事前訓練モデルの言語拡張とその効果

Language Expansion of a Pretrained Model and its Effects

国立研究開発法人情報通信研究機構　ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター　主任研究員
今村　賢治

2004年奈良先端科学技術大学院大学情報科学研究科情報処理学専攻博士課程修了。1985年日本電信電話株式会社。2014年株式会社ATR-Trekより NICTに出向。機械翻訳の研究に従事。

国立研究開発法人情報通信研究機構　ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター　副センター長
隅田　英一郎

1999年京都大学大学院博士（工学）取得。1982年日本アイ・ビー・エム。2002年ATR。2010年NICT（2016年同フェロー）。機械翻訳の研究に従事。

1 はじめに

事前訓練モデルは、大規模なコーパスでモデルを事前に訓練しておき、それを微調整（fine-tuning）することによって、さまざまなタスクに適用させる技術である。機械翻訳を含むさまざまなタスクにおけるベースモデルとして利用され、精度向上に大きく寄与してきた。

機械翻訳では、原言語と目的言語は異なるので、事前訓練モデルも多言語で訓練されたもの（多言語モデル）が使用される。しかし、原言語または目的言語が事前訓練モデルの対象に含まれていない場合、基本的には使用することができず、何らかの対処を必要とする。

本稿では、事前訓練モデルの一つであるmBARTモデル（multilingual sequence-to-sequence denoising auto-encoder）[1]を、新たな言語対の翻訳に適用するため、対象言語を拡張し、追加訓練を行った。

モデルは、WAT-2021 [2]のNICT-SAP共有タスクで評価した。これは、英語⇔ヒンディー語、インドネシア語、マレー語、タイ語間の翻訳タスクである。なお、公開されているmBARTモデル[1]（以下、オリジナルmBARTモデルと呼ぶ）は、英語、ヒンディー語を含んでいるが、インドネシア語、マレー語、タイ語は含んでいない。

2 mBARTモデルの言語拡張

2.1 mBARTモデル

mBARTはBART（(Bidirectional and Auto-Regressive Transformers）[3]の多言語モデルである。エンコーダー・デコーダー形式のTransformer [4]で、デコーダーが自己再帰方式であることが特徴である（図1）。

mBARTは、BARTを多言語化するため、エンコーダー入力の最後尾、およびデコーダー入力の先頭に言語タグをつけて訓練、使用する。これによって、複数の言語を1つのモデルで訓練・使用することができる。

オリジナルmBARTモデルは、エンコーダー、デコー

1 https://dl.fbaipublicfiles.com/fairseq/models/mbart/mbart.cc25.v2.tar.gz

ダーともに 12 層、モデル次元数 1024、16 ヘッドである。このモデルは Common Crawl コーパス［5］のうち、25 言語の単言語コーパスを使って訓練されている。

表 1　事前訓練モデル用の単言語訓練データ

言語	文数	トークン数
英語（En）	7,000,000	1.74 億
ヒンディー語（Hi）	1,968,984	0.54 億
インドネシア語（Id）	6,997,907	1.51 億
マレー語（Ms）	2,723,230	0.57 億
タイ語（Th）	2,233,566	0.60 億

2.2　言語拡張と追加訓練

前述のように、オリジナル mBART モデルにはインドネシア語、マレー語、タイ語が含まれていない。そこで我々は、この 3 言語を含むようにモデルを拡張し、英語、ヒンディー語を含む 5 言語で追加訓練を行った。

mBART モデルの追加訓練は、オリジナル mBART large モデルの単語埋込に、ランダム初期化した言語タグを追加して、訓練した。これは、コーパスとハイパーパラメーターを除いて、mBART-50［6］の訓練手順と同じである。

事前訓練コーパスは、Wikipedia のダンプファイルから抽出したデータを使用した。コーパスサイズを表 1 に示す。抽出した文は、英語だけ突出して多く（約 1.5 億文）、他の言語と水準を合わせるため、700 万文だけサンプリングして使用した。

訓練は、Fairseq 翻訳器［7］[2] で行った。訓練時間は、NVIDIA V100 GPU8 個で約 15 日だった。

2　https://github.com/pytorch/fairseq

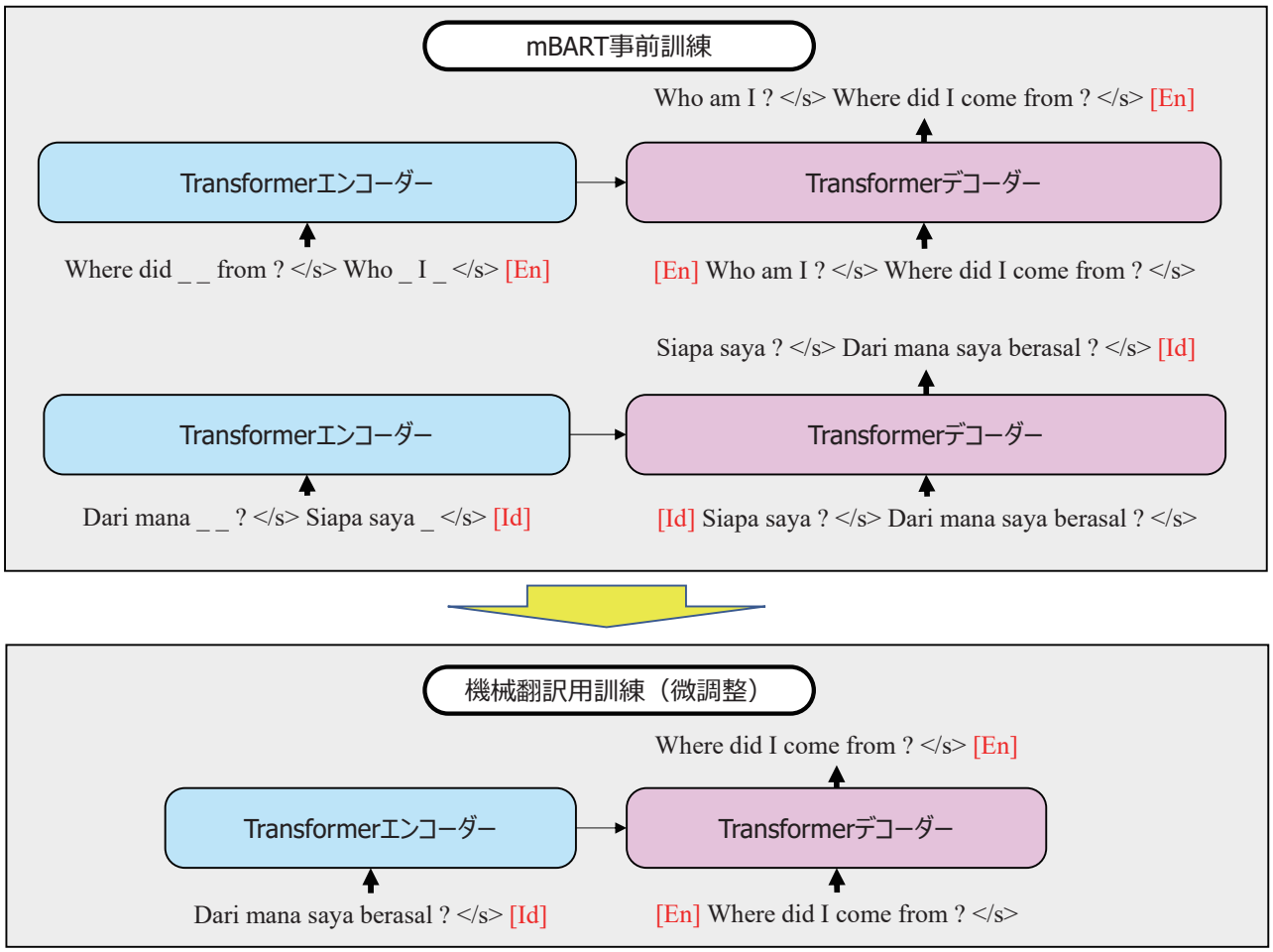

図 1　mBART の事前訓練と機械翻訳タスクへの適用例（［1］の図を改変）。インドネシア語から英語への翻訳の場合。

3 翻訳実験

言語を追加した mBART モデルは、対訳コーパスで微調整することによって、翻訳モデルとなる。今回は WAT-2021 の NICT-SAP 共有タスクで評価した。

3.1 実験設定

NICT-SAP 共有タスク［2］は、対訳が比較的少ない 4 言語（ヒンディー語（Hi）、インドネシア語（Id）、マレー語（Ms）、タイ語（Th））と英語との翻訳タスクである。これには 2 種類のドメインを含んでいる。

- ALT（Asian Language Treebank）ドメインは、WikiNews から作成した対訳データである。訓練文数は各 1 万 8 千文と、とても少ない。
- IT ドメインは、ソフトウェア文書の対訳データである。訓練データは言語対によって 7 万文から 50 万文程度である。

なお、訓練セットには、対訳として不適切と思われるデータが含まれていたため、それらを削除して使用した。

翻訳システムは、事前訓練と同じく Fairseq を使った。比較システムは、事前訓練モデル未使用（Transformer ベースモデル）と、事前訓練モデルの単語埋め込みに、言語タグを追加しただけもの（オリジナル）の 2 種類を使用した。

3.2 実験結果

表 2 は、ALT ドメイン、IT ドメインでの、WAT 公式 BLEU スコア［8］である。これを見ると、言語拡張と追加訓練を行った事前訓練モデルは、すべての言語対において、事前訓練未使用の BLEU スコアを上回っている。

オリジナル mBART モデルに含まれていなかったインドネシア語、マレー語、タイ語に着目しても、すべての言語対において、事前訓練モデル（言語拡張）は 10 ポイント以上、事前訓練モデルなしを上回った。したがって、新しい言語対の翻訳には言語拡張・追加訓練が有効である。

オリジナル mBART モデルと言語拡張して追加訓練した mBART モデルを比べた場合、ほとんどの場合、追加訓練によって翻訳品質が若干向上した。元の mBART モデルが Common Crawl で訓練されていたのに対して、本稿の拡張モデルは、Wikipedia で追加訓練しているため、モデルが許容する言語が増えたとも考えられるが、これについては、さらなる検証が必要であると考えている。

なお、テストセットと表 1 の事前訓練コーパスを比較したところ、ALT ドメインでは同一文はまったく含まれていなかった（IT ドメインでは、言語によって 0 ～ 10% の同一文が存在）。したがって、この品質向上は、事前訓練モデルがテストセット文を記憶したためではない。

4 おわりに

本稿では、事前訓練モデルの一つである mBART に対して、言語を拡張し、追加訓練を行った。事前訓練モデル自体は、小規模リソースの翻訳品質改善に大きく寄与するが、拡張言語に対しても、同様に翻訳品質が改善した。言語拡張・追加訓練は、新しい言語の翻訳に有効である。

表 2　WAT-2021 における公式 BLEU スコア

ドメイン	事前訓練モデル	ヒンディー語		インドネシア語		マレー語		タイ語	
		En→Hi	Hi→En	En→Id	Id→En	En→Ms	Ms→En	En→Th	Th→En
ALT	未使用	12.26	8.23	24.71	23.65	31.02	27.52	13.73	2.04
	オリジナル	34.26	33.62	40.45	40.62	44.27	41.93	54.95	25.56
	言語拡張＋追加訓練	**34.97**	**35.21**	**41.15**	**43.90**	**45.17**	**44.53**	**55.69**	**28.96**
IT	未使用	7.97	4.92	23.33	22.64	29.62	26.53	10.24	0.99
	オリジナル	27.77	34.72	42.12	39.03	40.37	37.00	**51.58**	20.00
	言語拡張＋追加訓練	**29.05**	**35.32**	**43.25**	**40.69**	**40.76**	**38.42**	50.91	**21.89**

謝辞

本件は、総務省の「ICT 重点技術の研究開発プロジェクト（JPMI00316）」における「多言語翻訳技術の高度化に関する研究開発」による委託を受けて実施した研究開発による成果です。

参考文献

[1] Liu, Y., et al., 2020. Multilingual Denoising Pre-training for Neural Machine Translation. arXiv 2001.08210.

[2] Nakazawa, T. et al., 2021. Overview of the 8th Workshop on Asian Translation. In Proc. of WAT 2021.

[3] Lewis, M., et al., 2020. BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension. In Proc. of ACL 2020, pp. 7871–7880.

[4] Vaswani, A. et al., 2017. Attention is All You Need. CoRR, abs/1706.03762.

[5] Wenzek, G. et al., 2019. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. arXiv 1911.00359.

[6] Tang, Y. et al., 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. arXiv 2008.00401.

[7] Ott, M. et al., 2019. Fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In Proc. of NAACL-2019 (Demonstrations), pp. 48–53.

[8] Papineni, K. et al., 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proc. of ACL-2002, pp. 311–318.