計算資源が限られた複数組織での出力選択による協働の検討

伊藤郁海 [1]　伊藤拓海 [1,2]　鈴木潤 [1,3]　乾健太郎 [1,3]
[1] 東北大学 [2]Langsmith 株式会社 [3] 理化学研究所
ikumi.ito.p8@dc.tohoku.ac.jp　{t-ito,jun.suzuki,kentaro.inui}@tohoku.ac.jp

概要

潤沢な計算資源を持たない組織一つでは，高性能な大規模ニューラルモデルの開発は困難である．本稿では，そのような組織がそれぞれ開発したモデルを集結させて組み合わせた場合の性能は，潤沢な計算資源を使い一箇所で訓練されたモデルの性能に匹敵するのかどうか検証する．実験では，独立に訓練された少訓練データかつ小サイズの小モデル群と，多訓練データかつ大サイズの単独大モデルの英日翻訳の性能比較を行った．実験の結果，小モデル群が生成する複数の出力の中から適切な出力を選択した場合，小モデル単体と比較して 10 倍の訓練データ量かつ 3 倍のモデルサイズの大モデルの能力に匹敵する性能を達成した．

1　はじめに

近年，自然言語処理分野において大規模ニューラルモデルが様々なタスクで高い性能を達成している．その背景には，より多くの訓練データを用いて，より大きなモデルを訓練することで，モデルの性能がべき乗則に従って向上し続けることが期待できるという実験的知見がある [1]．今後も高性能モデルの実現を目的とした訓練データやモデルサイズのさらなる大規模化が進む可能性は十分考えられる．こうした大規模モデルの開発には多くの計算資源が必須となるため，潤沢な計算資源を持つ一部の企業や研究機関（大組織）のみがモデル開発を行っているというのが実情である．しかし，大学の研究室など計算資源が潤沢ではない組織（小組織）が高性能なモデルの開発に関与できないという状況は望ましくない [2]．限られた計算資源を最大限に活用する方法を模索し，より広く多くの研究者や開発者がモデルの性能向上において競争力を持つことは重要である．

では，どのようにすれば計算資源で劣る組織が性能の高いモデルを開発することができるだろうか．本稿では，計算資源で劣る複数の小組織が力を合わせた場合に，大組織が作るモデルに匹敵する性能を達成できるかどうかを検証する．具体的には，独立に訓練された少訓練データかつ小サイズの小モデル群と多訓練データかつ大サイズの単独大モデルの性能比較を行う．本稿では，小モデル群の活用方法として，各小モデルの出力の中から最適であると推定される出力を小モデル群全体の出力として採用するという戦略をとる．訓練済みモデルを多数集め，それらのモデルの出力の中から最適であると推定される出力を選択するという戦略は，各モデルの訓練は独立に行うことができるという点で有用である．モデルの訓練を独立に実行できることには二つの利点がある．一つ目は，訓練データの共有が不要な点である．これによりプライバシー保護が必要な組織（例：医療機関，金融機関など）が持つデータを活用できる．二つ目は，他の組織の開発やモデル共有後の工程に対する考慮が不要な点である．個々の組織は，既存のモデル開発ノウハウをそれぞれの状況に合わせて自由に適用することができる．

本稿では，英日翻訳タスクを題材として小組織群と大組織間の比較を行った．模擬実験としての簡易化のために，以下の制約のもとに実験を行った．

- 各小組織は，大組織用の訓練データを等分割したものの一部を訓練データとして用いる．
- 訓練，推論の独立性は保った上で，小組織間で同一の実験設定を用いる．

実験の結果，複数の小モデルの出力集合に対して，参照なし評価指標を用いた出力選択を行ったところ，大モデル（小モデル単体と比較して 10 倍の訓練データ量かつ 3 倍のモデルサイズ）の性能に匹敵する結果となった．

2　関連研究

計算資源が限られている場合でも，高精度なモデルを構築するため，さまざまな研究が行われてい

る．例えば，モデル性能を保ちながら効果的な訓練データを選択することで，訓練データ量の削減を試みた研究がある [2, 3]．また，量子化などによりパラメータサイズを削減した上でモデルを訓練するという方法も提案されている [4, 5]．

他にも，本研究と類似した取り組みとして，独立した複数の計算機の協働による分散的なモデル訓練方法も研究されている．例えば，連合学習は，モデルの訓練データが分散して存在する場合に，それぞれの場所で独立にモデルの訓練を行った後，各モデルの更新情報を集約することにより，全体として一つの大きなモデルを訓練する手法である [6]．また，訓練の独立性を保ちながら複数モデルの合成を繰り返し，性能向上を試みる新たな事前訓練手法も提案されている [7]．

また，本研究と同様に，複数のモデルを推論時に用いることで良い出力を得る方法も研究されている．例えば，アンサンブル [8] は，複数のモデルの予測確率を共有し，その平均から最終的な出力を得る手法である．そのため，複数の組織間でアンサンブルを適用するためには事前にモデルの辞書を共有することが必要である．

モデルの推論時ではく，推論後の出力集合をリランキングやその集合から最も良い出力を選択するという方法も研究されている [9]．

3 タスク定義

本稿では，モデルの訓練に使用するデータ量とモデルサイズを制御することにより，使える計算資源が少ない小組織と，豊富な計算資源を使える大組織を再現する．本実験の概要を図 1 に示す．

小組織の総数を n とし，大組織，小組織が使用する訓練データをそれぞれ D，$d_i(1 \leq i \leq n)$ とすると，本実験における D と d_i の関係は以下のように表される[1)]．

$$D = d_1 \cup d_2 \cup \ldots \cup d_n, \quad d_i \cap d_j = \emptyset \, (i \neq j) \tag{1}$$

また，大組織，小組織が使用するモデルをそれぞれ M，m_i とすると，モデルサイズに関して，$M > m$ が成り立つ．実験にて使用する具体的なデータセットおよびモデルは，4.2 節，4.1 節にて述べる．

大組織は，訓練データ D を用いてモデル M の訓練を行い，評価データ X に対する出力を評価指標

1) 組織間において共通の訓練データが存在し，かつ訓練データの一部のみが各組織独自のデータであるような状況も想定され得るが，本実験の対象外とする．

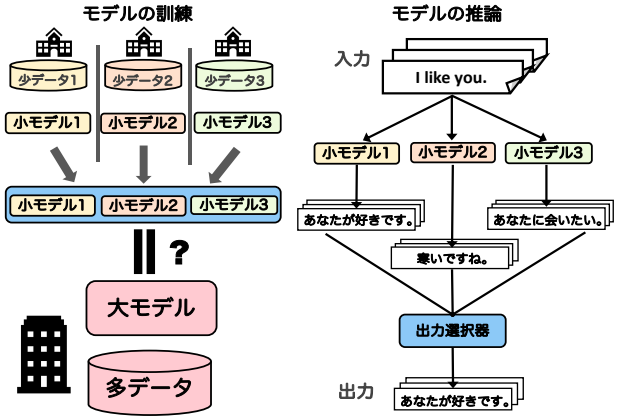

図 1 実験設定の概要

Eval(·) により評価する．大組織のモデルのスコアは次の式で表される．

$$\text{score}_{\text{Large}} = \frac{1}{|X|} \sum_{n=1}^{|X|} \underset{x_n \in X}{\text{Eval}}(M(x_n)) \tag{2}$$

各小組織は，訓練データ d_i を用いてモデル m_i の訓練を独立に行う．その後，評価データ $x_j \in X(1 \leq j \leq |X|)$ に対する小モデルたちの候補文集合 Y_j を得る．

$$Y_j = \{y_1 = m_1(x_j), y_2 = m_2(x_j), \ldots, y_i = m_i(x_j)\} \tag{3}$$

出力選択関数を $f_{\text{sel}}(\cdot)$ とすると，評価文 x_j に対して最終的な出力 h_j は以下のようになる．

$$h_j = \underset{y \in Y_j}{\arg\max} \, f_{\text{sel}}(y) \tag{4}$$

評価データ X に対する小モデル群のスコアは次の式で表される．

$$\text{score}_{\text{small}} = \frac{1}{|X|} \sum_{n=1}^{|X|} \text{Eval}(h_n) \tag{5}$$

実験にて使用する具体的な Eval および f_{sel} は，4.3 節にて述べる．

本稿では，$\text{score}_{\text{Small}}$ が $\text{score}_{\text{Large}}$ に匹敵する値となるかどうかを検証する．

4 実験

本稿では，英日翻訳タスクにて実験を行う．

4.1 モデル

翻訳モデルとして，Transformer [10] を使用した．モデルパラメータが約 100M のモデルを小組織が扱う小モデル，約 300M のモデルを大組織が扱う大モデルとみなした．

4.2 訓練用/評価用データセット

訓練データとして JParaCrawlv3.0 [11]（約 25M）を使用した．小モデルは，25M の訓練データを等分割したもの[2]の一部を用いてモデルの訓練を行う．大モデルは，25M の訓練データを全て用いてモデルの訓練を行う．また，評価用データとして WMT2022 の newstest2022 [12] を使用した．サブワード分割器およびモデルの辞書の作成は，分割前の全訓練データをもとに行い，小組織，大組織含め全モデルで同一のものを使用した．また、評価指標として，COMET[13][3]と BLEURT [14, 15][4]，sacreBLEU [16] を使用する．

4.3 出力選択方法

本実験では，式 4 の出力選択関数 $f_{\text{sel}}(\cdot)$ として，参照あり評価指標である COMET と参照なし評価指標である COMETKIWI [17][5]を使用する．参照あり手法は評価時に参照テキストを使用し出力選択をしているため，実際の設定では使用することができない．また，COMET は評価指標として使用したものと同じモデルを使用した．つまり，COMET による出力選択はオラクル設定である．一方で，COMETKIWI は出力選択時に，翻訳モデルへの原文と翻訳モデルの出力のみが必要であり，現実の設定でも利用可能な方法である．[6]

4.4 比較設定

本稿では，以下の設定の翻訳性能を比較する．

- `Small`：小モデル．それぞれ小モデル群の平均値（`Mean`），最大値（`Max`），最小値（`Min`）を報告する．
- `Large`：大モデル単体．
- `Est`：小モデルの候補文集合から COMETKIWI で最大スコアとなる翻訳文を選択する方法．
- `PostCosE`：小モデルの候補文集合に含まれる各文を単語分散表現を用いてベクトル化し，他の文とのコサイン類似度が最も高い翻訳文を選択する方法 [9]．単語分散表現は，fastText[19] により，JParaCrawlv3.0 の日本語データを用いて作成．
- `MaxRef`：小モデルの候補文集合から COMET で最大スコアとなる翻訳文を選択する方法．

個々の小モデルのハイパーパラメータ等の設定は全て揃えて実験を行った．なお，ハイパーパラメータ等の詳細な設定は付録 A に示す．

表 1　訓練データを 10 等分割した場合の結果（newstest2022）．

		COMET	BLEURT	sacreBLEU
`Small`	`(Mean)`	0.7605	0.5926	18.4
	`(Max)`	0.7717	0.6055	19.8
	`(Min)`	0.7455	0.5801	17.7
`Large`		0.8026	0.6439	**22.3**
`Est`		**0.8130**	**0.6474**	20.7
`PostCosE`		0.7796	0.6140	20.1
`MaxRef`		0.8333	0.6683	23.2

4.5 実験結果・考察

表 1 に，訓練データを 10 等分割し，小モデルが 10 個存在する状態における WMT2022 の newstest2022 に対する各手法の評価結果を示す．[7]

`Small` は，いずれも `Large` に劣る性能を示した．[8]`Est` は，COMET と BLEURT において `Large` を超える性能を示し，`MaxRef` は全ての評価指標において `Large` を超える性能を示した．これらの結果より，モデルを複数用意した上で，それぞれの出力の中から最適な選択ができた場合，個々のモデルの性能から大きな性能向上が期待できる．また，`Est` と `MaxRef` の性能差は COMET と COMETKIWI の性能差でもある．

他方，`PostCosE` は `Est` と比較して小さな性能向上に留まった．その要因として，それぞれの手法の出力選択戦略の違いが考えられる．`PostCosE` は，モデル群の中央値的な出力を採用する手法である．この手法は，モデル群の中で及第点となる性能を持つモデルが多数派となるような状況において有効である

2） 本実験では，分割においてドメインやトークン数など各データ集合の性質に関する考慮は行わない．
3） https://unbabel-experimental-models.s3.amazonaws.com/comet/wmt22/wmt22-comet-da.tar.gz（11 月 22 日時点で公開されていたモデル）
4） https://storage.googleapis.com/bleurt-oss-21/BLEURT-20.zip
5） https://unbabel-experimental-models.s3.amazonaws.com/comet/wmt22/wmt22-cometkiwi-da.tar.gz（11 月 23 日時点で公開されていたモデル）
6） wmt22-comet-da の訓練データは，2017 年から 2020 年の WMT News Translation Task における人手評価データである．また，wmt22-cometkiwi-da の訓練データは，wmt22-comet-da の訓練データと MLQE-PE データ [18] である．
7） WMT2022 の newsdev2022，WMT2020 の newsdev2020 に対する評価も行い，結果を付録 C に示した．
8） 付録 B に，全訓練データで訓練した `Small` モデルの性能と 10 分割データで訓練した `Large` モデルの性能を示す．`Large` が `Small` を超える性能を示した要因は，訓練データの量とモデルサイズの両方の影響であると示唆される．.

表 2　訓練データを 4 等分割した場合の結果（newstest2022）．

	COMET	BLEURT	sacreBLEU
Small-0	0.7846	0.6200	20.0
Small-1	0.7916	0.6289	20.9
Small-2	0.7722	0.6100	20.7
Small-3	0.7822	0.6163	20.2
Small-0+1	0.7919	0.6296	21.1
Small-0+2	0.7853	0.6253	21.2
Small-0+3	0.7915	0.6281	21.2
Small-1+2	0.7869	0.6266	21.8
Small-1+3	0.7944	0.6313	21.4
Small-2+3	0.7847	0.6208	21.2
Small-0+1+2	0.7898	0.6292	21.5
Small-0+1+3	0.7942	0.6325	21.5
Small-0+2+3	0.7896	0.6278	21.5
Small-1+2+3	0.7910	0.6308	21.7
Small-0+1+2+3	0.7933	0.6315	21.6
Large	0.8026	0.6439	**22.3**
Est(4)	0.8138	0.6523	21.8
Est(15)	**0.8197**	**0.6586**	**22.3**
PostCosE(4)	0.7943	0.6315	21.4
PostCosE(15)	0.7934	0.6315	21.7
MaxRef(4)	0.8283	0.6649	23.7
MaxRef(15)	0.8355	0.6734	24.7

と考えられる．一方で，Est は，モデル群の中で最良と推定される出力を採用する手法である．この手法は，モデル群の中で性能の低いモデルが多数派であっても，高性能なモデルが一つでも存在すればそのモデルの出力が選択される．訓練データが全く異なるなど，本実験のように各モデルの性能にばらつきが生じる得る状況の場合は，Est の方が PostCosE よりも好相性であることが示唆される．

4.6　発展：アンサンブル

使用できるモデルが複数ある場合に，それらのモデルを活用する方法の一つとして推論時のアンサンブル [8] が挙げられる．しかし，一般的なアンサンブルではモデル間でモデルの辞書を共有する必要がある．本稿では各小モデルが独立に訓練される場面を想定しているため，アンサンブルは比較設定に含まなかった．しかしながら，実験前に全ての小組織間で調整を行い，アンサンブルの実施が可能な場合は有効な手法である．そこで，本章では追加でアンサンブルを含む実験を行い，その結果を報告する．なお，章 4.5 とは異なり，訓練データを 4 等分割し，小モデルが 4 個存在する状態にて実験する．

表 2 に，WMT2022 の newstest2022 に対する各手法の評価結果を示す．Small-*のハイフン以下はモデルの ID を示す．また，0+1 は Small-0 と Small-1 をアンサンブルした場合の結果である．***(4)，***(15) は出力選択の対象となるモデル数の違いを示す．つまり，Est(4) は Small-0 から Small-3 の 4 モデル、Est(15) は Small-0 から Small-0+1+2+3 までの 15 モデルで候補文集合を得て出力選択を行った設定である．

全ての評価指標において，Est(4) よりも Est(15)，MaxRef(4) よりも MaxRef(15) の方が高い性能を示した．これは，モデルアンサンブルを適用することで，候補文集合内により良い出力を得られたことを意味する．

また，アンサンブルに参加するモデル内に，性能が低いモデルが含まれる場合は，性能が向上しない場合もあることがわかった．4 つの小モデル単体の性能に着目すると，Small-2 の COMET が他のモデルと比較して低く，Small-2 がアンサンブルに関わる場合の COMET の値に注目すると，Small-2 が参加しない場合の方が高い性能を示している（例：Small-0+1：0.7919, Small-0+1+2：0.7898 ）．

5　おわりに

本稿では，独立に訓練された少訓練データかつ小サイズの小モデル群は，多訓練データかつ大サイズの単独大モデルと同等の性能を達成できるのか，という研究課題の検証を英日翻訳の性能比較により行った．複数の小モデルの出力集合に対して，参照なし評価指標を用いた出力選択を行った場合の性能は，大モデルの性能に匹敵することを示した．本実験結果は，計算資源の制約により単独組織が構築できるモデルの性能に限界があったとしても，複数の組織が協働することで，潤沢な計算資源を持つ単独組織が作成する大規模モデルに匹敵する性能を達成できる可能性があることを示唆するものである．

候補文集合から出力選択を行う戦略においては，出力選択器の性能が鍵となるため，出力選択器のさらなる性能向上が望まれる．今後の展望として，翻訳タスク以外でも実験を行うことを計画している．また，GPT-3 [20] のような汎用言語モデルによる複数タスクの性能比較においても本実験と同様の結果が得られるのかも検証する予定である．本実験では各小組織の訓練データが全て異なる設定を用いたが，訓練データの一部を共有するなど，より現実に即した設定での実験を検討する．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2011（fundamental research），JSPS 科研費 JP21J14152 の助成を受けて実施されたものです．また，本研究を進めるにあたり，有益な助言を頂きました東北大学　乾・坂口・徳久研究室の仲村祐希氏へ感謝いたします．

参考文献

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, 2020.

[2] 鈴木潤, 全炳河, 賀沢秀人. ニューラル言語モデルの効率的な学習に向けた代表データ集合の獲得. 言語処理学会第 28 回年次大会, 2022.

[3] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset Distillation, 2018.

[4] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations. **J. Mach. Learn. Res.**, Vol. 18, No. 1, p. 6869–6898, jan 2017.

[5] Ron Banner, Itay Hubara, Elad Hoffer, and Daniel Soudry. Scalable Methods for 8-Bit Training of Neural Networks. In **Proceedings of the 32nd International Conference on Neural Information Processing Systems**, NIPS'18, p. 5151–5159, Red Hook, NY, USA, 2018. Curran Associates Inc.

[6] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, **Proceedings of the 20th International Conference on Artificial Intelligence and Statistics**, Vol. 54 of **Proceedings of Machine Learning Research**, pp. 1273–1282. PMLR, 2017.

[7] Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. Branch-Train-Merge: Embarrassingly Parallel Training of Expert Language Models, 2022.

[8] Lior Rokach. Ensemble-based classifiers. **Artificial intelligence review**, Vol. 33, No. 1, pp. 1–39, 2010.

[9] Hayato Kobayashi. Frustratingly Easy Model Ensemble for Abstractive Summarization. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 4165–4176, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[11] Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 6704–6710, Marseille, France, June 2022. European Language Resources Association.

[12] Tom Kocmi, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, et al. Findings of the 2022 Conference on Machine Translation (WMT22).

[13] Ricardo Rei, José GC de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In **Proceedings of the Seventh Conference on Machine Translation, Abu Dhabi. Association for Computational Linguistics**, 2022.

[14] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning Robust Metrics for Text Generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[15] Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. Learning Compact Metrics for MT. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 751–762, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[16] Matt Post. A Call for Clarity in Reporting BLEU Scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[17] Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC de Souza, Taisiya Glushkova, Duarte M Alves, Alon Lavie, et al. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. **arXiv preprint arXiv:2209.06243**, 2022.

[18] Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. MLQE-PE: A Multilingual Quality Estimation and Post-Editing Dataset. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 4963–4974, Marseille, France, June 2022. European Language Resources Association.

[19] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. **Transactions of the Association for Computational Linguistics**, Vol. 5, pp. 135–146, 2017.

[20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[21] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

A　実験設定の詳細

翻訳モデルの構築には fairseq [21] を使用した．どのモデルでも共通して，最適化手法には Adam，Label-Smoothing は $\epsilon_{ls} = 0.1$，Warmup Steps は更新回数の三分の一とした．また，推論時のビーム探索幅は 5 とした．付録 B で使用する 2B モデルは，大モデルのアーキテクチャに対して、隠れ層と順伝播層の次元増やしたモデルである．表 3，表 4，表 5 に Small，Large，2B でのハイパーパラメータで異なる設定を示す．

表 3　Small モデルのハイパーパラメーターの一覧．

設定名	値
翻訳モデル	transformer
学習率	5e-4
ミニバッチサイズ	512,000
エポック数	200
モデル平均化	20 エポック毎にモデルを保存 最終 5 モデルの平均

表 4　Large モデルのハイパーパラメーターの一覧．

設定名	値
翻訳モデル	transformer-wmt-en-de-big
学習率	5e-4
ミニバッチサイズ	512,000
エポック数	50
モデル平均化	5 エポック毎にモデルを保存 最終 5 モデルで平均化

表 5　2 B モデルのハイパーパラメーターの一覧

設定名	値
翻訳モデル	transformer-wmt-en-de-big（embedding を含む隠れ層の次元：2048，順伝播層の次元：32,768）
学習率	1e-04
ミニバッチサイズ	512,000
更新回数	30,000
モデル平均化	500 回パラメータ更新単位で保存し 最終 10 モデルの平均

B　データ量とモデルサイズの影響

表 6 に訓練データ量とモデルサイズを変化させた場合のモデル性能の変化を示す．***-full は，25M の訓練データで訓練したモデルの性能を示す．***-div10 は，25M の訓練データを 10 分割したものの一部を用いて訓練したモデルの性能を示す．また，比較対象として新たに 2B のモデルを加えた．[9)]

-div10 と-full の比較より，訓練データ量の違い（***-div10：2.5M，***-full：25M）がモデルの性能に影響していることがわかる．Small-div10 と Large-div10 の比較より，モデルサイズの違い（Small：100M，Large：300M）もモデルの性能に影響していることが示唆される．Est（Small）と Est（Large）の比較より，出力選択は個々のモデルの性能が向上した場合でも有効であることがわかる．

表 6　訓練データ量とモデルサイズを変化させた場合の各評価指標の値（newstest2022）．

		COMET	BLEURT	sacreBLEU
Small-div10	(Mean)	0.7605	0.5926	18.4
	(Max)	0.7717	0.6055	19.8
	(Min)	0.7455	0.5801	17.7
Est（Small）		0.8048	0.6392	20.2
MaxRef（Small）		0.8171	0.6530	22.1
Large-div10	(Mean)	0.7823	0.6168	20.1
	(Max)	0.7916	0.6266	21.3
	(Min)	0.7627	0.5981	19.5
Est（Large）		0.8258	0.6640	22.3
MaxRef（Large）		0.8455	0.6842	25.3
Small-full		0.7901	0.6273	20.9
Large-full		**0.8026**	**0.6439**	22.3
2B		0.7981	0.6371	**22.5**

C　10 分割設定における性能評価

表 1 の実験と同様の実験を，newsdev2022 と newsdev2020 を評価データとして行った．表 7 に newsdev2022 に対する結果を，表 8 に newsdev2020 に対する結果を示す．いずれの評価データにおいても，全ての評価指標で Est が Large を上回ることはなかった．しかし，Est と Small の比較から出力選択が性能向上に寄与していることがわかる．

表 7　訓練データを 10 等分割した場合の各評価指標の値（newsdev2022）

		COMET	BLEURT	sacreBLEU
Small	(Mean)	0.7342	0.5146	17.9
	(Max)	0.7568	0.5385	19.7
	(Min)	0.7052	0.4843	16.5
Large		**0.8101**	**0.6042**	**23.9**
Est		0.8014	0.5873	20.7
PostCosE		0.7580	0.5394	20.0
MaxRef		0.8225	0.5993	22.0

表 8　訓練データを 10 等分割した場合の各評価指標の値（newsdev2020）

		COMET	BLEURT	sacreBLEU
Small	(Mean)	0.7517	0.5400	15.2
	(Max)	0.7676	0.5608	16.0
	(Min)	0.7336	0.5181	13.6
Large		**0.8143**	**0.6180**	**20.3**
Est		0.8106	0.6026	16.9
PostCosE		0.7727	0.5641	16.3
MaxRef		0.8305	0.6177	18.9

9）Large の性能を超えられなかったため，今回の訓練データに対しては訓練がうまくいかなかった．