One-to-Many/Many-to-Many 翻訳アプローチ [12, 1]，および，事前訓練済みモデルに対して，転移学習あるいは fine-tuning 等，何らかの手法を用いることにより，(低資源言語等の)NMT タスクにおいて利用するアプローチ [16, 4, 11, 15, 8] が知られている．

Many-to-One/One-to-Many/Many-to-Many 翻訳アプローチの一つとして，Tan ら [12] は，Aharoni ら [1] の Many-to-One/One-to-Many 翻訳において，語族および

*Translation of Low Resource Languages by Many-to-Many NMT using Multilingual BERT as Word Embeddings

†Tomoya Isobe, Yizhen Wei, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Takuya Tamura, College of Engineering Systems, University of Tsukuba
§Masaaki Nagata, NTT Communication Science Laboratories, NTT Corporation, Japan

単語埋め込みの類似した言語間で翻訳パラメータを共有することが有効であることを示した．これに関連して，本論文では，Many-to-One/Many-to-Many 翻訳において，対訳資源が豊富な言語対における異分野の大規模対訳文集合を利用することにより，低資源言語を含む日-to-越翻訳，および，越-to-日翻訳の翻訳精度が改善できることを示す．

一方，事前訓練済みモデルを (低資源言語等の)NMT タスクに適用するアプローチの研究事例として，Zoph ら [16] は，大規模な仏英対訳コーパスを用いて訓練されたモデルに対して，小規模なウズベク語-英語対訳コーパスを用いた転移学習により，ウズベク語から英語への翻訳精度が改善することを示した．Dabre ら [4] は，同じ語族に属する言語の対訳コーパスを用いて事前訓練されたモデルを転移学習に用いることが有効であることを示した．これらのアプローチの研究の一つの方向性として，近年では，大規模言語資源を用いて事前訓練したモデルを NMT タスクに対して利用するアプローチの研究が知られている．それらのモデルのアーキテクチャは，mBERT [5]，XLM [6]，XLM-R [3] 等，エンコーダのみで構成されるもの，および，mBART [8]，MARGE [7]，mT5 [14]，MASS [11] 等，エンコーダ・デコーダで構成されるものとに大別される．一例として，例えば，Zhu ら [15] においては，Transformer のエンコーダ・デコーダにおいて，多言語の非対訳コーパスによって事前訓練された mBERT の出力する単語埋め込みを用いる BERT-fused NMT の枠組みを提案し，翻訳精度が改善することを示した．また，同様に，多言語の非対訳コーパスによって事前訓練された mBART [8]，および，MASS [11] 等のモデルにおいては，fine-tuning タスクの事例として NMT タスクへの適用事例が報告されている．

これらに関連して，本論文では，低資源言語を対象とした Many-to-One/Many-to-Many 翻訳のアプローチにおいて，Zhu ら [15] の BERT-fused NMT の枠組みを採用し，低資源言語を含む翻訳タスクの翻訳精度が改善することを示す．

3 データセット

3.1 Asian Language Treebank Parallel Corpus [10]

Asian Language Treebank Parallel Corpus (ALT) は，BPPT, I2R, IOIT, NECTEC, NIPTICT, PUP, UCSY, NICT の 8 機関の共同プロジェクトにより作成された対訳コーパスである．英語 Wikinews の 20,106 文を他の 12 言語[1]に翻訳して作成した 13 言語の対訳コーパスであり，18,088 文の訓練文対，1,000 文の開発文対，1,018 文の評価文対から構成されている．本論文の評価においては，表 1 に示す日本語-越語 (JV)，英語-越語 (EV)，越語-日本語 (VJ)，英語-日本語 (EJ) の訓練文対集合，および，日本語-越語 (JV)，越語-日本語 (VJ) の開発文対集合，評価文対集合を用いた．英語，越語の Tokenization には英語用の Moses Tokenizer[2]を，日本語の Tokenization には MeCab[3]を，それぞれ用いた[4]．

3.2 TED Talks Parallel Corpus [2]

TED Talks Parallel Corpus (TED Talks) は，カリフォルニアに拠点を置く TED 会議の講演の英語字幕から作成された 80 言語以上の対訳コーパスである．本論文の評価においては，表 1 に示す英語-越語 (EV)，日本語-英語 (JE)，中国語-英語 (ZE)，英語-日本語 (EJ)，英語-中国語 (EZ) の対訳文集合を用いた．英語，越語の Tokenization には英語用の Moses Tokenizer を，日本語の Tokenization には MeCab を，中国語の Tokenization には Jieba[5]を，それぞれ用いた．

4 Many-to-One 翻訳

本論文では，表 1 に示す対訳文集合のうち，日本語，および，越語を目的言語とする対訳文集合を対象として，Google Many-to-One 翻訳手法 [1] を適用した．Google Many-to-One モデルにおいては，表 1 中の対訳文集合のうち，目的言語が共通となる対訳文集合のみをまとめて使用し，一つの NMT モデルを訓練する．具体的には，ALT 日-to-越対訳文集合 1.8 万文を用いて訓練した翻訳モデル (図 1(a) 左)，および，ALT 越-to-日対訳文集合 1.8 万文を用いて訓練した翻訳モデル (図 1(b) 左) をベースラインとすると，本論文の Google Many-to-One モデル [1] の一例は，図 1(a) 中，および，図 1(b) 中となる[6]．これらのモデルの訓練時，および，翻訳評価時には，目的言語が越語の場合には “<2VI>” タグを，目的言語が日本語の場合には “<2JA>” タグを，それぞれ原言語文の先頭に付与して，モデルの訓練・翻訳評価を行う．

[1]ベンガル語，フィリピン語，ヒンディー語，インドネシア語，日本語，クメール語，ラオス語，マレー語，ミャンマー語，タイ語，ベトナム語 (越語)，中国語．

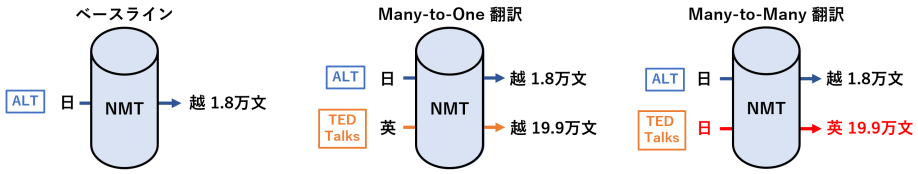
[2]https://github.com/moses-smt/mosesdecoder/

[3]https://taku910.github.io/mecab/

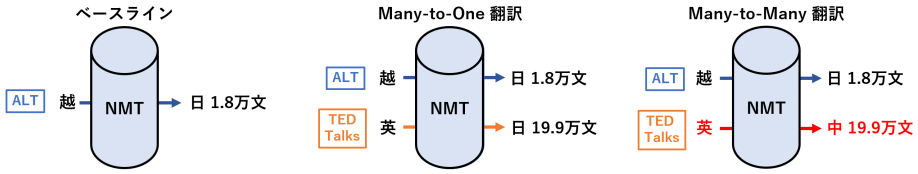
[4]越語の文においては，漢字の発音がアルファベット表記されて漢字一文字分ごとに空白で分かち書きされたものや英語由来の単語が混在していることをふまえて，英語用の Moses Tokenizer をそのまま用いる．

[5]https://github.com/fxsjy/jieba

[6]例えば，図 1(a) 中のモデルでは，ALT 日-to-越対訳文集合 1.8 万文に加えて，他ドメインである TED Talks 英-to-越対訳文集合 19.9 万文をまとめて使用して一つの NMT モデルを訓練する．



(a) 日-to-越 翻訳モデル



(b) 越-to-日 翻訳モデル

図 1: Many-to-One / Many-to-Many 翻訳

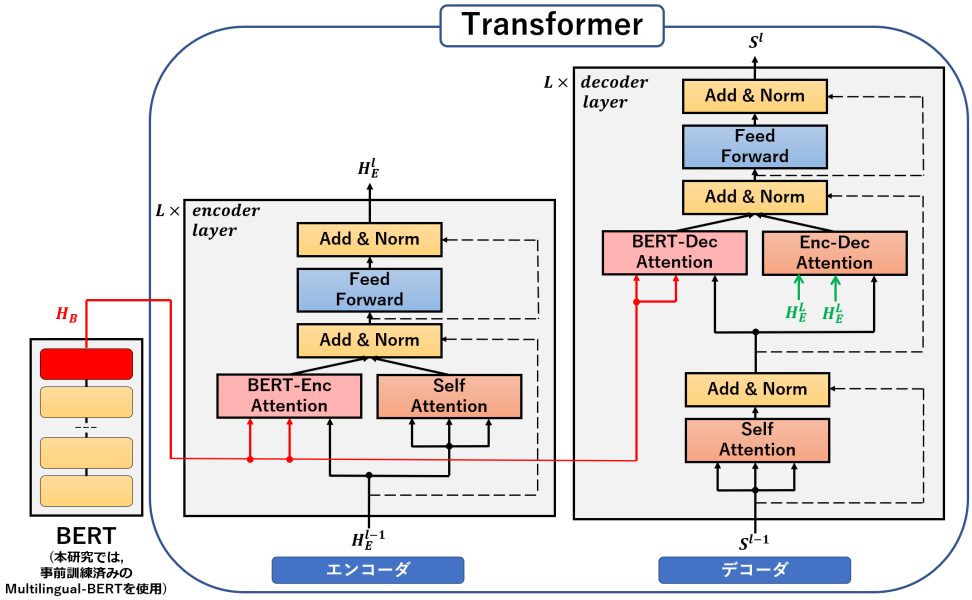


図 2: BERT-fused NMT [15]

5 Many-to-Many 翻訳

Google Many-to-Many 翻訳手法 [1] を適用する際には，表 1 に示す対訳文集合のうち，評価文対の目的言語とは異なる言語を目的言語とする対訳文集合も追加して一つの NMT モデルを訓練する．具体的には，本論文の Google Many-to-Many モデル [1] の一例は，図 1(a) 右，および，図 1(b) 右となる[7]．前節の Google Many-to-One モデルの場合と同様に，これらのモデルの訓練時，および，翻訳評価時には，目的言語が越語の場合には “<2VI>” タグを，目的言語が日本語の場合には “<2JA>” タグを，それぞれ原言語文の先頭に付与して，モデルの訓練・翻訳評価を行う．この Google Many-to-Many モデルの評価においては，低資源言語である越語を含む翻訳タスクにおいて，評価対象の ALT コーパスの分野とは異分野であり，かつ，低資源言語である越語以外の言語のみから構成される訓練用対訳文を混合することによって，翻訳精度がどこまで改善可能かの評価を行うことが最大の目的となる．

[7]例えば，図 1(a) 右のモデルでは，ALT 日-to-越対訳文集合 1.8 万文に加えて，他ドメインである TED Talks 日-to-英対訳文集合 19.9 万文をまとめて使用して一つの NMT モデルを訓練する．

6 BERT-fused NMT [15]

Many-to-One 翻訳，および，Many-to-Many 翻訳に対して，BERT-fused NMT [15] の実装を用いて事前訓練済み多言語言語モデル mBERT を適用した．図 2 に示す BERT-fused NMT のネットワーク構成は，図中青枠の通常の Transformer [13] に対して，事前訓練済み多

言語言語モデル mBERT が出力する単語埋め込み，および，BERT-Enc Attention 機構，BERT-Dec Attention 機構を加えた構造となっている．

7 評価

評価においては，fairseq ツールキット [9] 上で実装された Transformer モデル [13] を用いた[8]．評価結果を表 2 に示す[9]．図 1(a) 左 (日-to-越翻訳)，および，図 1(b) 左 (越-to-日翻訳) のベースラインの翻訳精度 (BLEU) は，それぞれ，4.73，および，5.39 と極めて低いが，Many-to-One 翻訳・Many-to-Many 翻訳いずれにおいても，評価文集合とは異分野である TED Talks コーパスの 19.9 万文対の訓練文対を混合して NMT モデルを訓練することにより，BLEU が 10〜15 程度改善することが分かる．特に，図 1(a) 右 (日-to-越翻訳)，および，図 1(b) 中右 (越-to-日翻訳) の設定においては，評価文集合とは異分野である TED Talks コーパスにおいて，低資源言語である越語が含まれなくても，BLEU が 10〜15 程度改善している．また，図 1(a) 右 (日-to-越翻訳)，および，図 1(b) 右 (越-to-日翻訳) の Many-to-Many 翻訳の設定においては，評価文集合とは目的言語・分野が異なる訓練文対集合を混合したとしても，BLEU が 10〜15 程度改善しており，提案手法の有効性を示している．さらに，表 2 において，評価文集合とは異分野である TED Talks コーパスを混合した場合においては，いずれの場合においても，BERT-fused NMT [15] によって Transformer の翻訳精度を有意 ($p < 0.05$) に改善しており，Many-to-One 翻訳・Many-to-Many 翻訳の枠組みにおいて BERT-fused NMT [15] を適用する本論文のアプローチが有効であることを示している．

8 おわりに

本論文では，Many-to-One 翻訳・Many-to-Many 翻訳アプローチ [1] によって，日-to-越，越-to-日の低資源言語翻訳の翻訳精度を大幅に改善できることを示した．さらに，このアプローチにおいて BERT-fused NMT [15] を適用することによって，低資源言語の翻訳精度が有意に改善することを示した．今後の課題としては，低資源言語を対象とした Many-to-One・Many-to-Many 翻訳において，エンコーダ・デコーダ型アーキテクチャによって構成される事前訓練済みモデル [8, 14] に対する fine-tuning のアプローチを適用することが挙げられる．また，中国語・英語由来の単語が多く含まれる越語に対する本論文の評価結果をふまえて，越語以外の低資源アジア言語と日本語の間の翻訳に対して，本論文の手法の有効性を評価することが挙げられる．

表 1: 訓練用対訳文集合

ID	コーパス	原言語	目的言語	文対数
JV	ALT [10]	日本語	越語	1.8 万
EV		英語	越語	
VJ		越語	日本語	
EJ		英語	日本語	
EV	TED Talks [2]	英語	越語	19.9 万
JE		日本語	英語	
ZE		中国語	英語	
EJ		英語	日本語	
EZ		英語	中国語	

表 2: 評価結果 (「+ BERT fused」が Transformer に対して有意差がある ($p < 0.05$) 場合を † で示す)

(a) 日本語-to-越語翻訳

ALT	TED Talks		Transformer	+ BERT fused
JV	—		4.73	4.73
JV, EV	—		3.97	3.97
	EV		20.45	21.81†
	EV	JE	21.08	**23.61**†
		ZE	18.66	20.89†
		EJ	17.99	20.62†
		EZ	20.60	21.78†
JV	EV		**19.60**	**21.40**†
	EV	JE	**21.51**	**22.95**†
		ZE	20.49	21.84†
		EJ	21.00	22.32†
		EZ	21.16	21.77†
JV	JE		**19.15**	**22.86**†
	ZE		18.80	20.50†
	EJ		17.44	18.56†
	EZ		16.18	18.78†

(b) 越語-to-日本語翻訳

ALT	TED Talks		Transformer	+ BERT fused
VJ	—		5.39	5.39
VJ, EJ	—		6.04	6.04
	EJ		17.15	19.13†
	EJ	EV	18.09	**19.89**†
		JE	18.13	19.14†
		ZE	18.18	19.14†
		EZ	**18.32**	**19.80**†
VJ	EJ		**16.62**	**19.37**†
	EJ	EV	17.45	19.10†
		JE	**17.83**	18.65†
		ZE	17.67	19.04†
		EZ	**17.83**	19.26†
VJ	EV		13.81	16.28†
	JE		15.44	17.13†
	ZE		14.72	16.69†
	EZ		**15.51**	**17.71**†

[8]head の数を 4，エンコーダとデコーダを各 6 層，単語分散表現を 512 次元，隠れ層を 1,024 次元，ドロップアウトを 0.3，学習率を 0.0003 とし，Adam optimizer を使用した．150,000 ステップの訓練を行い，ALT コーパスに対して，開発文対 1,000 文に対する損失が最小となるモデルを選択し，評価文対 1,018 文に対して評価を行った．ハードウェアとして，TITAN RTX 24GB GPU 2 枚を使用した．BLEU スコアの評価，および，有意差検定においては，Moses デコーダのスクリプト (`multi-bleu.perl`)，および，mteval Toolkit (`https://github.com/odashi/mteval`) をそれぞれ用いた．

[9]同一条件または訓練文対のより少ない条件のもとで翻訳精度最大のものを**太字**で示す．

参考文献

[1] R. Aharoni, M. Johnson, and O. Firat. Massively multilingual neural machine translation. In *Proc. NAACL-HLT*, pp. 3874–3884, 2019.

[2] M. Cettolo, C. Girardi, and M. Federico. WIT^3: Web inventory of transcribed and translated talks. In *Proc. 16th EAMT*, pp. 261–268, 2012.

[3] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F.Guzman ´ , E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proc. 58th ACL*, pp. 8440–8451, 2020.

[4] R. Dabre, T. Nakagawa, and H. Kazawa. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proc. 31st PACLIC*, pp. 282–286, 2017.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[6] G. Lample and A. Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

[7] M. Lewis, M. Ghazvininejad, G. Ghosh, A. Aghajanyan, S. Wang, and L. Zettlemoyer. Pre-training via paraphrasing. In *Proc. 34th NeurIPS*, pp. 1–14, 2020.

[8] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 726–742, 2020.

[9] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proc. NAACL-HLT: Demonstrations*, pp. 48–53, 2019.

[10] H. Riza, M. Purwoadi, Gunarso, T. Uliniansyah, A. A. Ti, S. M. Aljunied, L. C. Mai, V. T. Thang, N. P. Thai, V. Chea, R. Sun, S. Sam, S. Seng, K. M. Soe, K. T. Nwet, M. Utiyama, and C. Ding. Introduction of the Asian language treebank. In *Proc. 19th O-COCOSDA*, pp. 1–6, 2016.

[11] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. MASS: Masked sequence to sequence pre-training for language generation. In *Proc. 36th ICML*, pp. 5926–5936, 2019.

[12] X. Tan, J. Chen, D. He, Y. Xia, T. Qin, and T.-Y. Liu. Multilingual neural machine translation with language clustering. In *Proc. EMNLP-IJCNLP*, pp. 963–973, 2019.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. 30th NIPS*, pp. 5998–6008, 2017.

[14] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

[15] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T.-Y. Liu. Incorporating BERT into neural machine translation. In *Proc. 8th ICLR*, pp. 1–18, 2020.

[16] B. Zoph, D. Yuret, J. May, and K. Knight. Transfer learning for low-resource neural machine translation. In *Proc. EMNLP*, pp. 1568–1575, 2016.