多言語文符号化器の言語表現と意味表現の分離に基づく機械翻訳の品質推定

黒田 勇斗 [1]　梶原 智之 [2]　荒瀬 由紀 [3]　二宮 崇 [2]

[1] 愛媛大学工学部　[2] 愛媛大学大学院理工学研究科　[3] 大阪大学大学院情報科学研究科
kuroda@ai.cs.ehime-u.ac.jp　kajiwara@cs.ehime-u.ac.jp
arase@ist.osaka-u.ac.jp　ninomiya@cs.ehime-u.ac.jp

概要

本研究では，多言語文符号化器に基づく文表現から言語固有の情報を取り除き，言語非依存な意味表現を抽出する．機械翻訳の品質推定における教師なし設定での実験の結果，提案手法は既存の多言語文符号化器に基づく手法を上回る性能を達成した．他のアプローチと比較しても，提案手法は少資源言語対において人手評価に対する最も高い相関を得た．

1　はじめに

機械翻訳の研究開発の場では BLEU [1] などの参照訳に基づく自動評価が行われているものの，実世界における機械翻訳の使用者は参照訳を事前に用意できない場合が多い．本研究では，機械翻訳の実世界での利用を進める上で重要な，参照訳を用いない生成文の自動評価（品質推定） [2] に取り組む．

機械翻訳に関する国際会議 WMT における品質推定タスク [3] を中心に，多言語 BERT（mBERT） [4] や XLM-RoBERTa（XLM-R） [5] などの事前訓練された多言語文符号化器に基づく教師あり品質推定モデル [6–8] が提案されてきた．しかし，これらの教師あり品質推定モデルは，再訓練に「原言語文・目的言語文・人手評価値」の 3 つ組を必要とする．このようなデータセットの作成は，原言語と目的言語の両方に精通したアノテータが必要となるため，非常にコストが高い．そのため，教師あり品質推定モデルは，わずかな言語対でしか構築できない．

人手評価値を用いず対訳コーパスのみで訓練する教師なし品質推定 [9] では，多言語文符号化器の言語特異性が問題となる．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，再訓練なしでは言語を超えての文間の意味的類似度を正確に推定できない．この課題に対して先行研究の DREAM [9] では，正例としての対訳コーパスおよび擬似的な負例を用いて，多言語文符号化器から得られる文のベクトル表現を言語固有の言語表現と言語非依存の意味表現に分離した．DREAM の意味表現は教師なし品質推定タスクにおいて人手評価との高い相関を達成したが，意味表現に言語固有の情報が含まれないことは保証されていない．

本研究では，意味表現に言語固有の情報が含まれないことを保証するための敵対的学習を用いて，多言語文符号化器から得られる文のベクトル表現を意味表現と言語表現に分離する．提案手法は訓練時に負例を必要としないため，先行研究よりも単純な構造で訓練できるという利点を持つ．WMT20 の品質推定タスク [3] における実験の結果，提案手法は既存の多言語文符号化器に基づく教師なし品質推定手法よりも高い人手評価との相関を達成した．また，復号器を用いる他のアプローチと比較しても，提案手法は少資源言語対において最高性能を更新した．

2　提案手法

図 1 に示すように，本研究では MLP_L および MLP_M の 2 つの多層パーセプトロンからなる自己符号化器を用いて，多言語文符号化器から得た文のベクトル表現を言語固有の言語表現と言語非依存の意味表現に分離する．MLP_L は言語表現を抽出し，MLP_M は意味表現を抽出するものである．言語表現と意味表現を足し合わせることで，文のベクトル表現が復元される．提案手法では，これらの多層パーセプトロンを，以下の 4 つの損失関数に基づく多言語のマルチタスク学習によって訓練する．

$$L = L_R + L_C + L_L + L_A \tag{1}$$

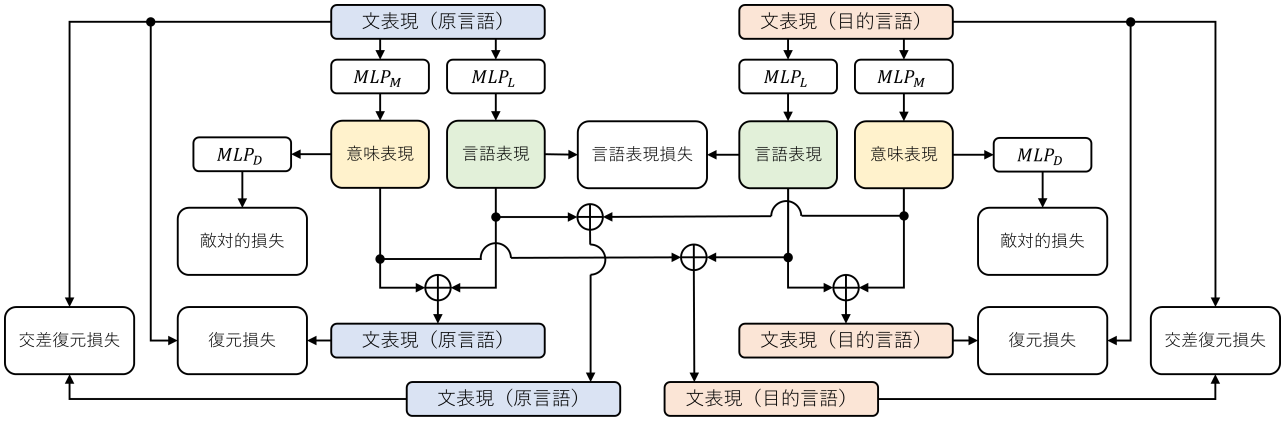

図 1　提案手法の概要

2.1　復元損失 L_R

復元損失は我々の自己符号化器を学習するための基本的な損失関数であり，意味表現 $\hat{\boldsymbol{e}}_M \in \mathbb{R}^d$ と言語表現 $\hat{\boldsymbol{e}}_L \in \mathbb{R}^d$ から文表現 $\boldsymbol{e} \in \mathbb{R}^d$ を復元できることを表している．ただし，d は文表現の次元数である．以下のように復元損失を定義[1)]する．

$$L_R = 1 - \cos(\boldsymbol{e}, (\hat{\boldsymbol{e}}_M + \hat{\boldsymbol{e}}_L)), \tag{2}$$

ここで，$\hat{\boldsymbol{e}}_M$ と $\hat{\boldsymbol{e}}_L$ は，それぞれ $\mathrm{MLP}_M(\cdot)$ と $\mathrm{MLP}_L(\cdot)$ によって抽出された意味表現と言語表現である．

$$\hat{\boldsymbol{e}}_M = \mathrm{MLP}_M(\boldsymbol{e}), \tag{3}$$

$$\hat{\boldsymbol{e}}_L = \mathrm{MLP}_L(\boldsymbol{e}). \tag{4}$$

2.2　交差復元損失 L_C

対訳文の原言語文 s と目的言語文 t は意味的に等価である．そこで，対訳文において意味表現同士を置換できることを保証するために，交差復元損失を用いる．これは，原言語の言語表現 $\hat{\boldsymbol{s}}_L$ と目的言語の意味表現 $\hat{\boldsymbol{t}}_M$ から原言語の文表現 $\boldsymbol{s}$ を復元でき，同様に目的言語の言語表現 $\hat{\boldsymbol{t}}_L$ と原言語の意味表現 $\hat{\boldsymbol{s}}_M$ から目的言語の文表現 $\boldsymbol{t}$ を復元できることを表している．以下のように交差復元損失を定義する．

$$L_C = 2 - \cos(\boldsymbol{s}, (\hat{\boldsymbol{s}}_L + \hat{\boldsymbol{t}}_M)) - \cos(\boldsymbol{t}, (\hat{\boldsymbol{s}}_M + \hat{\boldsymbol{t}}_L)). \tag{5}$$

2.3　言語表現損失 L_L

対訳文の原言語文 s と目的言語文 t は異なる言語の文である．そこで，対訳文において言語表現同士が類似しないことを保証するために，言語表現損失を用いる．以下のように言語表現損失を定義する．

$$L_L = \max(0, \cos(\hat{\boldsymbol{s}}_L, \hat{\boldsymbol{t}}_L)). \tag{6}$$

2.4　敵対的損失 L_A

本研究では，多言語文符号化器から得られる文表現を言語表現と意味表現に分離することによって，言語非依存の意味的類似度推定を実現したい．そこで，意味表現に言語固有の情報が含まれないことを保証するために，敵対的損失を用いる．これは，意味表現 $\hat{\boldsymbol{e}}_M$ から入力文の言語を識別できないことを表している．

敵対的訓練として言語を識別するため，新たに多層パーセプトロン MLP_D を用意する．以下のように意味表現から言語を識別する N クラス分類を行う．

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\mathrm{MLP}_D(\hat{\boldsymbol{e}}_M)), \tag{7}$$

ここで，$\mathrm{softmax}(\cdot)$ は softmax 関数を表す．MLP_D は，以下のように多クラス交差エントロピー損失を用いて訓練する．

$$L_D = -\sum_j \boldsymbol{y}_j \log \hat{\boldsymbol{y}}_j, \tag{8}$$

ここで，式 (8) は MLP_D を訓練するための損失関数であり，MLP_M および MLP_L を訓練するための式 (1) には含まれないことに注意されたい．

この敵対的モデルに対して，本研究では意味表現から言語を識別できないことを目指すため，言語識別における $\hat{\boldsymbol{y}}$ の分布を一様分布に近づける．以下のように敵対的損失を定義する．

$$L_A = -\sum_j \frac{1}{N} \log \hat{\boldsymbol{y}}_j, \tag{9}$$

ここで，N は訓練用データに含まれる言語の種類数である．言語の識別に関しては，意味表現から言語

1)　余弦類似度はベクトルの方向のみを考慮するため，本手法は厳密にはベクトルの復元を評価していない．しかし，先行研究 [9] で用いられた平均二乗誤差よりも高い性能が得られるため，本手法を採用した．詳しい検証は今後の課題である．

を識別可能にする式 (8) の訓練と，意味表現から言語を識別不可能にする式 (9) の訓練の両方によって，敵対的な訓練を行う．

3 評価実験

WMT20 の品質推定タスク [3] において提案手法の性能を評価する．提案手法では，原言語文および機械翻訳による出力文（目的言語文）を多言語文符号化器によって文表現に変換し，それぞれの意味表現を抽出する．品質推定には，意味表現間の余弦類似度を用いる．公式の評価方法に従い，モデルが推定した翻訳品質と人手評価値の間のピアソン相関によって性能評価を行う．

3.1 実験設定

データセット　WMT20 の品質推定タスクには，6 つの言語対[2]が含まれる．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）の少資源言語対である．各言語対において，1,000 文対の原文および機械翻訳の出力文と，人手評価値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[3] [10] を用いて訓練された Transformer モデル [11] である．

我々のモデルの訓練には，WMT20 の品質推定タスクで利用可能な対訳コーパスの一部[4]を使用した．多資源言語対は 100 万文対ずつ，中資源言語対は 20 万文対ずつ，少資源言語対は 5 万文対ずつの対訳コーパスを用いて訓練した．

モデル　本研究では，全ての MLP（MLP_M, MLP_L, MLP_D）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．文表現を得るための多言語文符号化器には，先行研究 [9] において品質推定の性能が最高であった LaBSE[5] [12] を用いた．文表現には，LaBSE の [CLS] トークンに対応する最終層の出力を用いた．なお，対訳コーパスを使用して訓練するのは MLP のみであり，LaBSE は再訓練しない．

我々のモデルは，バッチサイズを 512，最適化手法を Adam [13]，学習率を $1e-5$ として HuggingFace Transformers [14] を用いて訓練した．検証用データにおける式 (1) の損失が 10 エポック改善しない場合に訓練を終了した．検証用データは，訓練用データから 10% を無作為抽出して作成した．

比較手法　本実験では，教師なし品質推定の既存手法と提案手法を比較する．LaBSE [12] のベースラインは，提案手法による意味表現の抽出を行う前の文表現を用いて品質推定を行う．LaBSE から意味表現を抽出する先行研究として，DREAM[6] [9] と比較する．その他の多言語文符号化器による品質推定として，LASER[7] [15, 16]・mSBERT[8] [17]・BERTScore[9] [18] の 3 手法と比較する．なお，各モデルの事前訓練の設定に従い，LASER では双方向 LSTM の最終層の最大プーリング，mSBERT では最終層の平均プーリングを，それぞれ文表現として用いた．BERTScore には，最高性能が報告されている xlm-roberta-large モデルを使用した．

また，参考のために，系列変換モデルに基づく教師なし品質推定手法である D-TP [19] および Prism[10] [20]，WMT20 の品質推定タスクでベースラインとして採用されている教師あり品質推定手法である Predictor-Estimator[11] [22] とも比較する．

3.2 実験結果

表 1 に実験結果を示す．上段には，LaBSE ベースラインおよび LaBSE から抽出した意味表現による品質推定の結果を示している．まず，LaBSE と提案手法を比較すると，提案手法によって全ての言語対において性能が向上しており，本研究での意味表現の抽出が有効であることがわかる．また，DREAM との比較においても，全ての言語対において改善が見られるため，提案手法によってより良い意味表現を抽出できていると考えられる．

表 1 の中段には，多言語文符号化器による教師なし品質推定の比較手法の性能を示している．これらと比較して，提案手法は少資源言語対を中心に優れた結果を示しており，6 言語対の平均値としては人手評価との最も高い相関を達成している．

表 1 の下段には，他のアプローチによる品質推定

2） https://github.com/facebookresearch/mlqe
3） https://github.com/pytorch/fairseq
4） 先行研究 [9] と同程度の量を http://www.statmt.org/wmt20/quality-estimation-task.html から無作為抽出した．
5） https://huggingface.co/sentence-transformers/LaBSE
6） https://github.com/nattaptiy/qe_disentangled
7） https://github.com/facebookresearch/LASER
8） https://huggingface.co/sentence-transformers/stsb-xlm-r-multilingual
9） https://github.com/Tiiiger/bert_score
10） https://github.com/thompsonb/prism
11） Predictor-Estimator モデルの実装には OpenKiwi (https://github.com/Unbabel/OpenKiwi) [21] を用いた．

表 1　WMT20 品質推定タスクにおける人手評価とのピアソン相関係数

	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	ro-en	et-en	ne-en	si-en	Avg.
LaBSE	0.084	0.036	0.705	0.550	0.545	0.455	0.396
DREAM	0.151	0.156	0.711	0.549	0.627	0.552	0.458
提案手法	**0.216**	0.222	0.718	**0.587**	**0.634**	**0.571**	**0.491**
LASER	0.105	0.106	0.705	0.463	-	0.325	0.341
mSBERT	0.130	**0.287**	**0.766**	0.512	0.467	0.418	0.430
BERTScore	0.134	0.143	0.746	0.568	0.562	0.549	0.450
D-TP	0.259	0.321	0.693	0.642	0.558	0.460	0.489
Prism	0.464	0.303	0.829	0.694	-	-	0.573
Predictor-Estimator	0.145	0.190	0.685	0.477	0.386	0.374	0.376

の性能を示している．全言語対において，提案手法は教師ありベースラインである Predictor-Estimator の性能を上回った．多言語文符号化器に基づく教師なし品質推定手法（上段および中段）の中で，全ての言語対において教師ありベースラインを上回るのは提案手法のみであり，提案手法の有効性が明らかになった．また，教師なし品質推定手法である D-TP および Prism と比較して，少資源言語対においては提案手法が優れた性能を示しており，提案手法は少資源言語対における教師なし品質推定の最高性能を更新した．なお，D-TP は評価対象の機械翻訳の内部状態にアクセスする必要があり，Prism は大規模な対訳コーパスを用いて訓練する必要があるため，オンライン機械翻訳サービスなどのブラックボックス機械翻訳の品質推定や少資源言語対における品質推定において，提案手法は特に有効であると言える．

3.3　アブレーション分析

提案手法で用いている式 (1) の 4 種類の損失関数の影響を考察するために，同じく WMT20 の品質推定タスクにおいてアブレーション分析を行った．表 2 に分析結果を示す．なお，表 2 の上段では基本となる復元損失 L_R とその他の 1 種類の損失を組み合わせた際の性能を，下段では 1 種類ずつの損失を取り除いた際の性能を，それぞれ示している．

全ての損失を用いた際の性能は 0.491 であるため，(g) および (h) から，復元損失 L_R および交差復元損失 L_C の影響は小さいことがわかる．上段の結果から，交差復元損失 L_C および言語表現損失 L_L とは異なり，敵対的損失 L_A は基本の復元損失 L_R のみとの組み合わせでも有効であることがわかる．下段の結果から，言語表現損失 L_L を除外した際に最も性能が低下するため，言語表現損失からは他の損失で補完できない重要な情報が学習されると言える．本分析から，言語表現同士が類似しないことを保証する言語表現損失および意味表現に言語固有の情報が含まれないことを保証する敵対的損失の 2 つが，提案手法の性能改善に特に貢献することがわかった．

表 2　アブレーション分析におけるピアソン相関係数

	L_R	L_C	L_L	L_A	Avg.
(a)	✓				0.390
(b)	✓	✓			0.072
(c)	✓		✓		0.082
(d)	✓			✓	0.434
(e)	✓	✓	✓		0.439
(f)	✓	✓		✓	0.327
(g)	✓		✓	✓	0.483
(h)		✓	✓	✓	0.488

4　おわりに

本研究では，事前訓練された多言語文符号化器に基づく教師なし品質推定に取り組んだ．提案手法は，文表現から言語固有の情報を取り除くことで言語非依存の意味表現を抽出し，言語を超えた文間の意味的類似度の推定を可能にした．6 つの言語対における実験の結果，提案手法は最先端の多言語文符号化器の性能を一貫して改善し，特に少資源言語対において教師なし品質推定の最高性能を達成した．

謝辞

本研究は JSPS 科研費（若手研究，課題番号：JP20K19861）の助成を受けたものです.

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In **Proc. of ACL**, p. 311–318, 2002.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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. 6000–6010, 2017.

[12] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT Sentence Embedding. **arXiv:2007.01852**, 2020.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In **Proc. of ICLR**, 2015.

[14] 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.

[15] Mikel Artetxe and Holger Schwenk. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. In **Proc. of ACL**, pp. 3197–3203, 2019.

[16] Mikel Artetxe and Holger Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 597–610, 2019.

[17] Nils Reimers and Iryna Gurevych. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In **Proc. of EMNLP**, pp. 4512–4525, 2020.

[18] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In **Proc. of ICLR**, pp. 1–43, 2020.

[19] 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. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 539–555, 2020.

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

[21] Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. OpenKiwi: An Open Source Framework for Quality Estimation. In **Proc. of ACL**, pp. 117–122, 2019.

[22] Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. Predictor-Estimator: Neural Quality Estimation Based on Target Word Prediction for Machine Translation. **ACM Transactions on Asian and Low-Resource Language Information Processing**, Vol. 17, No. 1, pp. 1–22, 2017.