対訳データを用いた継続事前訓練による大規模言語モデルの翻訳精度評価

近藤海夏斗 [1]　宇津呂武仁 [1]　森下睦 [2]　永田昌明 [2]
[1] 筑波大学大学院 システム情報工学研究群　[2]NTT コミュニケーション科学基礎研究所
s2320743_@_u.tsukuba.ac.jp　utsuro_@_iit.tsukuba.ac.jp
{makoto.morishita, masaaki.nagata}_@_ntt.com

概要

大規模言語モデルが，多くの自然言語処理タスクで成果を収めている．しかし，パラメーター数が100 億前後の大規模言語モデルでは，既存手法である Encoder-Decoder モデルより，翻訳精度が大きく劣ることが報告されている．そこで本論文では，対訳データを用いた継続事前訓練を提案する．対訳データで大規模言語モデルを継続事前訓練した後，少量の人手データで Supervised Fine-Tuning し，WMT22 のテストデータをはじめとする 12 種のテストセットで評価した．その結果，提案手法を適用したモデルは，対訳データで訓練された Encoder-Decoder モデルより BLEU・COMET の両方で統計的に有意な差が示された．

1　はじめに

機械翻訳では，2017 年に Google から Transformer [1] が提案されて以降，Transformer ベースの Encoder と Decoder で構成されたモデルが使用されている．一方で，GPT [2,3] をはじめとする大規模言語モデルは，Transformer の Decoder のみで構成されている．そして，XGLM [4]，PaLM [5]，そして LLaMA-2 [6] など多言語で事前訓練された大規模言語モデルが提案された．しかし，パラメーター数が 100 億前後の大規模言語モデルの翻訳精度は，既存手法である Encoder-Decoder モデルに大きく劣ることが報告されている．例えば，日英翻訳において XGLM-7.5B の 8-shot は，対訳データで訓練された Encoder-Decoder モデルである NLLB-1.7B と比較すると，BLEU で 25 ポイント以上，COMET で 34 ポイント以上低いスコアであることが報告されている [7]．

この問題に対し本論文では，対訳データを用いた継続事前訓練を提案する．大規模言語モデルを対訳データで継続事前訓練し，少量の人手データで Supervised Fine-Tuning(以下 SFT という) を行ったモデルを，WMT22 [8] のテストデータをはじめとする 12 種のテストセットで評価した結果，対訳データで訓練された Encoder-Decoder モデルより BLEU・COMET の両方で統計的に有意な差が示された．

2　関連研究

Briakou ら [9] は，翻訳例などの偶発的な bilingual データが，大規模言語モデルの翻訳性能に与える影響を調査した．具体的には，パラメーター数が 10 億および 80 億の大規模言語モデルに対し，事前訓練データにおける bilingual データの有無による翻訳性能の影響について調査した．その結果，bilingual データが存在することで，大規模言語モデルの翻訳性能が向上することが明らかとなった．

また，Xu ら [10] は，LLaMA-2 に対して 2 段階の fine-tuning を行う ALMA という手法を提案した．ALMA は第 1 段階目に単言語データを用いた fine-tuning，そして第 2 段階目に少量の高品質な対訳データで fine-tuning を行う．この ALMA により，SoTA モデルに匹敵する翻訳精度を達成した．

そして，Zhou ら [11] は，大規模言語モデルのほぼ全ての知識は事前訓練中に学習され，モデルから高品質の出力を得るために必要なのは限られた instruction tuning データのみであるという仮説を提唱した．この仮説を検証するため，LLaMA-2 を人手で作成された 1,000 件のデータのみで instruction tuning した LIMA というモデルを提案した．LIMA はたった 1,000 件のデータで訓練されたのにもかかわらず，RLHF で訓練されたモデルより優れた性能を示した．

本論文ではこれらの報告に着目し，対訳データを用いた継続事前訓練を提案する．

3 対訳データを用いた継続事前訓練

事前訓練済みの大規模言語モデルをそのまま対訳データで SFT すると，既存の翻訳モデルより翻訳精度が劣ることが報告されている [10]．本論文でも予備実験を行った結果，同様の現象が確認された．[1] この現象の原因として，Zhou らが提唱した仮説をもとに，単言語データで事前訓練された大規模言語モデルは，翻訳に必要な知識を有していないという仮説を提唱する．

この仮説をもとに本論文では，対訳データを用いて継続事前訓練することを提案する．この継続事前訓練により，大規模言語モデルが翻訳に必要な知識を獲得することを期待する．そして，継続事前訓練したモデルに対して SFT を行うことで，対訳データで訓練された Encoder-Decoder モデルを凌駕したことを報告する．

4 評価手順

4.1 概要

今回の評価実験では，単言語データで事前訓練された大規模言語モデルとして `rinna/bilingual-gpt-neox-4b` [2] (以下 rinna-4b という) を使用した．このモデルに対して，対訳データを用いた継続事前訓練を行った後，SFT を行った際の翻訳性能を評価した．また，baseline として，JParaCrawl v3.0 で　訓練された transformer モデル [12] を使用した．さらに，rinna-4b をそのまま SFT した場合についても評価を行った．なお，SFT では full fine-tuning と LoRA [13] チューニングの両方とも評価を行った．推論時は，開発データの誤差が最小となるモデルを用いて greedy decoding で生成した．

4.2 データセット

4.2.1 継続事前訓練

継続事前訓練のデータとして，JParaCrawl v3.0 [12]，開発データとして，WMT20 [14] の開発，テストデータ，および WMT21 [15] のテストデータを使用した．なお，継続事前訓練で使用する JParaCrawl v3.0 は，LEALLA-large[3] [16] で取得した文埋め込みベクトルのコサイン類似度をもとに，2,080 万文対をサンプリングした．[4]

本論文では，対訳データを用いた継続事前訓練の最適なデータ形式を調査するため，以下 5 通りの継続事前訓練を行った．

1. monolingual: 対訳データを日本語と英語の単言語データとみなす．
2. En-Ja: 英文 1 サンプルの直後に，和訳文を結合する．すなわち，英日方向のみ対訳となる．
3. Ja-En: 和文 1 サンプルの直後に，英訳文を結合する．すなわち，日英方向のみ対訳となる．
4. mix: 2. および 3. で作成したデータから，重複しないよう 50%ずつランダムサンプリングする．
5. En-Ja2mix: 2. の形式で継続事前訓練したモデルを，3. のデータを 99%，2. のデータを 1%重複しないようランダムサンプリングしたデータでさらに継続事前訓練する．[5]

4.2.2 SFT

SFT の訓練データは，WMT20 および Flores-200 [18] の開発，テストデータ，そして KFTT [19] の訓練データから作成した．[6] なお，KFTT の訓練データは，全データから 10,000 件をランダムサンプリングした．作成した訓練データの数は，英日と日英それぞれ約 15,000 件ずつとなった．また，SFT の開発データは WMT21 のテストデータを使用した．これらのデータに対し，ALMA の実装[7] を参考に以下のプロンプトを適用した．本論文では，プロンプトを原言語文側の言語とし，プロンプト部分の誤差は除外して訓練を行った．そしてテストセットの推論時にも同じプロンプトを使用した．

英日翻訳のプロンプト

Translate this from English to Japanese:
English: { 原言語文 }
Japanese: { 目的言語文 }

日英翻訳のプロンプト

これを日本語から英語に翻訳してください:
英語: { 原言語文 }
日本語: { 目的言語文 }

1) 詳細は付録 A を参照．
2) https://huggingface.co/rinna/bilingual-gpt-neox-4b
3) https://huggingface.co/setu4993/LEALLA-large
4) サンプリングの詳細は付録 B を参照．
5) 2. のデータを 1%含めた理由は破壊的忘却を防ぐためであり，1%という割合は既存研究 [17] を参考にした．
6) SFT では，データの品質が重要であることが報告されている [10, 11]．そのため，現存する対訳コーパスの中でも，人手で翻訳された高品質なデータを選択した．
7) https://github.com/fe1ixxu/ALMA

表 1　英日翻訳の評価結果 (BLEU / COMET)．5 つの継続事前訓練済みモデルの詳細は 4.2.1 節を参照．各データで最高スコアを**太字**，baseline を上回るスコアを下線で示す．*は baseline と有意差あり ($p < 0.05$)．

テストセット	fine-tuning 手法	baseline	original	継続事前訓練済み				
				1. monolingual	2. En-Ja	3. Ja-En	4. mix	5. En-Ja2mix
ASPEC	full	**19.8** / 88.5	5.2 / 79.0	5.1 / 78.7	19.1 / 88.6	6.6 / 80.5	18.4 / 88.1	17.6 / 87.9
	LoRA		5.5 / 79.2	4.6 / 77.1	19.0 / **88.7**	6.4 / 80.7	18.5 / 88.2	17.2 / 87.9
JESC	full	6.2 / 72.6	3.6 / 71.9	3.6 / 71.4	**7.4*** / **76.0***	4.3 / 72.4	**7.4*** / 75.7*	7.3* / 75.1*
	LoRA		3.7 / 71.9	3.4 / 70.8	7.3* / 75.7*	3.9 / 72.4	7.0* / 75.8*	6.8* / 75.5*
KFTT	full	10.2 / 82.4	6.8 / 76.4	6.7 / 76.4	**15.5*** / **84.4***	7.1 / 76.4	14.1* / 83.3*	13.5* / 82.8
	LoRA		6.8 / 76.4	6.1 / 75.7	15.0* / 84.3*	6.3 / 76.1	13.5* / 83.7*	12.4* / 82.6
TED (tst2015)	full	11.4 / 78.9	5.4 / 77.2	5.4 / 76.4	12.7* / **83.6***	6.7 / 78.1	12.3* / 83.0*	11.8 / 82.7
	LoRA		5.3 / 76.6	5.1 / 75.1	12.8* / 83.2*	6.3 / 77.7	**12.9*** / 83.0*	11.9* / 82.5*
Business Scene Dialogue Corpus	full	12.6 / 85.5	7.6 / 81.7	7.5 / 81.3	14.4* / 87.5*	8.6 / 82.9	14.1* / 87.0*	13.6* / 86.8*
	LoRA		8.0 / 81.8	7.5 / 81.2	**15.5*** / **87.7***	8.6 / 83.1	15.2* / 87.4*	14.2* / 87.1*
WMT19 Robustness En-Ja	full	13.9 / 76.6	6.7 / 75.1	6.0 / 74.2	**15.2*** / **81.7***	6.7 / 75.3	14.4 / 81.2*	14.2 / 80.5*
	LoRA		6.1 / 74.4	5.5 / 73.1	15.1* / 81.5*	6.7 / 75.1	14.7* / 81.0*	14.3 / 80.9*
WMT19 Robustness Ja-En	full	12.3 / 79.2	6.9 / 75.7	5.9 / 75.0	13.7* / **82.5***	7.4 / 76.1	**14.2*** / 81.7*	**14.2*** / 81.4*
	LoRA		6.6 / 75.3	5.4 / 73.9	13.5* / 81.8*	7.1 / 76.2	13.6* / 81.8*	13.1 / 81.2*
WMT20 Robustness Set1 En-Ja	full	17.6 / 67.1	7.7 / 66.5	7.1 / 64.6	18.4* / **76.8***	7.5 / 65.6	17.5 / 76.2*	15.3 / 73.7*
	LoRA		7.7 / 67.1	6.5 / 63.7	**19.5*** / 76.3*	8.0 / 66.3	18.7* / 75.4*	17.2 / 75.2*
WMT20 Robustness Set2 En-Ja	full	13.8 / 75.5	6.1 / 74.3	5.5 / 72.7	**14.8*** / **81.6***	6.8 / 74.8	13.9 / 80.6*	13.4 / 79.6*
	LoRA		5.2 / 73.5	4.9 / 72.2	**14.8*** / 81.1*	6.8 / 74.6	13.7 / 80.4*	13.7 / 80.2*
WMT20 Robustness Set2 Ja-En	full	7.3 / 79.9	3.7 / 76.6	3.7 / 75.8	9.3* / **83.2***	4.3 / 76.6	9.3* / 82.9*	8.8* / 82.0*
	LoRA		3.6 / 76.1	3.6 / 75.4	**9.8*** / **83.2***	4.0 / 77.1	9.0* / 82.8*	8.5* / 82.2*
IWSLT21 En-Ja Dev	full	12.8 / 82.6	5.6 / 80.6	5.3 / 79.4	**13.2** / **86.2***	6.6 / 81.3	12.8 / 85.8*	12.3 / 85.3*
	LoRA		5.7 / 80.0	5.2 / 78.8	**13.2** / 85.9*	7.0 / 81.2	12.8 / 85.8*	12.6 / 85.5*
WMT22 News En-Ja	full	21.8 / 85.4	10.3 / 81.3	11.0 / 80.8	**23.1*** / **88.3***	11.9 / 82.1	22.0 / 87.8*	21.2 / 87.0*
	LoRA		10.9 / 81.8	9.8 / 79.8	**23.1*** / **88.3***	11.7 / 81.0	22.1 / 87.9*	21.3 / 87.6*

4.2.3　テストセット

SFT を行ったモデルの翻訳性能を評価するため，JParaCrawl v3.0 の評価で用いられたテストセットを使用した．なお，WMT20 および WMT21 のテストデータは，継続事前訓練および SFT の訓練・開発データに含まれるため除外し，WMT22 のテストデータを追加した．これにより，テストセットは全 12 種となった．[8)]

4.3　ハイパーパラメータ

以下にハイパーパラメータを一部抜粋して示す．より詳細な設定は付録 C に示す．

4.3.1　継続事前訓練

継続事前訓練では，optimizer として AdamW [20] を使用した．そして，通常の事前訓練と同様に固定長 2,048 トークンをモデルへ入力し，次単語を予測するよう訓練を行った．また，エポック数を 1，バッチサイズを 256 とした．

4.3.2　SFT

SFT も継続事前訓練と同様に，optimizer として AdamW を使用した．また，エポック数を 5，バッチサイズを 64 とした．また，LoRA チューニングでは，学習可能パラメータは約 640 万となり，オリジナルモデルのパラメータ数の約 0.17%となった．

4.4　評価指標

評価指標として，BLEU [21] および COMET[9)] [22] を使用した．BLEU は sacreBLEU[10)] [23] を用いて計測した．COMET のモデルは `wmt22-comet-da` を使用した．

8)　テストセットの詳細は付録 D を参照．

9)　https://github.com/Unbabel/COMET

10)　https://github.com/mjpost/sacrebleu

表 2　日英翻訳の評価結果 (BLEU / COMET).

テストセット	fine-tuning 手法	baseline	original	継続事前訓練済み				
				1. monolingual	2. En-Ja	3. Ja-En	4. mix	5. En-Ja2mix
ASPEC	full	**21.4 / 82.8**	8.9 / 75.0	8.8 / 74.8	9.4 / 75.3	20.3 / 82.5	19.1 / 81.9	20.0 / 82.3
	LoRA		8.0 / 74.4	7.9 / 74.2	8.5 / 74.8	20.4 / 82.5	19.4 / 82.1	19.9 / 82.4
JESC	full	**9.2** / 68.1	4.5 / 64.6	4.6 / 64.3	4.1 / 64.3	8.5 / 69.2*	7.9 / 68.7*	8.6 / 69.1*
	LoRA		4.0 / 64.0	4.0 / 63.2	4.3 / 63.5	8.8 / **69.3***	7.7 / 68.6*	8.3 / 69.0*
KFTT	full	17.2 / 74.5	10.0 / 71.1	9.6 / 70.1	10.6 / 70.5	**19.9*** / **78.2***	18.5* / 77.4*	**19.9*** / 77.9*
	LoRA		8.8 / 69.9	8.4 / 69.4	9.7 / 70.3	19.0* / 77.8*	17.4 / 76.6*	**19.9*** / 77.8*
TED (tst2015)	full	12.3 / 75.8	7.4 / 72.0	7.4 / 71.2	6.9 / 71.2	14.7* / **78.7***	14.3* / 78.3*	**14.8*** / **78.7***
	LoRA		6.6 / 71.1	6.3 / 70.4	6.6 / 70.9	15.2* / **78.7***	14.4* / 78.1*	14.7* / **78.7***
Business Scene Dialogue Corpus	full	**20.5** / 81.4	9.0 / 74.6	9.4 / 74.4	8.5 / 74.1	20.1 / **82.9***	18.7 / 81.4	20.4 / 82.1*
	LoRA		9.0 / 74.4	8.8 / 74.1	9.2 / 74.2	20.4 / 82.0*	18.6 / 81.1	20.1 / 81.9*
WMT19 Robustness En-Ja	full	17.6 / 78.2	7.7 / 71.1	7.0 / 70.2	6.6 / 70.3	17.8 / 79.1*	17.0 / 78.5	17.8 / **79.2***
	LoRA		8.0 / 71.0	6.5 / 70.2	7.0 / 69.6	**18.3** / 79.0*	17.1 / 78.5	17.7 / **79.2***
WMT19 Robustness Ja-En	full	17.6 / 74.8	8.5 / 70.4	8.3 / 69.6	8.3 / 69.0	**18.0** / **76.8***	16.4 / 76.5*	17.2 / 76.5*
	LoRA		7.9 / 69.7	6.9 / 68.6	7.1 / 68.4	17.1 / 76.3*	16.5 / 76.2*	16.5 / 76.4*
WMT20 Robustness Set1 En-Ja	full	22.6 / 72.7	12.2 / 66.2	10.3 / 64.2	9.1 / 62.7	23.6* / 74.9*	22.6 / 74.2*	23.3* / 74.5*
	LoRA		11.2 / 66.0	10.4 / 64.0	10.5 / 63.9	**24.3*** / **75.2***	22.3 / 73.9*	24.2* / 74.7*
WMT20 Robustness Set2 En-Ja	full	**18.4** / 78.2	7.2 / 71.3	7.4 / 70.8	6.5 / 69.7	17.7 / 79.2*	16.5 / 78.5	17.8 / 79.2*
	LoRA		7.2 / 70.6	6.5 / 70.1	6.4 / 69.7	18.2 / **79.3***	16.8 / 78.5	18.2 / **79.3***
WMT20 Robustness Set2 Ja-En	full	**14.7** / 70.7	6.0 / 66.1	5.6 / 65.7	5.4 / 65.0	13.9 / 72.8*	13.0 / 72.2*	13.6 / 72.5*
	LoRA		5.3 / 65.3	5.4 / 64.8	5.4 / 65.0	14.1 / **73.0***	13.0 / 72.1*	14.3 / 72.8*
IWSLT21 En-Ja Dev	full	14.7 / 81.1	7.4 / 75.8	7.3 / 74.9	6.2 / 74.8	14.5 / 81.9*	13.8 / 81.3	14.9 / 81.9*
	LoRA		6.7 / 75.0	6.7 / 74.2	6.7 / 74.9	**15.3*** / **82.0***	14.0 / 81.5	14.9 / **82.0***
WMT22 News Ja-En	full	**21.6** / 80.1	9.8 / 73.7	9.5 / 73.4	9.3 / 72.7	20.8 / 81.0*	19.1 / 80.5*	20.9 / 81.1*
	LoRA		9.7 / 73.0	9.4 / 72.6	9.9 / 72.9	21.1 / **82.0***	19.3 / 80.4*	20.3 / 81.1*

5　評価結果

表 1，表 2 にそれぞれ英日方向，日英方向の BLEU および COMET を示す．まず，オリジナルのモデルと，対訳データを単言語データとみなして継続事前訓練したモデルをそれぞれ SFT した場合，両者に翻訳精度の違いは見られなかった．一方で，英日方向もしくは日英方向のみ対訳となるように継続事前訓練したモデルは，継続事前訓練した言語方向のみ baseline と同等もしくはそれ以上のスコアとなった．しかし，継続事前訓練で対訳になっていない言語方向については，オリジナルのモデルをそのまま SFT した場合と翻訳精度に違いは見られなかった．そして，英日・日英方向の両方とも継続事前訓練したモデルは，英日・日英の両方向とも baseline と同等もしくはそれ以上のスコアとなった．

以上のことから，対訳データを用いた継続事前訓練によって，大規模言語モデルに対して翻訳に必要な知識を学習させることが可能であることが示唆された．そして，継続事前訓練によって翻訳能力を得たモデルに対し，少量の人手データで SFT を行うことで，対訳データで訓練された Encoder-Decoder モデルを凌駕することが明らかとなった．また，継続事前訓練によって大規模言語モデルが獲得する翻訳能力は，対訳文対を結合する際の言語方向に依存することも明らかとなった．

6　おわりに

本論文では，大規模言語モデルをそのまま SFT しても，対訳データで訓練された Encoder-Decoder モデルより翻訳精度が劣るという問題に対し，対訳データを用いた継続事前訓練を提案した．この提案手法の効果を検証するため，継続事前訓練を行ったモデルを SFT し，WMT22 のテストデータをはじめとする 12 種のテストデータで評価した．その結果，提案手法を適用したモデルは，対訳データで訓練された Encoder-Decoder モデルより BLEU・COMET の両方で統計的に有意に有意な差が示された．

参考文献

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

[2] T. Brown, et al. Language models are few-shot learners. In **Proc. 33rd NIPS**, pp. 1877–1901, 2020.

[3] OpenAI. GPT-4 technical report. **arXiv:2303.08774**, 2023.

[4] X. Lin, et al. Few-shot learning with multilingual generative language models. In **Proc. EMNLP**, pp. 9019–9052, 2022.

[5] A. Chowdhery, et al. PaLM: Scaling language modeling with pathways. **arXiv:2204.02311**, 2022.

[6] H. Touvron, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv:2307.09288**, 2023.

[7] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li. Multilingual machine translation with large language models: Empirical results and analysis. **arXiv:2304.04675**, 2023.

[8] T. Kocmi, et al. Findings of the 2022 conference on machine translation (WMT22). In **Proc. 7th WMT**, pp. 1–45, 2022.

[9] E. Briakou, C. Cherry, and G. Foster. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In **Proc. 61st ACL**, pp. 9432–9452, 2023.

[10] H. Xu, Y. Kim, A. Sharaf, and H. Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. **arXiv:2309.11674**, 2023.

[11] C. Zhou, et al. LIMA: Less is more for alignment. In **Proc. 37th NIPS**, 2023.

[12] M. Morishita, K. Chousa, J. Suzuki, and M. Nagata. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In **Proc. 13th LREC**, pp. 6704–6710, 2022.

[13] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In **Proc. 10th ICLR**, 2022.

[14] L. Barrault, et al. Findings of the 2020 conference on machine translation (WMT20). In **Proc. 5th WMT**, pp. 1–55, 2020.

[15] F. Akhbardeh, et al. Findings of the 2021 conference on machine translation (WMT21). In **Proc. of the 6th WMT**, pp. 1–88, 2021.

[16] Z. Mao and T. Nakagawa. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. In **Proc. 17th EACL**, pp. 1886–1894, 2023.

[17] T. Scialom, T. Chakrabarty, and S. Muresan. Fine-tuned language models are continual learners. In **Proc. EMNLP**, pp. 6107–6122, 2022.

[18] NLLB Team, et al. No language left behind: Scaling human-centered machine translation. **arXiv:2207.04672**, 2022.

[19] G. Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

[20] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In **Proc. 7th ICLR**, 2019.

[21] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proc. 40th ACL**, pp. 311–318, 2002.

[22] R. Rei, J. C. de Souza, D. Alves, C. Zerva, A. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In **Proc. 7th WMT**, pp. 578–585, 2022.

[23] M. Post. A call for clarity in reporting BLEU scores. In **Proc. 3rd WMT**, pp. 186–191, 2018.

[24] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara. ASPEC: Asian scientific paper excerpt corpus. In **Proc. 10th LREC**, pp. 2204–2208, 2016.

[25] R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz. JESC: Japanese-English subtitle corpus. In **Proc. 11th LREC**, 2018.

[26] M. Cettolo, C. Girardi, and M. Federico. WIT3: Web inventory of transcribed and translated talks. In **Proc. 16th EAMT**, pp. 261–268, 2012.

[27] M. Rikters, R. Ri, T. Li, and T. Nakazawa. Designing the business conversation corpus. In **Proc. 6th WAT**, pp. 54–61, 2019.

[28] Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. Findings of the first shared task on machine translation robustness. In **Proc. 4th WMT**, pp. 91–102, 2019.

[29] L. Specia, Z. Li, J. Pino, V. Chaudhary, F. Guzmán, G. Neubig, N. Durrani, Y. Belinkov, P. Koehn, H. Sajjad, P. Michel, and X. Li. Findings of the WMT 2020 shared task on machine translation robustness. In **Proc. 5th WMT**, pp. 76–91, 2020.

[30] A. Anastasopoulos, et al. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In **Proc. 18th IWSLT**, pp. 1–29, 2021.

A　予備実験の詳細

JParaCrawl v3.0 で LoRA チューニングによる SFT を行った rinna-4b を，WMT22 のテストデータで評価した．エポック数は 1 とし，使用した JParaCrawl v3.0 は LEALLA-large で取得した文埋め込みベクトルのコサイン類似度が 0.76 以上 0.95 未満の 1,000 万文対をサンプリングし，訓練を行った．開発データは WMT20 および 21 の開発，テストデータとし，開発データの誤差が最小となるモデルで greedy decoding による推論を行った．また，rinna-4b は，英日・日英の両方向をそれぞれ別々に訓練を行い，学習可能パラメータは 1,680 万である．その結果を表 3 に示す．rinna-4b は単言語データで事前訓練されており，かつ 1,000 万文対で LoRA チューニングを行ったのにもかかわらず，BLEU と COMET ともに baseline に劣る結果となった．

表 3　WMT22 のテストデータの結果 (BLEU / COMET).

モデル	英日翻訳	日英翻訳
baseline	21.8 / 85.4	21.6 / 80.1
rinna-4b (LoRA)	17.2 / 83.3	15.5 / 76.5

B　JParaCrawl v3.0 のサンプリング

継続事前訓練に用いる JParaCrawl v3.0 は，LEALLA-large で取得した文埋め込みベクトルのコサイン類似度が 0.4 以上 0.95 未満の文対をサンプリングした．類似度が 0.4 未満のサンプルは，和文と英文とで文章の長さが不相応というような，不適切なサンプルが目視で散見されたため除外した．また，類似度が 0.95 以上の文対は，和文と英文がほぼ同一の文章となっているサンプルがほとんどであったため除外した．このサンプリングにより，継続事前訓練の総トークン数は rinna-4b の tokenizer で約 18 億となった．

C　ハイパーパラメータの詳細

継続事前訓練で用いる AdamW は，$\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 10^{-8}$ とした．そして，weight decay を 0.1，gradient clipping を 1.0 とした．また，最大の学習率を 1.5×10^{-4} とし，warmup ratio を 1%の cosine scheduler とした．

SFT では，継続事前訓練で用いた AdamW のパラメータのうち，$\beta_2 = 0.999$ に変更した．そして，weight decay と gradient clipping は継続事前訓練と同じ値とした．最大の学習率を full fine-tuning では 3.0×10^{-5}，LoRA チューニングでは 2.0×10^{-4} とし，warmup ratio を 1%の inverse square scheduler とした．LoRA では，$r = 16, \alpha = 32$，dropout を 0.05 とし，multi-head attention に存在する Query，Key，Value の Linear 層に適用した．

D　テストセットの詳細

表 4　テストセット各データの分野および対訳文数

テストセット	分野	対訳文数
ASPEC [24]	科学技術論文	1,812
JESC [25]	映画字幕	2,000
KFTT [19]	Wikipedia 記事	1,160
TED (tst2015) [26]	TED Talk	1,194
Business Scene Dialogue Corpus [27]	対話	2,120
WMT19 Robustness En-Ja (MTNT2019) [28]	Reddit	1,392
WMT19 Robustness En-Ja (MTNT2019) [28]	Reddit	1,111
WMT20 Robustness Set1 En-Ja [29]	Wikipedia コメント	1,100
WMT20 Robustness Set2 En-Ja [29]	Reddit	1,376
WMT20 Robustness Set2 Ja-En [29]	Reddit	997
IWSLT21 Simultaneous Translation En-Ja Dev [30]	TED Talk	1,442
WMT22 News En-Ja [8]	ニュース	2,037
WMT22 News Ja-En [8]	ニュース	2,008