多言語事前訓練機械翻訳の語彙選択による高速化

Speeding up of Multilingual Pretrained Machine Translation by Vocabulary Selection

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。

1　はじめに

ニューラルネットワークの事前訓練モデルは、低リソース環境での自然言語処理タスクの精度向上に有効である。機械翻訳でも、エンコーダ・デコーダ型の多言語事前訓練モデルがリリースされ、低リソース言語の翻訳品質向上に役立っている。多言語事前訓練モデルは、大量のコーパスで事前訓練されているため、一般的にはモデルも大規模になっている場合が多い。

1.1　mBART-50 モデル

たとえば、本稿で対象とするmBART-50[1] [1]はエンコーダ・デコーダモデルで、52言語の単言語コーパスで訓練されている。開発者は、このモデルを自分が所有する対訳コーパスでファインチューニングすることで、翻訳器を構成する。大規模コーパスで事前訓練されているため、ファインチューニングする対訳コーパスの規模が小さくても、比較的高精度にしやすい。

mBART-50は、52言語を取り扱うため、モデルは、12層、分散表現次元数768で、入出力の語彙（サブワード[2]の種類数。単語埋込のエントリ数と同じ）は25万語と、非常に大きなものを使用している。これは、モデルサイズ増大と、処理時間増大の原因となっている。なお、Transformerの基本モデル[3]は、6層、512次元で、単語埋込は数万の場合が多い。

1　https://dl.fbaipublicfiles.com/fairseq/models/mbart50/mbart50.pretrained.tar.gz

1.2　本稿の目的

しかし、翻訳対象の言語対が確定した場合、対象言語以外の単語埋込エントリは翻訳で使用されない。これを削減すると、メモリ使用量が削減されるとともに、デコーダ出力のSoftMax操作の処理量も削減され、翻訳速度が向上する。

図1は、mBART-50（実体はTransformer）の構成図である。語彙に関連する部分は、図の赤枠部分（エンコーダ単語埋込、デコーダ単語埋込、デコーダ出力への線形変換）の3箇所である。本稿では、mBART-50をベースに日英・英日翻訳を構築する際、語彙を制限することで上記単語埋込を縮小し、どの程度翻訳速度が向上するか、モデルパラメータが減少するのか、検証を行った。

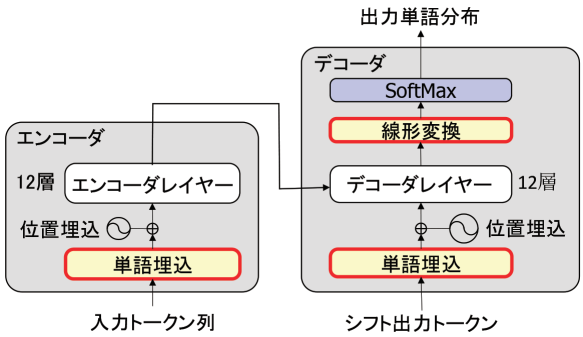

図1　mBART-50の構成と単語埋込の位置

2 実験

2.1 実験設定

今回、日英、英日翻訳で確認した。使用したコーパスは JParaCrawl 3.0 [4] で、クリーニング後の文数は約2,500万文対である。開発・テストセットには、国際ワークショップ WMT-21 [5] で使用された Dev、Test セットを使用した。

翻訳品質は sacreBLEU [6] で測定した。翻訳速度は、1 秒あたりの処理トークン数で測定した。ただし、モデルのロード、初期化時間は含まない。

mBART-50 は Fairseq 翻訳器 [7] を基にしており、今回はこれを使用した。mBART-50 のファインチューニングモデルをベースとして、日英、英日の各方向に対して、デコーダの語彙のみ縮小した場合、エンコーダ、デコーダ双方の語彙を縮小した場合について実験を行った。mBART-50 の 25 万語彙のうち、JParaCrawl に出現する語彙（サブワード）は日本語 7 万 7 千語、英語 8 万 5 千語である。

なお、比較のため Transformer 基本モデルでも翻訳品質、翻訳速度を測定した。日本語・英語の語彙数は、どちらも 32 万サブワードとした。

2.2 実験結果

各条件でのモデルパラメータ数、BLEU スコア、翻訳速度の測定結果を表 1 に示す。最もパラメータ数が少なく、かつ翻訳速度が高速なのは、Transformer 基本モデルである。mBART50 はモデルが巨大であるため、翻訳速度も遅くなるが、BLEU スコアが英日で 2.3 ポイント、日英で 1.0 ポイント程度向上し、多言語事前訓練モデルは JParaCrawl にも効果があることがわかる。

デコーダおよびエンコーダの語彙を制限（単語埋込を縮小）した場合、BLEU スコアをほとんど変えることなく、翻訳速度を英日で 1.7 倍、日英で 1.6 倍以上高速化した。デコーダだけでなく、エンコーダの語彙も制限する方が、モデルパラメータは減少するが、翻訳速度にはほとんど影響がない。エンコーダの語彙は、そのまま翻訳可能な入力となるため、メモリ使用量を気にせずに翻訳速度を向上させたい場合は、エンコーダ語彙を無理に制限する必要はない。

まとめると、mBART-50 の語彙を適切に制限すると、翻訳品質を変えずに翻訳の高速化と、モデルパラメータを削減することができる。

3 おわりに

大規模な多言語事前訓練モデルは、特に低リソース環境での機械翻訳の精度向上に効果的であるが、言語対を限った場合、冗長なパラメータが含まれている。

本稿では、デコーダの単語埋込を対象言語に限ることで、翻訳品質を変えずに翻訳速度を向上させることができることを確認した。また、エンコーダの単語埋込を限定することで、パラメータ数も減少させることができるが、翻訳器が翻訳可能な語彙を制限することになるため、こちらの方は注意が必要である。

表 1　mBART-50 の語彙縮小時の翻訳品質と翻訳速度

言語対	モデル	パラメータ数	BLEU	翻訳速度（tokens/s）
英日	Transformer 基本モデル	93M	22.7	2,060
	mBART-50	611M	25.0	602
	デコーダ単語埋込縮小	690M	24.9	1,120
	エンコーダ・デコーダ単語埋込縮小	521M	24.9	1,070
日英	Transformer 基本モデル	93M	20.3	2,090
	mBART-50	611M	21.3	595
	デコーダ単語埋込縮小	698M	21.2	971
	エンコーダ・デコーダ単語埋込縮小	521M	21.2	974

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

[2] Sennrich, R. et al., 2016. Neural Machine Translation of Rare Words with Subword Units. In Proc. of ACL-2016, pp.1715-1725.

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

[4] 森下他 , 2022. JParaCrawl v3.0: 大規模日英対訳コーパス . 言語処理学会第 28 回年次大会 , pp. 1750-1755.

[5] Akhbardeh, F. et al., 2021. Findings of the 2021 Conference on Machine Translation（WMT21）. In Proc. of WMT-21, pp. 1-88.

[6] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In Proc. of WMT-18, pp. 186-191.

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