最適輸送に基づく擬似訓練データを用いた機械翻訳の品質推定

黒田 勇斗 [1]　藤田 篤 [2]　梶原 智之 [1]
[1] 愛媛大学大学院理工学研究科　[2] 情報通信研究機構
{kuroda@ai., kajiwara@}cs.ehime-u.ac.jp　atsushi.fujita@nict.go.jp

概要

人間が作成した参照訳を用いずに機械翻訳文のどの部分に修正が必要かを推定するタスクを単語レベルの品質推定という．単語レベルの品質推定の性能を向上させるため，擬似訓練データの活用に関する研究が行われている．擬似訓練データは，機械翻訳文中の各語のラベルを参照訳との表層的な対応に基づいて定めることが多いが，この方法では表層が異なる単語が実際に翻訳品質を損なうか否かを適切に判断できない．そこで本稿では，最適輸送に基づく擬似訓練データを用いた品質推定モデルを提案する．WMT21 のデータセットを用いた実験の結果，提案モデルは，従来の擬似訓練データを用いた品質推定モデルを上回る性能を示した．

1　はじめに

機械翻訳の品質推定（Translation Quality Estimation; TQE）[1] とは，参照訳を用いず起点言語文とそれに対する機械翻訳文のみを参照して，機械翻訳文の品質を推定する技術である．機械翻訳文の品質を測る単位としては，文レベルおよび単語レベルの 2 種類が主に研究されている．いずれの評価単位についても，人手評価との相関が高い TQE 手法を開発することにより，機械翻訳文をそのまま使用するか，後編集を行うか，他の機械翻訳を利用するか等の判断を支援できる．単語レベルの TQE では，機械翻訳文のどの部分に修正が必要であるかという，細かい粒度の情報による支援が可能である．本稿では，単語レベルの品質推定に取り組む．

単語レベルの TQE の性能を向上させるため，擬似訓練データを活用した手法 [2, 3, 4] が研究されている．擬似訓練データの作成には，主に Translation Edit Rate (TER) toolkit [5] が用いられる，TER toolkit では，表層の一致に基づき，編集量が最小となるように，機械翻訳文と参照訳の間の単語レベルのアライメントを同定する．そのためこの手法は，表層は異なるものの翻訳品質を損なわないような単語も誤りとみなしてしまう．逆に，表層的には一致するものの単語の位置が不適切であり翻訳品質を損なうような例は，誤りであると適切に判断できない．

本稿では，単語埋め込みに基づく擬似訓練データの作成に焦点を当て，最適輸送（Optimal Transport; OT）を用いる手法について提案する．WMT21 の TQE タスクにおける実験の結果，最適輸送に基づく擬似訓練データにより TQE 向け事前訓練を行ったモデルは，TER に基づく擬似訓練データを用いたモデルを上回る性能を示した．

2　提案手法

2.1　擬似訓練データの作成

先行研究 [2, 3, 4] と同様に，擬似訓練データの作成には，対訳コーパス $D_{\text{para}} = (S_k, T_k)_{k=1}^{N}$ と機械翻訳器を用いる．まず，対訳コーパスの起点言語文を機械翻訳することで，擬似対訳文 $T'_k = \text{MT}(S_k)$ を得る．次に，T'_k の各単語に対する擬似ラベルを，OTAlign [6] から着想を得て，(T_k, T'_k) の間の最適輸送に基づいて定める．

最適輸送では，ある確率分布を別の確率分布と一致させるための最も効率的な輸送計画を探索する．まず，目標言語文 $T = [t_1, \ldots, t_n]$ と擬似対訳文 $T' = [t'_1, \ldots, t'_m]$ とし，最適輸送問題における輸送コストおよび確率分布の質量を次のように定義する．

輸送コスト　任意の語の間の輸送コスト $c(t_i, t'_j)$ を各語の単語埋め込みの非類似度により定義し，輸送コスト行列 $C_{i,j} = c(t_i, t'_j)$ を求める．

確率分布の質量　目標言語文および擬似対訳文のそれぞれの質量を $a \in \mathbb{R}_+^n$，$b \in \mathbb{R}_+^m$ とする．

そして，総輸送コストを最小化する最適輸送行列 P を次式で求める．

$$P = \underset{P' \in U(a,b)}{\arg\min} \sum_{i,j} C_{i,j} P'_{i,j} \tag{1}$$

U は最適化問題の制約をすべて満たす実行可能解の集合である．得られた最適輸送行列 P から，擬似対訳文の各単語に対する擬似ラベルとして，$y^{\text{soft}} \in [0,1]$ の連続値（ソフトラベル）および $y^{\text{hard}} \in \{\text{OK}, \text{BAD}\}$ の 2 値（ハードラベル）の 2 種類を定める．ソフトラベルは次のように定義する．

$$y_j^{\text{soft}} = \max([P_{0,j}, \ldots, P_{n,j}]) \tag{2}$$

$$Y^{\text{soft}} = [y_1^{\text{soft}}, \ldots, y_m^{\text{soft}}] \tag{3}$$

TER との比較のために，最適輸送行列から TER と同様のハードラベルを次のように定義する．

$$y_j^{\text{hard}} = \begin{cases} \text{OK} & y_j^{\text{soft}} > \lambda \\ \text{BAD} & \text{otherwise} \end{cases} \tag{4}$$

$$Y^{\text{hard}} = [y_1^{\text{hard}}, \ldots, y_m^{\text{hard}}] \tag{5}$$

得られた擬似ラベルを用いて，ソフトな擬似訓練データ $D_{\text{QE}}^{\text{soft}} = (S_k, T'_k, Y_k^{\text{soft}})_{k=1}^N$ およびハードな擬似訓練データ $D_{\text{QE}}^{\text{hard}} = (S_k, T'_k, Y_k^{\text{hard}})_{k=1}^N$ を作成する．

2.2 TQE モデルの事前訓練

擬似訓練データおよび事前訓練済みモデルを用いて，TQE 向け事前訓練を行う．ソフトラベルを予測するための回帰モデルと，ハードラベルを予測するための分類モデルの 2 つを考える．

先行研究 [7] にならい，起点言語文 $S = [s_1, \ldots, s_l]$ と機械翻訳文 $T = [t_1, \ldots, t_m]$ を連結して事前訓練済みモデルに入力する．出力として，機械翻訳文の各単語に対する特徴量 $H = [h_1, \ldots, h_m]$ を得る．そして，特徴量から機械翻訳文に含まれる i 番目の単語のラベルを次式により予測する．

$$\hat{y}_i = \sigma(w h_i) \tag{6}$$

回帰モデルの場合は，$w \in \mathbb{R}^{d \times 1}$ の線形層，σ はシグモイド関数となる．分類モデルの場合は，$w \in \mathbb{R}^{d \times 2}$ の線形層，σ はソフトマックス関数となる．d は，事前訓練済みモデルの隠れ層の次元数である．

3 評価実験

提案手法の性能を，WMT21 の TQE Task 2 [8] において評価した．WMT21 の TQE Task 2 では，原言語文，機械翻訳文，および機械翻訳文の修正訳文と TER によって定められた単語レベルの品質ラベルが与えられている．単語レベルの品質ラベルは，target タグおよび gap タグからなる．target タグは，機械翻訳文中の個々の単語に対する品質ラベルである．正しく翻訳されている単語には OK，置換および削除が必要な誤った単語には BAD が与えられている．gap タグは，機械翻訳文中の個々の単語間の空白に対する品質ラベルである．当該単語間に新たに単語を挿入する必要がある場合は BAD，必要ない場合は OK が与えられている．最適輸送を用いた擬似ラベルでは単語の挿入が必要な箇所を特定できないため，本稿では，target タグのみを評価対象とする．

公式の評価方法に従い，各 TQE 手法により推定した翻訳品質，および機械翻訳文と修正訳文から TER によって定められた正解ラベルの間のマシューズ相関係数（MCC）によって性能を評価した．

3.1 実験設定

訓練および評価用データセット　モデルの訓練および評価には，WMT21 の TQE Task 2 のデータセットを用いた．WMT21 の TQE Task 2 には，英語からドイツ語（En-De）および英語から中国語（En-Zh）の多資源言語対，ルーマニア語から英語（Ro-En）およびエストニア語から英語（Et-En）の中資源言語対，ネパール語から英語（Ne-En）およびシンハラ語から英語（Si-En）の少資源言語対の 6 つの言語対[1]が含まれる．各言語対において，訓練用，開発用および評価用データとして，それぞれ 7,000 文対，1,000 文対および 1,000 文対の起点言語文および機械翻訳文と，単語レベルの品質ラベルの組が提供されている[2]．評価対象の機械翻訳は，fairseq[3] [9] を用いて訓練された Transformer モデル [10] である．

WMT21 TQE Task 2 において評価用データのみが提供されている zero-shot 言語対についても TQE の性能を比較した．zero-shot 言語対として，英語からチェコ語（En-Cs），英語から日本語（En-Ja），クメール語から英語（Km-En）およびパシュトー語から英語（Ps-En）が含まれる．評価対象の機械翻訳は mBART50 [11] である．

擬似訓練データ　擬似訓練データの作成には，WMT21 の TQE タスクで利用可能な対訳コーパス[4]（表 1）および M2M-100[5] [12] の機械翻訳器を用いた．ただし，M2M-100 は，表 1 の対訳コーパスの一

1) https://github.com/sheffieldnlp/mlqe-pe
2) ロシア語から英語（Ru-En）は訓練・開発用データも公開されているが，擬似訓練データの作成に使用できる対訳コーパスが公開されていないため zero-shot 言語対として扱った．
3) https://github.com/pytorch/fairseq
4) https://www.statmt.org/wmt21/quality-estimation-task.html
5) https://huggingface.co/facebook/m2m100_418M

表 1 対訳文数および擬似訓練データ数

	言語対	対訳文	擬似訓練データ
多資源言語対	En-De	23,360,441	22,701,552
	En-Zh	20,305,268	16,201,271
中資源言語対	Ro-En	3,901,501	3,027,243
	Et-En	877,769	855,680
少資源言語対	Ne-En	498,271	166,893
	Si-En	646,766	570,770

部を用いて，WMT21 の TQE タスクの翻訳方向ごとにファインチューニングした．具体的には，多資源言語対：100 万文対ずつ，中資源言語対：20 万文対ずつ，少資源言語対：5 万文対ずつを無作為に抽出し，両言語ともに 128 トークン以下である対訳文のみを使用した．HuggingFace Transformers [13] を用いてファインチューニングを実施した．バッチサイズを 16 文対，学習率を $3e-5$，最適化手法を AdamW [14]（$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon = 1e-8$）として，多資源言語対および中資源言語対は 1 エポック，少資源言語対は 3 エポックの訓練を行った．

対訳コーパスから起点言語文および目標言語文の双方が 128 トークン以下である文対を抽出し，重複する文対を削除した上で，翻訳方向ごとにファインチューニングを経て得た機械翻訳器を用いて起点言語文を翻訳して機械翻訳文を得た．ただし，機械翻訳文が 128 トークンを超える場合は除外した．

擬似ラベルの作成には，次式に示すエントロピー正則化付き最適輸送 [15] を用いた．

$$P = \underset{P' \in U(a,b)}{\arg\min} \sum_{i,j} C_{i,j} P'_{i,j} - \xi H(P') \tag{7}$$

制約条件は次のように定めた [16, 17]．

$$U(a,b) = \{P \in \mathbb{R}_+^{n \times m} \mid P\mathbb{1}_m \leq a, P^\top \mathbb{1}_n \leq b, \\ \mathbb{1}_m^\top P^\top \mathbb{1}_n = \lambda_m\} \tag{8}$$

単語埋め込みを得るための事前訓練済みモデルを InfoXLM$_{\text{Base}}$[6] [18]，コスト関数をコサイン距離，質量を一様分布，正則化項の ξ を 0.1 とした．ただし，λ_m および λ の値は，開発用データに対する MCC が最大となるように定めた．λ_m については [0, 1] の範囲を 0.02 間隔で，λ については [0, 1] の範囲を 0.01 間隔で探索した．そして，MCC が最大となったパラメタ（付録 A）を用いて擬似対訳データに対する擬似ラベルを最適輸送により定めた．

TQE モデル　TQE モデルの訓練は，擬似訓練データを用いた TQE 向け事前訓練と WMT21 の TQE Task 2 の訓練用データを用いたファインチューニング（FT）の 2 段階からなる．

TQE 向け事前訓練では，事前訓練済みモデルに InfoXLM$_{\text{Large}}$[7] を用いた．各モデルは，HuggingFace Transformers を用い，バッチサイズを 2,048 文対，最適化手法を Adam（$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon = 1e-8$）として 1 エポック訓練し，訓練済みモデルのパラメタも更新した．損失関数は，ソフトラベルを用いて回帰モデルを訓練する際は平均二乗誤差，ハードラベルを用いて分類モデルを訓練する際は交差エントロピー誤差を用いた．ただし，ラベルの不均衡に対応するため，交差エントロピー誤差の BAD の重みを 3.0 とした．回帰モデルを評価する際は，閾値を 0.5 として MCC を計算した．

FT では，WMT21 の TQE Task 2 の訓練データを用いて，TQE 向け事前訓練済みモデルを追加で訓練した．バッチサイズを 64 文対，最適化手法を Adam（$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon = 1e-8$）として 10 エポック訓練した．損失関数は，TQE 向け事前訓練時と同じものを使用した．0.5 エポックごとに検証用データに対する評価を行い，MCC が最大となるモデルを選択した．回帰モデルの選択時には，上記に加え，[0, 1] の予測値に対する閾値も 0.1 間隔で探索した．

比較手法　ベースラインとして，TQE 向け事前訓練をせず，FT のみを行ったモデルと比較した．また，TER に基づく従来法および self-training により作成した擬似訓練データを用いて，TQE 向け事前訓練をしたモデルと比較した．self-training による擬似ラベルの作成には，WMT21 の訓練データのみを用いて FT のみを行った分類モデルを使用した．

3.2 実験結果

表 2 に実験結果を示す．表中の “OT” は最適輸送に基づく手法を指す．FT なし（表の上段），FT あり（表の下段）のいずれの設定においても，OT によって定めたソフトラベルを用いた場合に他のラベルを用いる場合よりも高い性能を示した．特に，FT なしの設定では，すべての言語対において最も優れた性能を示した．また，FT ありの設定では，一部の例外を除いて，TQE 向け事前訓練後に FT したモデルの性能が FT のみのモデルの性能を上回った．このことから，擬似訓練データの作成方法によらず，擬似訓練データを用いた TQE 向け事前訓練が有効であることがわかった．

6） https://huggingface.co/microsoft/infoxlm-base

7） https://huggingface.co/microsoft/infoxlm-large

表 2 WMT21 の TQE タスクにおける MCC．Arch. はモデルのアーキテクチャ，擬似訓練データは TQE 向け事前訓練時に使用した擬似訓練データの種類，FT は WMT21 の訓練用データを用いた訓練の有無を表す．

	擬似訓練データ			非 zero-shot 言語対						zero-shot 言語対				
Arch.	手法	ラベル	FT	En-De	En-Zh	Ro-En	Et-En	Ne-En	Si-En	En-Cs	En-Ja	Km-En	Ps-En	Ru-En
分類	TER	ハード	—	0.218	0.131	0.251	0.309	0.287	0.357	0.229	0.087	0.194	0.242	0.177
分類	OT	ハード	—	0.237	0.112	0.258	0.330	0.353	0.365	0.234	0.101	0.212	0.239	0.177
回帰	OT	ソフト	—	**0.280**	**0.153**	**0.289**	**0.366**	**0.376**	**0.413**	**0.268**	**0.131**	**0.313**	**0.271**	**0.180**
分類	—	—	✓	0.435	0.318	0.635	0.578	0.539	0.554	0.320	0.157	0.435	0.349	0.320
回帰	—	—	✓	0.440	0.323	**0.655**	0.570	0.543	0.554	0.335	0.168	0.460	0.370	0.319
分類	self	ハード	✓	0.458	0.312	0.626	0.592	0.542	0.564	0.362	0.143	0.456	0.379	0.337
分類	TER	ハード	✓	0.484	0.332	0.630	0.600	0.550	**0.585**	0.370	0.161	0.465	0.366	**0.347**
分類	OT	ハード	✓	0.482	0.320	0.632	0.585	0.548	0.575	0.367	0.161	0.461	0.371	0.330
回帰	OT	ソフト	✓	**0.490**	**0.335**	0.642	**0.603**	**0.565**	0.583	**0.379**	**0.187**	**0.478**	**0.383**	0.346

表 3 対訳コーパスフィルタリングをした場合の，WMT21 の TQE タスクにおける MCC．Arch. はモデルのアーキテクチャ，擬似訓練データは TQE 向け事前訓練時に使用した擬似訓練データの種類，FT は WMT21 の訓練用データを用いた訓練の有無，閾値は対訳コーパスフィルタリング時の閾値を表す．

	擬似訓練データ										
Arch.	手法	ラベル	データ数	閾値	FT	En-De	En-Zh	Ro-En	Et-En	Ne-En	Si-En
回帰	OT	ソフト	43.5M	—	—	0.280	0.153	0.289	0.366	0.376	0.413
回帰	OT	ソフト	41.2M	0.5	—	0.318	0.158	0.428	0.353	0.383	0.412
回帰	OT	ソフト	38.8M	0.7	—	0.337	0.162	0.466	0.377	**0.408**	**0.425**
回帰	OT	ソフト	13.4M	0.9	—	**0.349**	**0.171**	**0.499**	**0.409**	0.393	0.410
回帰	OT	ソフト	43.5M	—	✓	**0.490**	0.335	0.642	0.603	**0.565**	**0.583**
回帰	OT	ソフト	41.2M	0.5	✓	0.485	**0.347**	0.659	**0.604**	0.543	0.574
回帰	OT	ソフト	38.8M	0.7	✓	0.482	0.344	0.658	0.587	0.550	0.571
回帰	OT	ソフト	13.4M	0.9	✓	0.457	0.328	**0.662**	0.587	0.549	0.558

zero-shot 言語対においては，評価対象の言語対が訓練用データに含まれていないにも関わらず，FT ありのモデルの性能が FT なしのモデルの性能を上回った．OT によって定めたソフトラベルを用いた手法は，事前訓練を含まない手法および self-training に基づく手法よりも高い性能を達成した．最良のモデルを用いた self-training によってさらに性能を改善できる可能性もある．

4 擬似訓練データの品質の影響

擬似訓練データの作成に使用した対訳コーパスには，対訳関係にない文対が含まれる．そこで，LaBSE[8)] [19] を用いて起点言語文と目標言語文を各々符号化し，それらのコサイン類似度が閾値未満の対訳文を除外することで，対訳コーパスの品質ひいては擬似訓練データの品質の TQE の性能への影響を調査した．

表 3 に実験結果を示す．FT なし（表の上段）の設定では，対訳コーパスフィルタリングによって性能が向上した．一方，FT あり（表の下段）の設定では，6 つの翻訳方向のうち 3 つにおいてフィルタリングによって性能が低下した．この結果から，FT を行う前提の TQE 向け事前訓練では，対訳コーパスの品質よりも量が重要であることが示唆される．

5 おわりに

本稿では，最適輸送に基づく擬似訓練データを用いた TQE 手法について述べた．WMT21 の TQE タスクにおける実験を通じて，既存の擬似訓練データの作成手法と比較して，優れた性能を達成することを確認した．また，提案手法は，zero-shot 言語対においても有効であることがわかった．分析の結果，FT なしの設定では，対訳コーパスのフィルタリングにより，性能が向上することも明らかになった．

今後の課題として，まず，最適輸送のパラメタに関する検討が挙げられる．擬似ラベルを定める際に TER に基づく擬似ラベルに類似するようにパラメタを定めたことの是非，文ごとの最適性について調査したい．また，target タグと同様に，gap タグや起点言語文に対するタグについても，より精緻な擬似ラベルを定め方について検討する必要がある．

8） https://huggingface.co/sentence-transformers/LaBSE

謝辞

本研究の成果は，国立研究開発法人情報通信研究機構（NICT）の委託研究（課題番号：22501）により得られたものです.

参考文献

[1] Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. Quality Estimation for Machine Translation. **Synthesis Lectures on Human Language Technologies**, Vol. 11, No. 1, pp. 1–162, 2018.

[2] Lemao Liu, Atsushi Fujita, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Translation Quality Estimation Using Only Bilingual Corpora. **IEEE/ACM TASLP**, Vol. 25, No. 9, pp. 1762–1772, 2017.

[3] Dongjun Lee. Two-Phase Cross-Lingual Language Model Fine-Tuning for Machine Translation Quality Estimation. In **Proc. of WMT**, pp. 1024–1028, 2020.

[4] Yi-Lin Tuan, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Francisco Guzmán, and Lucia Specia. Quality Estimation without Human-labeled Data. In **Proc. of EACL**, pp. 619–625, 2021.

[5] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In **Proc. of AMTA**, pp. 223–231, 2006.

[6] Yuki Arase, Han Bao, and Sho Yokoi. Unbalanced Optimal Transport for Unbalanced Word Alignment. In **Proc. of ACL**, pp. 3966–3986, 2023.

[7] Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In **Proc. of WMT**, pp. 634–645, 2022.

[8] Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. Findings of the WMT 2021 Shared Task on Quality Estimation. In **Proc. of WMT**, pp. 684–725, 2021.

[9] 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 **Proc. of NAACL**, pp. 48–53, 2019.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In **Proc. of NIPS**, pp. 5998–6008, 2017.

[11] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual Translation from Denoising Pre-Training. In **Findings of ACL: ACL-IJCNLP 2021**, pp. 3450–3466, 2021.

[12] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond English-Centric Multilingual Machine Translation. **Journal of Machine Learning Research**, Vol. 22, No. 1, pp. 4839–4886, 2022.

[13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In **Proc. of EMNLP**, pp. 38–45, 2020.

[14] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In **Proc. of ICLR**, 2019.

[15] Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In **Proc. NIPS**, 2013.

[16] Luis A. Caffarelli and Robert J. McCann. Free Boundaries in Optimal Transport and Monge-Ampère Obstacle Problems. **Annals of Mathematics**, Vol. 171, No. 2, pp. 673–730, 2010.

[17] Alessio Figalli. The Optimal Partial Transport Problem. **Archive for Rational Mechanics and Analysis**, Vol. 195, p. 533–560, 2008.

[18] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In **Proc. of NAACL**, pp. 3576–3588, 2021.

[19] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT Sentence Embedding. In **Proc. of ACL**, pp. 878–891, 2022.

A　最適輸送のパラメタの探索

擬似訓練データを作成するため，WMT21 の開発用データを用いて，最適輸送のパラメタを探索した．λ_m は $[0, 1]$ の範囲を 0.02 の間隔，λ は $[0, 1]$ の範囲を 0.01 の間隔で，開発用データに対してラベルを作成し，作成したラベルと正解ラベルの MCC が最大となるパラメタを探索した．一方，単語埋め込みを得るための事前訓練済みモデルは INFOXLM$_{\text{Base}}$，コスト関数はコサイン距離，質量は一様分布，正則化項の ξ は 0.1 に固定した．単語埋め込みには，単語のサブワードに対する事前訓練済みモデルの最終層の埋め込みの平均プーリングの値を用いた．MCC が最大となったパラメタを表 4 に示す．擬似対訳データに対するラベルは，表 4 の値に従って最適輸送により定めた．

表 4　WMT21 の開発用データにおける MCC および最適輸送のパラメタ．

言語対	λ_m	λ	MCC
En-De	0.02	0.37	0.870
En-Zh	0.24	0.51	0.832
Ro-En	0.14	0.33	0.876
Et-En	0.02	0.35	0.803
Ne-En	0.14	0.37	0.679
Si-En	0.02	0.36	0.698