Adversarial Training on Disentangling Meaning and Language Representations for Unsupervised Quality Estimation

Yuto Kuroda
Ehime University
kuroda@ai.cs.ehime-u.ac.jp

Tomoyuki Kajiwara
Ehime University
kajiwara@cs.ehime-u.ac.jp

Yuki Arase
Osaka University
arase@ist.osaka-u.ac.jp

Takashi Ninomiya
Ehime University
ninomiya@cs.ehime-u.ac.jp

Abstract

We propose a method to distill language-agnostic meaning embeddings from multilingual sentence encoders for unsupervised quality estimation of machine translation. Our method facilitates that the meaning embeddings focus on semantics by adversarial training that attempts to eliminate language-specific information. Experimental results on unsupervised quality estimation reveal that our method achieved higher correlations with human evaluations.

1 Introduction

Quality Estimation (QE) is a task of estimating translation quality without reference sentences (Specia et al., 2018). Reference-based automatic evaluation methods, such as BLEU (Papineni et al., 2002) and BLEURT (Sellam and Parikh, 2020), have contributed to research and development of machine translation; however, end-users of machine translation systems unlikely have such reference translations. Hence, the development of QE methods that correlate well with human evaluation is practically important.

Supervised QE models (Ranasinghe et al., 2020; Fomicheva et al., 2020a; Nakamachi et al., 2020) based on pre-trained multilingual sentence encoders (Conneau et al., 2020; Feng et al., 2022) have been actively proposed in the QE competitions (Specia et al., 2020). However, these models require bilingual sentence pairs with manually labeled translation quality scores for fine-tuning. Creating such a QE dataset is expensive because it requires annotators who are fluent in both of source and target languages. Therefore, supervised QE models are limited to several major language pairs included in the competitions.

In contrast, unsupervised QE allows quality estimation *without* human-assessed machine translation outputs. Instead of the annotated outputs, unsupervised QE utilizes widely available parallel corpora. Multilingual sentence encoders (Artetxe and Schwenk, 2019a,b; Reimers and Gurevych, 2020; Conneau et al., 2020; Feng et al., 2022) are promising for developing unsupervised QE models; however, their sentence embeddings are dominated by language-specific information. Due to this characteristic, these sentence embeddings form clusters by language rather than by meaning, which hinders precise estimation of semantic similarity across languages (Tiyajamorn et al., 2021). To address this problem, DREAM (Tiyajamorn et al., 2021) disentangles sentence embeddings to meaning and language embeddings. It conducts self-supervised learning using parallel sentence pairs in bilingual corpora as positive examples and random pairs as negative examples; meaning embeddings of positive pairs should be close while those of negative pairs should be distant. However, DREAM lacks direct supervision to eliminate language-specific information from the meaning embeddings and its architecture is complex.

We improve DREAM by introducing an adversarial training that attempts to remove language-specific information from the meaning embeddings.[1] Our adversarial training eliminates the random pairs that DREAM needs, which results in a simpler architecture and lighter computational costs for training. Experimental results on the WMT20 QE task (Specia et al., 2020) revealed that our method achieved higher correlations with human scores than previous unsupervised QE models based on multilingual sentence encoders. Com-

[1]The source code for this paper is available at https://github.com/kuro961/MEAT.

Proceedings of the 29th International Conference on Computational Linguistics, pages 5240–5245
October 12–17, 2022.

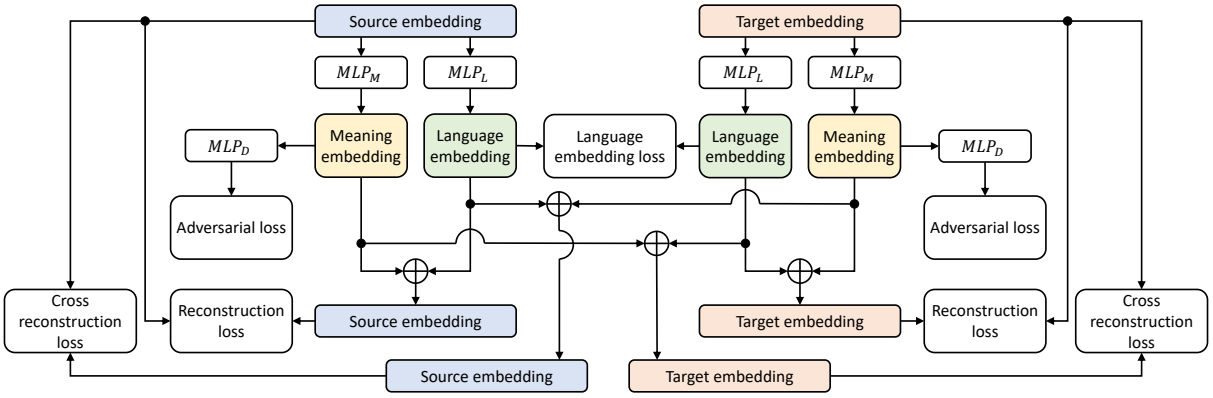

Figure 1: Multitask training for distilling meaning embeddings from multilingual sentence embeddings

pared to other approaches independent of multilingual sentence encoders (Fomicheva et al., 2020b; Thompson and Post, 2020), our method showed higher correlations in low-resource language pairs.

2 Proposed Method

Our model is an autoencoder comprising two multilayer perceptrons, MLP_M and MLP_L, trained with bilingual corpora as shown in Figure 1. The former is responsible for extracting meaning and the latter for extracting language-specific information. These outputs are summed to reconstruct the input sentence embedding. We train these MLPs using multilingual-multitask learning with the following four loss functions.

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

2.1 Reconstruction Loss L_R

L_R is the basis of the autoencoder training, which ensures that a meaning embedding $\hat{\boldsymbol{e}}_M \in \mathbb{R}^d$ and language embedding $\hat{\boldsymbol{e}}_L \in \mathbb{R}^d$ can reconstruct the input sentence embedding $\boldsymbol{e} \in \mathbb{R}^d$. Here, d is the dimension of the sentence embedding. We define reconstruction loss with cosine similarity[2] as:

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

2.2 Cross Reconstruction Loss L_C

Source and target sentences in Figure 1 are semantically equivalent as they are a parallel pair. Hence, their meaning embeddings should be interchangeable, for which we design a cross reconstruction loss L_C as:

$$L_C = 2 - \cos(\boldsymbol{s}, (\hat{\boldsymbol{s}}_L + \hat{\boldsymbol{t}}_M)) - \cos(\boldsymbol{t}, (\hat{\boldsymbol{t}}_L + \hat{\boldsymbol{s}}_M)). \tag{3}$$

The sentence embedding in the source language $\boldsymbol{s}$ should be reconstructed from its language embedding $\hat{\boldsymbol{s}}_L$ and the meaning embedding of the target language $\hat{\boldsymbol{t}}_M$. Similarly, the sentence embedding in the target language $\boldsymbol{t}$ should be reconstructed from its language embedding $\hat{\boldsymbol{t}}_L$ and the meaning embedding of the source language $\hat{\boldsymbol{s}}_M$.

2.3 Language Embedding Loss L_L

The source and target languages are different. To ensure that language embeddings of source and target are distinctive each other, we design a language embedding loss L_L as:

$$L_L = \max(0, \cos(\hat{\boldsymbol{s}}_L, \hat{\boldsymbol{t}}_L)). \tag{4}$$

2.4 Adversarial Loss L_A

We improve DREAM by giving direct supervision that eliminates language-specific information from the meaning embeddings. For this aim, we introduce an adversarial loss L_A that decrease language-identifiability from the meaning embeddings.

First, as an adversarial model that attempts to identify the language of the input sentence from its meaning embedding, we use the following multiclass classifier MLP_D:

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\mathrm{MLP}_D(\hat{\boldsymbol{e}}_M)). \tag{5}$$

MLP_D is trained using the cross-entropy loss:

$$L_D = -\sum_j \boldsymbol{y}_j \log \hat{\boldsymbol{y}}_j. \tag{6}$$

Note that Equation (6) is the loss function for training MLP_D, and is not included in Equation (1) for training MLP_M and MLP_L.

[2]This constraint does not strictly reconstruct the input embedding, because $\cos(\cdot)$ does not take into account the vector norm. We empirically employed $\cos(\cdot)$ for its higher performance than MSE used by Tiyajamorn et al. (2021).

With the adversarial model, we define L_A that supervises MLP_M to derive meaning embeddings from which languages are unidentifiable. Specifically, L_A makes the distribution of $\hat{\boldsymbol{y}}$ close to a uniform distribution:

$$L_A = -\frac{1}{N}\sum_j \log \hat{\boldsymbol{y}}_j, \tag{7}$$

where N is the number of language types in the training data. Adversarial training is performed simultaneously with the model training; Equation (6) trains the adversarial model to achieve higher language identifiability from the meaning embeddings while Equation (7) makes the meaning embeddings less language-identifiable.

2.5 Application to QE

Once our model is trained, we can use MLP_M to disentangle meaning embeddings from sentence representations generated by multilingual sentence encoders. We compute a QE score by a cosine similarity between meaning embeddings of source sentence s and translation output t:

$$\cos(\hat{\boldsymbol{s}}_M, \hat{\boldsymbol{t}}_M). \tag{8}$$

3 Evaluation

We evaluated the effectiveness of the proposed method in an unsupervised QE task.

3.1 Setting

Dataset Following the previous work (Tiyajamorn et al., 2021), we used six language pairs included in the WMT20 QE task[3] (Specia et al., 2020). For each language pair, the test set consists of 1k pairs of source and machine-translated output sentences manually labeled with a translation quality score. The evaluation metric is Pearson correlation coefficients between these human scores and model predictions.

We trained our model on the publicly available bilingual corpora that were used to train the target machine translation systems (Ott et al., 2019). We used bilingual corpora of 1M sentence pairs for high-resource (en-de and en-zh), 200k for medium-resource (ro-en and et-en), and 50k for low-resource (ne-en and si-en) language pairs.[4]

Model All the MLPs in our model are single-layer feedforward networks. As a multilingual sentence encoder to disentangle meaning embeddings, we used LaBSE[5] (Feng et al., 2022) with HuggingFace Transformers (Wolf et al., 2020), which achieved the best performance in DREAM (Tiyajamorn et al., 2021). We used a [CLS] embedding as a sentence embedding. The parameters of LaBSE were frozen and only those of MLPs in our method were updated during training using the parallel corpora.

We used a batch size of 512, Adam (Kingma and Ba, 2015) optimizer with a learning rate of $1e-5$. We employed early stopping for training with a patience of 10 using a validation loss of Equation (1). The validation set was created by randomly sub-sampling 10% of the training set.

Comparison We compared our method to DREAM.[6] Besides, we compared to unsupervised QE methods that compute cosine similarities of original sentence embeddings of LaBSE, LASER[7] (Artetxe and Schwenk, 2019a,b), mSBERT[8] (Reimers and Gurevych, 2020), and BERTScore[9] (Zhang et al., 2020). Following the pre-training setup of each model, max-pooling of final layer outputs of the BiLSTM was used as a sentence embedding on LASER, and similarly, mean-pooling was used on mSBERT.

We also compared to other approaches that do not depend on multilingual sentence encoders as a reference. D-TP (Fomicheva et al., 2020b) and Prism (Thompson and Post, 2020) are unsupervised QE methods based on an encoder-decoder model. Predictor-Estimator[10] (Kim et al., 2017; Kepler et al., 2019) is a supervised method employed as the baseline for the WMT20 QE task.

3.2 Result

The first set of rows in Table 1 indicates the performance of the original sentence embeddings from LaBSE and their meaning embeddings derived by DREAM and our method. While both DREAM

[3] https://github.com/facebookresearch/mlqe

[4] We sampled the same numbers of parallel sentences as in Tiyajamorn et al. (2021) from http://www.statmt.org/wmt20/quality-estimation-task.html for fair comparison.

[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/Unbabel/OpenKiwi

	High Resource		Medium Resource		Low Resource		
Model	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	0.196	0.197	0.724‡	0.578	**0.636**	0.568	0.483
Ours	**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
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

Table 1: Pearson correlation coefficients measured on WMT20 QE task (Superscripts of ‡ and † indicate statistically significant differences of $p < 0.01$ and 0.05, respectively, compared to DREAM.)

and our method consistently outperformed LaBSE, our method achieved larger improvements. These results confirm that our method with adversarial training further enhanced the ability of meaning embedding distillation of DREAM.

The second set of rows shows the performance of previous unsupervised methods based on multilingual sentence encoders. Our method outperformed these methods on most language pairs. Particularly, it showed higher scores on low-resource language pairs, and achieved the highest correlation with human scores on average for all language pairs. It is notable that our method outperformed mSBERT on four out of six language pairs, which had sentence similarity estimation in its pre-training.

The last set of rows shows the performance of other QE models independent of multilingual sentence encoders. Our method achieved higher scores than the supervised QE model, Predictor-Estimator, for all language pairs.

D-TP and Prism achieved higher scores than ours in high-resource language pairs, but our method outperformed them in low-resource language pairs. Although D-TP assumes that users can access to the parameters of a machine translatioin model for QE, such a situation is practically limited because in general, machine translation systems are black-box to end-users (*e.g.* online machine translation services). Prism requires a large-scale bilingual corpus for training its encoder and decoder from scratch, which restricts its applicability to low-resource language pairs.

	L_R	L_C	L_L	L_A	Avg.
(a)	✓				0.393
(b)	✓	✓			0.086
(c)	✓		✓		0.075
(d)	✓			✓	0.427
(e)	✓	✓	✓		0.439
(f)	✓	✓		✓	0.297
(g)	✓		✓	✓	0.482
(h)		✓	✓	✓	0.488

Table 2: Pearson correlation coefficients in ablation

3.3 Ablation Study

Table 2 shows the results of the ablation study. The upper rows show the performance when L_R is combined with one other loss function, and the lower rows show the performance when each loss function is excluded from the proposed method, measured on WMT20 QE task.

The first set of rows (rows (a) to (d)) shows that adversarial loss L_A has the largest contribution on its own. In contrast, cross reconstruction loss L_C and language embedding loss L_L largely deteriorated the performance of L_R. However, interestingly, the second set of rows (rows (e) to (h)) show that the performance drop is largest when L_L is removed. These results indicate that L_L is crucial when combining different loss functions. We presume that L_L has an effect that meaning and language information are separated into the corresponding embeddings. In other words, it prevents that meaning information leaks to language embed-

dings. In summary, these analyses revealed two loss functions that most contribute to the performance of the proposed method: L_L and L_A, and these should be used together.

4 Summary and Future Work

We introduced adversarial training to disentangle meaning embeddings from sentence representations of multilingual sentence encoders for unsupervised QE. Our method consistently improves the performance of a state-of-the-art multilingual sentence encoder.

Our future work includes exploring ways to utilize language-specific embeddings for QE in terms of fluency of sentences. Combined with the present method of assessing the adequacy of sentences, a better QE may be achieved. We will also apply our method for disentangling styles and meanings of sentences for the style-transfer research.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP20K19861. This research was also obtained from the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), Japan.

References

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203.

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

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.

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 (Volume 1: Long Papers)*, pages 878–891.

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

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020b. Unsupervised Quality Estimation for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An Open Source Framework for Quality Estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

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.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 311–318.

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.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

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

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.

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

Brian Thompson and Matt Post. 2020. Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121.

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.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*, pages 1–43.