疑似訓練データに基づく機械翻訳の教師なし品質推定

黒田 勇斗 [1]　藤田 篤 [2]　梶原 智之 [1]　二宮 崇 [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.cs.ehime-u.ac.jp　atsushi.fujita@nict.go.jp
{kajiwara, ninomiya}@cs.ehime-u.ac.jp

概要

人間が作成した参照訳なしに機械翻訳文の品質を推定する手法を品質推定という．品質推定のデータの構築には，非常に高いコストがかかるため，教師あり学習に基づく品質推定はわずかな言語対に対してしか適用できない．この課題を解決するために，教師なし品質推定の研究が行われている．本稿では，対訳コーパスから自動生成できる疑似訓練データのみを用いた教師なし品質推定について述べる．実験の結果，疑似訓練データを用いて訓練したモデルは，多資源および少資源言語対を中心に，既存の教師なし品質推定手法よりも高い性能を達成した．

1　はじめに

機械翻訳の品質推定（Translation Quality Estimation；TQE） [1] とは，参照訳を用いず原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．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という判断を支援できる．このように，機械翻訳の実世界での利用を進める上で TQE は重要な技術である．

機械翻訳に関する国際会議 WMT における TQE シェアードタスク [2, 3] を中心に，多くの TQE モデル [4–8] が提案されてきた．これらのうち多くの手法は，「原言語文・機械翻訳文・人手評価値」の組を教師データとして用いる教師あり手法である．このような教師データの作成には，原言語と目的言語の両方に精通したアノテータによる主観評価（DA スコア）あるいは機械翻訳文の後編集（必要な編集量；HTER [9]）が必要であるため，非常にコストが高い．さらに，機械翻訳器によって機械翻訳文の品質の分布は大きく異なりうるが，複数の機械翻訳器の各々に対して教師データを作成することも現実的ではない．そのため，WMT21 の TQE タスク [3] においても，TQE モデルの訓練に使用できる教師データはわずか 7 言語対分しか蓄積されていない．

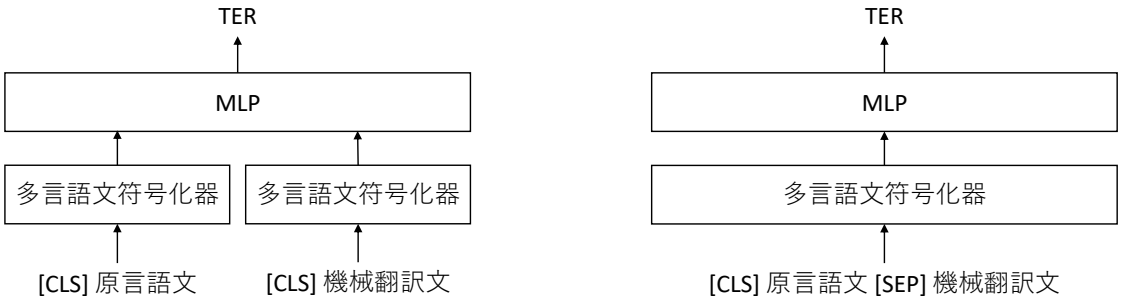
TQE における訓練データの作成コストの課題を解決するために，教師なし TQE 手法 [10–13] が研究されている．本研究では，対訳コーパスおよび訓練済みの機械翻訳器を用いて TQE 用の疑似的な訓練データを作成し，この疑似訓練データのみを用いて TQE モデルを訓練する手法について述べる．WMT20 の TQE タスクにおける実験の結果，疑似訓練データのみを用いて訓練したモデルは，既存の教師なし TQE 手法を上回る性能を示した．WMT21 の TQE タスクにおいても，少資源言語対を中心に既存の教師なし TQE 手法よりも高い性能を達成した．

2　先行研究

教師なし TQE の先行研究には，多言語文符号化器を用いる手法 [10, 11] および多言語系列変換器を用いる手法 [12, 13] がある．

多言語文符号化器を用いる手法は，原言語文およびその機械翻訳文をそれぞれベクトル化し，それらの余弦類似度を用いて機械翻訳文の品質を推定するものである．多言語文符号化器としては，例えば LaBSE [14] を用いることが考えられる．LaBSE は，複数言語の単言語コーパスを用いて事前訓練した文符号化器 mBERT [15] を，対訳コーパスを用いて新たな目的関数で再訓練したものである．訓練には人手評価値を使用していないため，LaBSE を用いた TQE は教師なし手法であるといえる．

しかし，多言語文符号化器による教師なし TQE では，多言語文符号化器の言語特異性 [10] が問題となる．つまり，多言語文符号化器から得られる文のベクトル表現は，意味よりも言語の影響を強く受けており，対象タスクに特化した再訓練なしでは言語が異なる文間の意味的類似度を正確に推定できない．この課題に対して DREAM [10] は対照学習，



(a) 文対をそれぞれ符号化（split）　**(b)** 文対を連結し符号化（concat）

図 1: モデルの概要

MEAT [11] は敵対的学習によって，多言語文符号化器から得られる文ベクトルを言語表現と意味表現に分離することを試みた．そして，原言語文およびその機械翻訳文のそれぞれに対する意味表現のみを TQE に用いて，LaBSE よりも高い性能を達成した．

もうひとつの教師なし TQE のアプローチとして，Prism [12] などの系列変換器に基づく手法がある．Prism では，自己注意ネットワークに基づく系列変換器 [16] を訓練し，そのモデルで原言語文から機械翻訳文への系列変換を forced decode した場合の文生成確率（forced decode スコア）を用いて機械翻訳文の品質を評価する．ただし，系列変換器に基づく手法は，訓練に大規模な対訳コーパスが必要であるため，少資源言語対に適用することは困難である．

3　疑似データを用いた訓練

本研究では，機械翻訳文の文レベルの品質として，HTER [9] を予測することを考える．HTER は，機械翻訳文から修正訳文（人間が機械翻訳文を後編集して作成した翻訳文）への編集距離である．

疑似訓練データの作成　疑似訓練データの作成には，対訳コーパス $C = (x_i, y_i)_{i=1}^{N}$ と機械翻訳器を用いる．対訳コーパスの原言語文 x を機械翻訳することで，疑似対訳文 $y' = \mathrm{MT}(x)$ を得る．「目的言語文 y・疑似対訳文 y'」の各対に対して TER [9] を求め，疑似訓練データ $D = (x_i, y'_i, \mathrm{TER}(y_i, y'_i))_{i=1}^{N}$ を作成する．ここで，TER は，疑似対訳文から目的言語文への編集距離である．

TQE モデル　作成した疑似訓練データおよび事前訓練済みの多言語文符号化器を用いて，回帰モデルの訓練を行う．本研究では，図 1 に示すように，2 種類の符号化方法を比較する．図 1(a) の方法では，2 節で述べた既存の手法と同様に，原言語文と機械翻訳文の文頭に [CLS] トークンを追加し，それぞれを独立に多言語文符号化器に入力する．図 1(b) の方法では，原言語文と機械翻訳文のトークン間の関連を捉えるために，文頭に [CLS] トークンを追加し，原言語文と機械翻訳文を [SEP] トークンで連結して多言語文符号化器に入力する．なお，両モデルとも，文頭の [CLS] トークンに対する多言語文符号化器の出力を多層パーセプトロン（MLP）に入力し，回帰モデルを訓練する．

表 1: 対訳文数および疑似訓練データ数

	言語対	対訳文	疑似訓練データ
多資源	en-de	23,360,441	22,268,661
言語対	en-zh	20,305,268	19,811,665
中資源	ro-en	3,901,501	3,178,631
言語対	et-en	877,769	863,973
少資源	ne-en	498,271	477,563
言語対	si-en	646,766	585,598

4　評価実験

3 節で述べた提案手法の性能を，WMT20 および WMT21 の TQE タスク [2, 3, 17] において評価した．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人間による機械翻訳文の修正訳文に基づいて与えられた HTER の間のピアソンの積率相関係数によって性能を評価した．

4.1　実験設定

評価データ　評価には WMT20 の TQE タスクの検証用データ[1)]および WMT21 の TQE タスクの評価用データを用いた．WMT20 の TQE タスクには，英語からドイツ語（en-de）および英語から中国語

1)　WMT20 の TQE タスクでは，評価用データの HTER が公開されていないため，検証用データを使用した．

表 2: WMT20 および WMT21 の TQE タスクにおける HTER とのピアソンの積率相関係数

	WMT20		WMT21					
	多資源言語対		多資源言語対		中資源言語対		少資源言語対	
モデル	en-de	en-zh	en-de	en-zh	ro-en	et-en	ne-en	si-en
LaBSE	-0.104	-0.238	-0.119	-0.122	-0.748	-0.547	-0.451	-0.420
MEAT（小規模）	-0.216	-0.387	-0.208	**-0.218**	-0.763	-0.616	-0.516	-0.502
MEAT（大規模）	-0.216	-0.384	-0.208	-0.215	-0.763	-0.614	-0.516	-0.501
M2M-100（FT なし）	-0.323	-0.333	-0.269	-0.189	**-0.808**	**-0.679**	-0.248	-0.386
M2M-100（FT あり）	-0.361	-0.340	-0.287	-0.200	-0.718	-0.674	-0.394	-0.389
LaBSE+MLP（split）	0.423	0.466	0.324	0.194	0.407	0.432	0.466	0.527
LaBSE+MLP（concat）	**0.483**	**0.548**	**0.352**	**0.218**	0.649	0.569	**0.598**	**0.611**
XLM-R_{Base}+MLP （concat）	0.456	0.492	0.282	0.180	0.593	0.480	0.434	0.536

（en-zh）の 2 言語対[2]が含まれる．WMT21 の TQE タスクには，6 つの言語対[3]が含まれる．WMT20 と同じ多資源言語対の 2 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）の少資源言語対である．各言語対において，1,000 文対の原言語文および機械翻訳文と，HTER の組が提供されている．評価対象の機械翻訳は，fairseq ツールキット[4] [18] を用いて訓練された Transformer モデル [16] である．

疑似訓練データ　疑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[5]（表 1）および M2M-100[6] [19] 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，MEAT [11] と同量（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 [20] 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 [21]（$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon = 1e-8$）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．疑似訓練データ中の各対の TER を SacreBLEU[7] [22] によって計算し，TER が 1 以下のもののみを使用した．

TQE モデル　多言語文符号化器に LaBSE[8] [14] または XLM-R_{Base}[9] [23]，MLP に 1 層のフィードフォワードニューラルネットワークを用いた．各モデルは，HuggingFace Transformers を用い，バッチサイズを 128 文対，損失関数を RMSE，最適化手法を AdamW（$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon = 1e-8$）として訓練し，訓練済みの多言語文符号化器のパラメタも更新した．1 万ステップごとに検証用データに対する損失を計算し，この損失が 3 回改善しない場合に訓練を終了した．検証用データは，表 1 の疑似訓練データから 10% を無作為に抽出して作成した．学習率を $5e-5$，$1e-5$，$5e-6$ および $1e-6$ として 4 つのモデルを訓練し，これらのうち，検証用データにおいて最も損失の小さいモデルを選択した．

LaBSE では [CLS] トークンが事前に最適化されているのに対して，XLM-R では最適化されていない．この 2 モデルの比較により，事前訓練による [CLS] トークンの最適化の有効性について考察する．

比較手法　既存の教師なし TQE 手法として，次の 5 種類を用いた．LaBSE [14] のベースラインは，原言語文と機械翻訳文をそれぞれベクトル化し，それらの余弦類似度を機械翻訳文の品質の推定値とする．MEAT [11] は，LaBSE のベクトル表現から意味表現を抽出して用いるものである．MEAT の訓練データとして，先行研究 [11] と同様に表 1 の対訳

2) https://statmt.org/wmt20/quality-estimation-task.html
3) https://github.com/sheffieldnlp/mlqe-pe
4) https://github.com/pytorch/fairseq
5) https://www.statmt.org/wmt21/quality-estimation-task.html
6) https://huggingface.co/facebook/m2m100_418M
7) https://github.com/mjpost/sacrebleu
8) https://huggingface.co/sentence-transformers/LaBSE
9) https://huggingface.co/xlm-roberta-base

表 3: 訓練データに含まれる翻訳方向数が異なる場合の HTER とのピアソンの積率相関係数の比較

	WMT20		WMT21					
	多資源言語対		多資源言語対		中資源言語対		少資源言語対	
モデル	en-de	en-zh	en-de	en-zh	ro-en	et-en	ne-en	si-en
LaBSE+MLP（concat，6 つの翻訳方向）	**0.483**	**0.548**	**0.352**	**0.218**	**0.649**	0.569	**0.598**	**0.611**
LaBSE+MLP（concat，1 つの翻訳方向）	0.466	0.475	0.344	0.184	0.511	**0.602**	0.268	0.586

コーパスの一部（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）を使用した場合および全文を使用した場合を比較した．さらに，系列変換に基づく教師なし TQE として，原言語文から機械翻訳文への forced decode スコアを用いる手法と比較した．機械翻訳器として M2M-100 を使用した．ただし，翻訳方向ごとにファインチューニング（FT）を行ったモデルと行っていないモデルの両方を比較した．前者は，疑似訓練データの作成に使用したモデルである．

良い機械翻訳文に対して，HTER は小さな値をとる一方で，余弦類似度および forced decode スコアは大きな値をとる．つまり，HTER と余弦類似度および forced decode スコアの間には負の相関がある．

4.2 実験結果

表 2 に実験結果を示す．疑似データを用いて訓練した LaBSE+MLP（concat）は，多資源および少資源言語対において LaBSE および MEAT よりも優れた結果を示した．LaBSE+MLP（concat）は，全翻訳方向において LaBSE+MLP（split）および XLM-R_{Base} の性能を上回った．このことから，文対を連結することで原言語文と機械翻訳文のトークン間の関連を捉えつつ符号化を行うこと，および事前訓練による [CLS] トークンの符号化の最適化が有効であることが示唆される．ただし，LaBSE のパラメタ数が 470M であるのに対して，XLM-R_{Base} のパラメタ数は 280M と小さい．したがって，より大きな XLM-R[10] を用いることで，性能が向上する可能性がある．また，MEAT の性能は，使用する対訳文数を増加させても向上しなかった．M2M-100 を用いた系列変換に基づく TQE と比較しても，多資源言語対および少資源言語対において優れた性能を示した．

10） https://huggingface.co/xlm-roberta-large など

5 分析

言語モデルやニューラル機械翻訳器は，単一言語および単一翻訳方向よりも複数言語および複数翻訳方向の訓練を行うことで，性能が向上することが報告されている [23, 24]．TQE タスクにおいても複数翻訳方向による訓練が有効であるかを明らかにするために，4 節と同じ TQE タスクにおいて，6 つの翻訳方向を訓練データに含めて 1 つのモデルを訓練した場合と翻訳方向ごとにモデルを訓練した場合を比較した．

実験結果を表 3 に示す．1 つの例外を除いて，6 つの翻訳方向の疑似訓練データを用いたモデルの方が優れた性能を示した．この結果から，疑似訓練データに基づく教師なし TQE において，複数の翻訳方向を含めることが有効であるとわかった．

6 今後の課題

本研究では翻訳品質として HTER を扱ったが，DA スコアも MT の実利用では有用であり，WMT20 および WMT21 にも DA スコアを予測する TQE タスクがある．DA スコアの予測において，本稿の提案手法は MEAT の性能を上回ることはできなかった．このことから，TER に基づく疑似訓練データでは，DA スコアとの相関を向上させることは困難であり，DA スコアを正確に推定するには異なる手法が必要であるといえる．

7 おわりに

本稿では，疑似訓練データを用いた教師なし TQE 手法について述べた．WMT20 および WMT21 の TQE タスクにおける実験を通じて，既存の教師なし TQE 手法と比較して，多資源および少資源言語対を中心に優れた性能を達成しうることを確認した．分析の結果，言語モデルやニューラル機械翻訳器と同様に，複数の翻訳方向の訓練データを用いることにより，性能が向上することも明らかになった．

謝辞

本研究は国立研究開発法人情報通信研究機構の委託研究（課題番号：225）の助成を受けたものです.

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] Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. Findings of the WMT 2020 Shared Task on Quality Estimation. In **Proc. of WMT**, pp. 743–764, 2020.

[3] 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.

[4] Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. TransQuest: Translation Quality Estimation with Cross-lingual Transformers. In **Proc. of COLING**, pp. 5070–5081, 2020.

[5] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Vishrav Chaudhary, Mark Fishel, Francisco Guzmán, and Lucia Specia. BERGAMOT-LATTE Submissions for the WMT20 Quality Estimation Shared Task. In **Proc. of WMT**, pp. 1010–1017, 2020.

[6] Akifumi Nakamachi, Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. TMUOU Submission for WMT20 Quality Estimation Shared Task. In **Proc. of WMT**, pp. 1037–1041, 2020.

[7] Jiayi Wang, Ke Wang, Boxing Chen, Yu Zhao, Weihua Luo, and Yuqi Zhang. QEMind: Alibaba's Submission to the WMT21 Quality Estimation Shared Task. In **Proc. of WMT**, pp. 948–954, 2021.

[8] Yimeng Chen, Chang Su, Yingtao Zhang, Yuxia Wang, Xiang Geng, Hao Yang, Shimin Tao, Guo Jiaxin, Wang Minghan, Min Zhang, Yujia Liu, and Shujian Huang. HW-TSC's Participation at WMT 2021 Quality Estimation Shared Task. In **Proc. of WMT**, pp. 890–896, 2021.

[9] 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.

[10] Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation. In **Proc. of EMNLP**, pp. 7764–7774, 2021.

[11] Yuto Kuroda, Tomoyuki Kajiwara, Yuki Arase, and Takashi Ninomiya. Adversarial Training on Disentangling Meaning and Language Representations for Unsupervised Quality Estimation. In **Proc. of COLING**, pp. 5240–5245, 2022.

[12] Brian Thompson and Matt Post. Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. In **Proc. of EMNLP**, pp. 90–121, 2020.

[13] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised Quality Estimation for Neural Machine Translation. **TACL**, Vol. 8, pp. 539–555, 2020.

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

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proc. of NAACL**, pp. 4171–4186, 2019.

[16] 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.

[17] 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 **Proc. of LREC**, pp. 4963–4974, 2022.

[18] 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.

[19] 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.

[20] 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.

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

[22] Matt Post. A Call for Clarity in Reporting BLEU Scores. In **Proc. of WMT**, pp. 186–191, 2018.

[23] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In **Proc. of ACL**, pp. 8440–8451, 2020.

[24] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. **arXiv:2008.00401**, 2020.