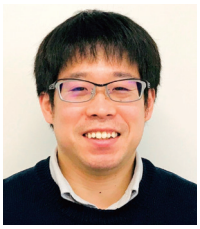
多言語文符号化器による機械翻訳の教師なし品質推定の研究

Studies on Unsupervised Quality Estimation for Machine Translation with Multilingual Sentence Encoders



愛媛大学大学院理工学研究科助教

梶原　智之

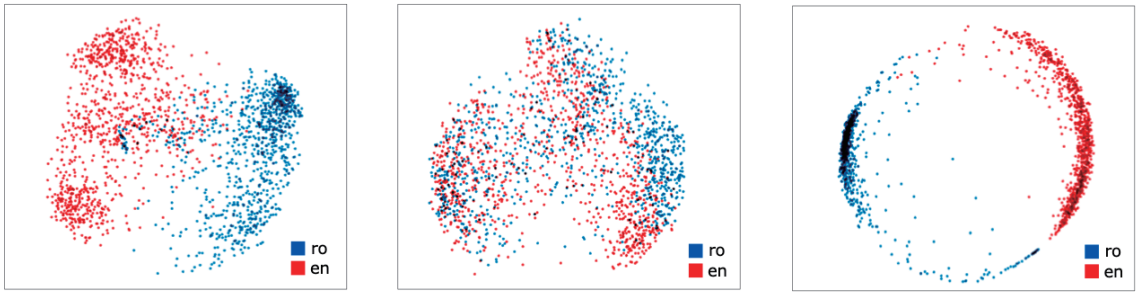
2018 年首都大学東京大学院システムデザイン研究科博士後期課程修了。博士（工学）。大阪大学データビリティフロンティア機構の特任助教を経て、2021 年より現職。言い換え生成や品質推定など自然言語処理の研究に従事。

1　はじめに

機械翻訳の研究開発の場では、翻訳品質は BLEU [1] や METEOR [2] などの参照訳に基づく自動評価によって推定されている。これらの自動評価指標は、ベンチマーク上での機械翻訳の性能改善に貢献してきたが、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。本稿では、機械翻訳の実世界での利用を進める上で重要な、参照訳を用いない品質推定（Quality Estimation; QE）[3] について解説する。品質推定では、原文とそれに対応する翻訳文を比較することで、機械翻訳の品質を推定する。人手評価との相関が高い品質推定の手法を開発することにより、機械翻訳の出力文をそのまま使用するか、手動または自動で後編集するか、他の機械翻訳を使用するかという判断を支援できる。

機械翻訳に関する国際会議 WMT（Workshop on Statistical Machine Translation ／ Conference on Machine Translation）にて開催されている品質推定コンペティション [4] を中心に、これまで多くの教師あり品質推定の手法 [5-9] 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の訓練には、「原言語文・機械翻訳の出力文・人手評価値」の 3 つ組が必要である。このような品質推定の訓練用データの構築は、翻訳者などの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、WMT の品質推定コンペティションに含まれるような、わずかな言語対でしか教師あり品質推定モデルを得られないのが現状である。

教師なし品質推定は、このような人手評価値のラベル付きデータの代わりに、機械翻訳の訓練に使用されるのと同じ対訳コーパス、つまり、原言語文と目的言語文の対のみを用いて訓練する。多言語 BERT [10] をはじめとする多言語文符号化器の活用は、教師なし品質推定のための有望なアプローチである。しかし、多言語文符号化器から得られる文のベクトル表現は、図 1（左）に示すように意味よりも言語の影響を強く受けており、再訓



mBERT　　意味表現　　言語表現

図 1　英語とルーマニア語のパラレルコーパスから無作為抽出した 1,000 文対の PCA による可視化

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。

本稿では、国際会議 EMNLP-2021（The 2021 Conference on Empirical Methods in Natural Language Processing）および COLING-2022（The 29th International Conference on Computational Linguistics）において我々が提案した教師なし品質推定の手法 [11-12] について解説する。これらの提案手法では、多言語文符号化器から得られる文のベクトル表現を、図 1（右）に示す言語固有の「言語表現」と図 1（中）に示す言語非依存の「意味表現」に分離する。この意味表現間の余弦類似度によって、言語を超えて文間の意味的類似度を推定でき、人手評価値のラベル付きデータを用いることなく機械翻訳の品質推定を実現できる。

2 DREAM:Disentangled REpresentatioin for language-Agnostic Meaning (EMNLP-2021)

多言語文符号化器は、品質推定をはじめとするクロスリンガル言語理解のために有用である。しかし、図 1（左）に示したように、多言語文符号化器から得られる文表現は、言語固有の情報による影響を強く受け、再訓練なしでは言語を超えての文間の意味的類似度の推定が難しい。

我々の目的は、多言語文符号化器から得られる文表現から言語固有の情報を取り除き、言語非依存の意味情報を表現する「意味表現」を抽出することである。本研究では、対訳コーパスを用いてこれを実現する方法を提案する。

提案手法は、図 2 に示すように、2 つの多層パーセプトロン（Multi-Layer Perceptron; MLP）から構成される。多言語文符号化器から得られる文のベクトル表現を入力として、MLP_L が言語固有の情報（言語表現）を抽出し、MLP_M が言語非依存の情報（意味情報）を抽出する。そして、言語表現と意味表現のベクトルを足し合わせると、元の多言語文符号化器から得た文ベクトルが復元されるという制約のもとで、2 つの多層パーセプトロンを学習する。

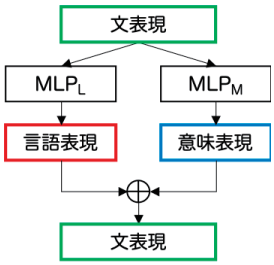


図 2　アイデアの中核となる自己符号化器

これらの多層パーセプトロンを、上記の復元損失 L_R の他に、意味に関する損失 L_M および言語に関する損失 L_L を用いて、対訳コーパス上で訓練する。つまり、式（1）を最小化するような多言語のマルチタスク学習を行う。

$$L = L_R + L_M + L_L \qquad (1)$$

図 3 に提案手法の概要を示す。本手法では、対訳コーパスに含まれる（a）原言語文および（b）目的言語文の文対に加えて、（c）原言語の中から無作為に選択された文および（d）目的言語の中から無作為に選択され

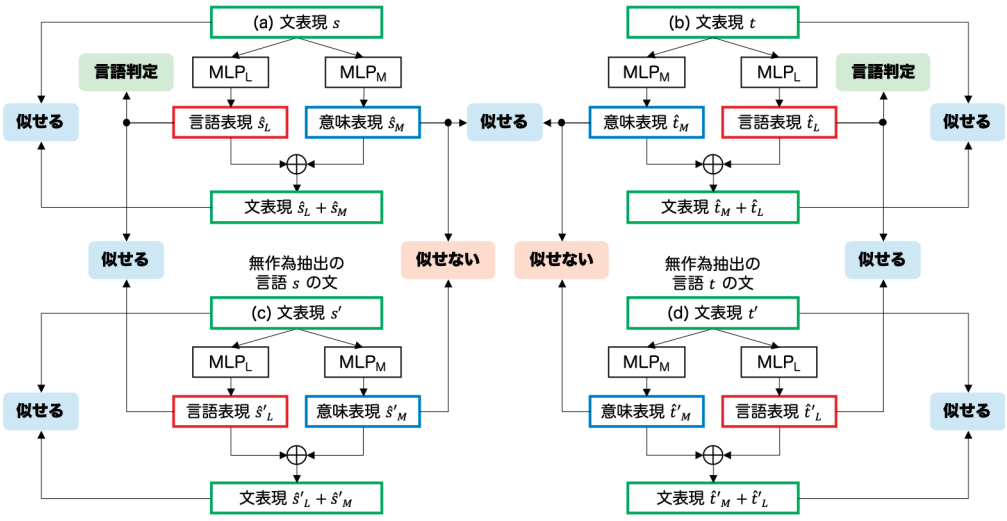


図 3　DREAM における意味表現の抽出

た文の４文を用いて MLP_M および MLP_L を訓練する。(a) と (b) は意味的に対応するが異言語の文対、(a) と (c) および (b) と (d) は意味的な対応は持たないが同一言語の文対である。これらの文間の関係を考慮して、(a) と (b) の間では意味表現を近づけ、(a) と (c) および (b) と (d) の間では言語表現を近づけつつ意味表現を遠ざける。さらに、言語表現が各言語に固有の情報を持つことを保証するために、言語表現からはその文が何語なのかを判定する。

2.1 復元損失 L_R

復元損失 L_R は、図２の自己符号化器を訓練するための基本的な損失関数であり、意味表現 $\hat{e}_M \in \mathbb{R}^d$ と言語表現 $\hat{e}_L \in \mathbb{R}^d$ から元の文表現 $e \in \mathbb{R}^d$ を復元できることを表している。これらの文表現の次元数を d として、復元損失を以下のように定義する。

$$L_R = \frac{1}{d}\|e - (\hat{e}_M + \hat{e}_L)\|_2^2 \qquad (2)$$

ここで、$\hat{e}_M$ と $\hat{e}_L$ は、それぞれ $\mathrm{MLP}_M(\cdot)$ と $\mathrm{MLP}_L(\cdot)$ によって元の文表現 e から抽出された意味表現と言語表現である。

$$\hat{e}_M = \mathrm{MLP}_M(e) \qquad (3)$$
$$\hat{e}_L = \mathrm{MLP}_L(e) \qquad (4)$$

2.2 意味に関する損失 L_M

式 (1) における意味に関する損失 L_M は、$\mathrm{MLP}_M(\cdot)$ が言語非依存の表現 $\hat{e}_M$ を抽出するための制約である。意味的に対応する (a) と (b) の文間では意味表現を近づけ、意味的に対応しない (a) と (c) および (b) と (d) の文間では意味表現を遠ざける。それぞれ、以下の式 (5) における L_M^x および L_M^m の各損失に対応する。

$$L_M = L_M^x + L_M^m \qquad (5)$$

対訳文間で意味表現を近づけるための損失 L_M^x は、原言語文から抽出した意味表現 $\hat{s}_M \in \mathbb{R}^d$ と目的言語文から抽出した意味表現 $\hat{t}_M \in \mathbb{R}^d$ の間の類似度によって以下のように定義される。なお、本研究では $\phi(\cdot)$ として余弦類似度を計算する。

$$L_M^x = 1 - \phi(\hat{s}_M, \hat{t}_M) \qquad (6)$$

無作為に選択された同一言語内の文間で意味表現を遠ざけるための損失 L_M^m は、原言語文から抽出した意味表現 $\hat{s}_M$ および $\hat{s}'_M \in \mathbb{R}^d$ と目的言語文から抽出した意味表現 $\hat{t}_M$ および $\hat{t}'_M \in \mathbb{R}^d$ の間の類似度によって以下のように定義される。

$$L_M^m = \max\left(0, \phi(\hat{s}_M, \hat{s}'_M)\right) + \max\left(0, \phi(\hat{t}_M, \hat{t}'_M)\right) \qquad (7)$$

2.3 言語に関する損失 L_L

式 (1) における言語に関する損失 L_L は、$\mathrm{MLP}_L(\cdot)$ が言語固有の表現 $\hat{e}_L$ を抽出するための制約である。同一言語である (a) と (c) および (b) と (d) の文間で言語表現を近づける。また、$\hat{e}_L$ が各言語に固有の情報を持つことを保証するために、その文が何語なのかを判定する言語判定を行う。それぞれ、以下の式 (8) における L_M^m および L_L^i の各損失に対応する。

$$L_L = L_L^m + L_L^i \qquad (8)$$

同一言語内の文間で言語表現を近づけるための損失 L_L^m は、原言語文から抽出した言語表現 $\hat{s}_L \in \mathbb{R}^d$ と目的言語文から抽出した言語表現 $\hat{t}_L \in \mathbb{R}^d$ の間の類似度によって以下のように定義される。

$$L_L^m = 2 - \phi(\hat{s}_L, \hat{s}'_L) - \phi(\hat{t}_L, \hat{t}'_L) \qquad (9)$$

このように、式 (7) の制約と式 (9) の制約を用いて、無作為に選択された同一言語内の文間で対照的な学習を行うことによって、意味表現と言語表現を明確に分離することを意図している。

最後の損失 L_L^i は、言語判定の学習を通じて言語表現 $\hat{e}_L$ として各言語に固有の表現を得ることを目指すものである。新たに多層パーセプトロン MLP_I を用意して式 (10) のように言語判定を行い、多クラス分類におけるクロスエントロピー損失として式 (11) のように L_L^i を計算する。

$$\hat{y} = \mathrm{softmax}\left(\mathrm{MLP}_I(\hat{e}_L)\right) \qquad (10)$$

$$L_L^i = -\sum_j y_j \log \hat{y}_j \qquad (11)$$

表 1　WMT20 品質推定タスクにおける DREAM と人手評価とのピアソン相関

	多資源言語対		中資源言語対		少資源言語対		
	en-de	en-zh	ro-en	et-en	ne-en	si-en	平均
mBERT	0.071	0.010	0.182	0.009	0.025	-	0.056
mBERT + DREAM	**0.125**	**0.131**	**0.663**	**0.354**	**0.400**	-	**0.335**
XLM-R	0.061	0.007	0.151	0.016	0.008	0.148	0.063
XLM-R + DREAM	**0.093**	**0.120**	**0.647**	**0.334**	**0.310**	**0.227**	**0.289**
LaBSE	0.084	0.036	0.705	**0.550**	0.545	0.455	0.396
LaBSE + DREAM	**0.151**	**0.156**	**0.711**	0.549	**0.627**	**0.552**	**0.458**
LASER	0.105	0.106	0.705	0.463	-	0.325	0.341
Predictor-Estimator	0.145	0.190	0.685	0.477	0.386	0.374	0.376

2.4　実験設定

WMT20 の品質推定タスク [4] において提案手法の有効性を評価する。提案手法では、原言語文および機械翻訳の出力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、意味表現の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品質推定の性能をメタ評価する。

2.4.1　データセット

WMT20 の品質推定タスク [1] には、6 言語対が含まれる。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 文対の原言語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 [2] [13] を用いて訓練された Transformer モデル [14] である。

我々の品質推定モデルの訓練には、WMT20 の品質推定タスクで利用可能な対訳コーパスの一部を使用した。多資源言語対においては 100 万文対ずつ、中資源言語対においては 20 万文対ずつ、少資源言語対においては 5 万文対ずつの対訳コーパスを用いて訓練した。

2.4.2　モデル

本研究では、全ての MLP（$\mathrm{MLP}_M \cdot \mathrm{MLP}_L \cdot \mathrm{MLP}_I$）に、1 層のフィードフォーワードニューラルネットワークを用いた。多言語文符号化器としては、代表的な mBERT[3] [10]、XLM-R[4][15]、LaBSE[5][16] の 3 つを用いた。文表現には、それぞれの多言語文符号化器における [CLS] トークンに対応する最終層の出力ベクトルを用いた。なお、対訳コーパスを用いて訓練するのは MLP のみであり、多言語文符号化器そのものは再訓練しない。

我々の品質推定モデルは、バッチサイズを 512 文、最適化手法を Adam [17]、学習率を $1e{-}5$ として HuggingFace Transformers[18] 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。なお、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。

2.4.3　比較手法

WMT19 の品質推定タスク [19] でベースライン手法として採用されている LASER [20] および WMT20 の品質推定タスク [4] でベースラインとして採用されている Predictor-Estimator [6] を提案手法と比較する。LASER は再帰型ニューラルネットワークに基づく多言語機械翻訳の符号化器部分である。LASER による品質推定は教師なし手法であり、原言語文と機械翻訳の出力文をそれぞれ LASER によってベクトル化し、それらの余弦類似度によって翻訳品質を推定する。Predictor-Estimator は、対訳コーパス上で目的言語文の各単語を原言語文および目的言語文中の周辺単語から推定するように事前訓練された Predictor と、Predictor によっ

1　https://github.com/facebookresearch/mlqe
2　https://github.com/pytorch/fairseq
3　https://huggingface.co/bert-base-multilingual-cased
4　https://huggingface.co/xlm-roberta-large
5　https://huggingface.co/sentence-transformers/LaBSE

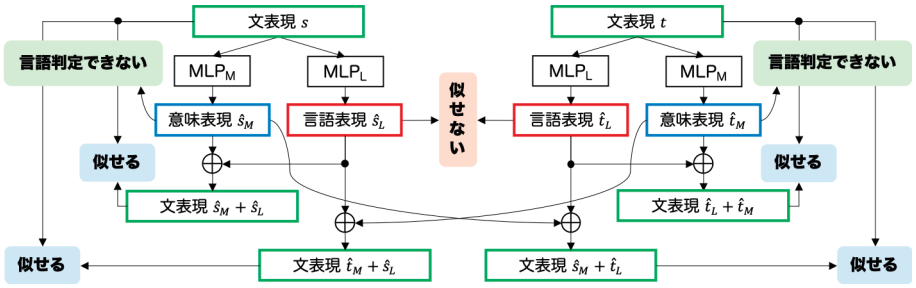


図 4　MEAT における意味表現の抽出

て得られる特徴表現から品質推定を行う Estimator で構成される教師あり品質推定の手法である。

2.5　実験結果

実験結果を表 1 に示す。mBERT、XLM-R、LaBSE の全ての多言語文符号化器に対して、提案手法の意味表現の抽出によって品質推定の性能が改善できることがわかる。特に、LaBSE の多言語文符号化器との組み合わせによって、最も高い性能を達成することができた。図 1（左）に示したように、mBERT や XLM-R は再訓練なしの状態での品質推定器としての性能は低いが、提案手法によって著しい性能改善が見られた。

LaBSE に提案手法を適用することで、教師なし品質推定の LASER や教師あり品質推定の Predictor-Estimator と比較して、多くの言語対において人手評価との高い相関を達成することができた。特に、少資源言語対において、著しい性能改善が見られた。

3　MEAT:Meaning Embedding by Adversarial Training (COLING-2022)

2 章では、正例としての対訳コーパスおよび無作為に選択された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する手法 DREAM を提案した。DREAM の意味表現は教師なし品質推定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。

本章では、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。本手法は、訓練時に負例を必要としないため、DREAM よりも単純な構造かつ少ない計算コストで訓練できる。

図 4 に示すように、本手法も 2 章と同じく 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。これらの言語表現と意味表現を足し合わせることで、元の文表現が復元される。本手法では、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。

$$L = L_R + L_C + L_L + L_A \tag{12}$$

3.1　復元損失 L_R

2 章とは異なり、余弦類似度によって復元損失を定義する。余弦類似度はベクトルの方向のみを考慮するため、厳密にはベクトルの復元を評価しないが、こちらの方が高い性能が得られるため、本章では以下の復元損失を採用する。

$$L_R = 1 - \phi\left(e - (\hat{e}_M + \hat{e}_L)\right) \tag{13}$$

3.2　交差復元損失 L_C

対訳コーパスに含まれる原言語文 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 $\hat{s}_L$ と目的言語の意味表現 $\hat{t}_M$ から原言語の文表現 s を復元でき、同様に目的言語の言語表現 $\hat{t}_L$ と原言語の意味表現 $\hat{s}_M$ から目的言語の文表現 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。

$$L_C = 2 - \phi\left(s, (\hat{s}_L + \hat{t}_M)\right) - \phi\left(t, (\hat{s}_M + \hat{t}_M)\right) \tag{14}$$

表 2　WMT20 品質推定タスクにおける MEAT と人手評価とのピアソン相関

	多資源言語対		中資源言語対		少資源言語対		
	en-de	en-zh	ro-en	et-en	ne-en	si-en	平均
LaBSE	0.084	0.036	0.705	0.550	0.545	0.455	0.396
LaBSE + DREAM	0.151	0.156	0.711	0.549	0.627	0.552	0.458
LaBSE + MEAT	**0.215**	**0.222**	**0.717**	**0.587**	**0.634**	**0.571**	**0.491**
LASER	0.105	0.106	0.705	0.463	-	0.325	0.341
Predictor-Estimator	0.145	0.190	0.685	0.477	0.386	0.374	0.376

3.3　言語表現損失 L_L

対訳コーパスに含まれる原言語文 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。

$$L_L = \max\left(0, \phi\left(\hat{s}_L, \hat{t}_L\right)\right) \tag{15}$$

3.4　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 $\hat{e}_M$ から入力文の言語を判定できないことを表している。

敵対的訓練として言語を判定するために、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判定する N クラス分類を行う。

$$\hat{y} = \mathrm{softmax}\left(\mathrm{MLP}_D(\hat{e}_M)\right) \tag{16}$$

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。

$$L_D = -\sum_j y_j \log \hat{y}_j \tag{17}$$

ここで、式（17）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（12）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判定できないことを目指すため、言語判定における $\hat{y}$ の分布を一様分布に近づける。以下にように敵対的損失を定義する。

$$L_A = -\sum_j \frac{1}{N} \log \hat{y}_j \tag{18}$$

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。言語の判定に関しては、意味表現から言語を判定可能にする式（17）の訓練と、意味表現から言語を判定不可能にする式（18）の訓練の両方によって、敵対的な訓練を行う。

3.5　評価実験

2.4 節と同じ設定で、提案手法の有効性を評価する。なお、本実験では多言語文符号化器として、2.5 節において最も高い性能を達成した LaBSE [16] を用いる。

実験結果を表 2 に示す。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、2 章で提案した 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、本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。

また、教師なし品質推定の LASER [20] や教師あり品質推定の Predictor-Estimator [6] と比較して、提案手法は全ての言語対において人手評価との高い相関を達成することができた。特に、少資源言語対において、著しい性能改善が見られた。

4　おわりに

本稿では、mBERT [10] や LaBSE [16] などの事前訓練された多言語文符号化器に基づく機械翻訳のための教師なし品質推定の手法について解説した。我々が提案する DREAM [11] および MEAT [12] は、多言語文符号化器によって得られる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、言語を超えて文間の意味的類似度推定を可能にした。

WMT20 の品質推定タスク [4] における 6 つの言語対における評価実験の結果、提案手法は最先端の多言語文符号化器である LaBSE の性能を一貫して改善し、人

手評価との高い相関を得た。特に、提案手法は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

本稿は、国際会議EMNLP-2021に採択された論文[11]および国際会議COLING-2022に採択された論文[12]に基づき、これらの論文を再構成して解説したものである。また、本研究はJSPS科研費（若手研究、課題番号：JP20K19861）および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）による助成を受けて実施した。

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 311–318.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72.

[3] Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality Estimation for Machine Translation. Synthesis Lectures on Human Language Technologies, 11（1）:1–162.

[4] Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 Shared Task on Quality Estimation. In Proceedings of the Fifth Conference on Machine Translation, pages 743–764.

[5] Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - A Translation Quality Estimation Framework. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 79–84.

[6] Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-Estimator: Neural Quality Estimation Based on Target Word Prediction for Machine Translation. ACM Transactions on Asian and Low-Resource Language Information Processing, 17（1）:1–22.

[7] Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation Quality Estimation with Cross-lingual Transformers. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5070–5081.

[8] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Vishrav Chaudhary, Mark Fishel, Francisco Guzmán, and Lucia Specia. 2020. BERGAMOT-LATTE Submissions for the WMT20 Quality Estimation Shared Task. In Proceedings of the Fifth Conference on Machine Translation, pages 1010–1017.

[9] Akifumi Nakamachi, Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. TMUOU Submission for WMT20 Quality Estimation Shared Task. In Proceedings of the Fifth Conference on Machine Translation, pages 1037–1041.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages

4171–4186.

[11] Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-Agnostic Representation from Multilingual Sentence Encoders for Cross-Lingual Similarity Estimation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7764–7774.

[12] Yuto Kuroda, Tomoyuki Kajiwara, Yuki Arase, and Takashi Ninomiya. 2022. Adversarial Training on Disentangling Meaning and Language Representations for Unsupervised Quality Estimation. In Proceedings of the 29th International Conference on Computational Linguistics.

[13] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. 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) , pages 48–53.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems, page 6000–6010.

[15] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451.

[16] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-Agnostic BERT Sentence Embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pages 878–891.

[17] Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, pages 1–15.

[18] 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. 2020. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45.

[19] Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 Shared Tasks on Quality Estimation. In Proceedings of the Fourth Conference on Machine Translation, pages 1–10.

[20] Mikel Artetxe and Holger Schwenk. 2019. Marginbased Parallel Corpus Mining with Multilingual Sentence Embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3197–3203.