Language-Agnostic Representation from Multilingual Sentence Encoders for Cross-Lingual Similarity Estimation

Nattapong Tiyajamorn[1], **Tomoyuki Kajiwara**[2], **Yuki Arase**[3], **Makoto Onizuka**[3]
[1] The University of Tokyo,* nat-tiyajamorn@g.ecc.u-tokyo.ac.jp
[2] Ehime University, kajiwara@cs.ehime-u.ac.jp
[3] Osaka University, {arase, onizuka}@ist.osaka-u.ac.jp

Abstract

We propose a method to distil language-agnostic meaning embedding using a multilingual sentence encoder. By removing language-specific information from the original embedding, we retrieve an embedding that fully represents the meaning of the sentence. The proposed method relies only on parallel corpora without any human annotations. Our meaning embedding allows for efficient cross-lingual sentence similarity estimation using a simple cosine similarity calculation. Experimental results of both the quality estimation of machine translation and cross-lingual semantic textual similarity tasks reveal that our method consistently outperforms the strong baselines using the original multilingual embeddings. The method also consistently improves the performance of any pre-trained multilingual sentence encoder, even in low-resource language pairs, where only tens of thousands of parallel sentence pairs are available.[1]

1 Introduction

Pre-trained sentence encoders (Kiros et al., 2015; Logeswaran and Lee, 2018; Cer et al., 2018; Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020) boost the performance of various natural language understanding (NLU) tasks (Wang et al., 2018). Among them, the combination of self-attention networks (SANs) (Vaswani et al., 2017) and masked language modelling (Devlin et al., 2019; Liu et al., 2019) has proved to be remarkably successful. These techniques are generalised across languages (K et al., 2020) and are even applied to cross-lingual and multilingual NLU tasks such as cross-lingual semantic textual similarity (STS) (Cer et al., 2017) and quality estimation (QE) of machine translation (Specia et al., 2020).

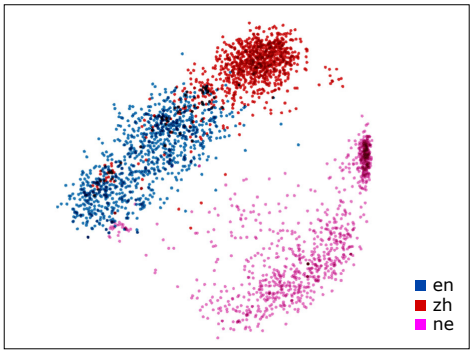

Figure 1: Visualisation of 1,000 mBERT embeddings of parallel sentences in three languages: English (en), Chinese (zh), and Nepalese (ne)

In the latest QE competitions at the conference on machine translation (WMT) (Specia et al., 2020), all top-ranked systems (Ranasinghe et al., 2020; Fomicheva et al., 2020a; Nakamachi et al., 2020) employed pre-trained multilingual sentence encoders, such as multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-RoBERTa (XLM-R) (Conneau and Lample, 2019; Conneau et al., 2020). These multilingual sentence encoders form a single self-attention network pre-trained on monolingual corpora in over 100 languages with the objective function of masked language modelling. Fine-tuning with a human-annotated corpus is mandatory for these models to enable them to estimate the semantic similarity between sentences across languages. Otherwise, these models are not sensitive to semantic similarity.

A sentence encoder that can estimate semantic similarity across languages without fine-tuning for the target task is desirable because bilingual corpora with human annotations are unavailable in most language pairs. Figure 1 plots embeddings of parallel sentences in three languages extracted from mBERT without fine-tuning. This visualisation implies that the mBERT embeddings form clusters by language rather than by meaning.

We propose a method for distilling language-

*This study was conducted when the first author was an undergraduate student at Osaka University.

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

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7764–7774
November 7–11, 2021.

agnostic meaning embeddings by removing language-specific information from sentence embeddings generated by off-the-shelf multilingual sentence encoders. Our embeddings allow efficient cross-lingual sentence similarity estimation using simple cosine similarity. Our method does not require human annotations specific to the target task and is based solely on the bilingual corpora. Experimental results on both the WMT20 QE task (Specia et al., 2020) and the SemEval-2017 cross-lingual STS task (Cer et al., 2017) in unsupervised settings revealed that our method consistently outperformed the strong baselines using the existing pre-trained multilingual sentence encoders.

2 Related Work

2.1 Multilingual Sentence Encoders

Early multilingual sentence encoders, such as LASER (Artetxe and Schwenk, 2019a,b), were encoder-decoder models based on recurrent neural networks. Similar to the evolution of monolingual sentence encoders (Kiros et al., 2015; Logeswaran and Lee, 2018; Cer et al., 2018; Reimers and Gurevych, 2019), multilingual sentence encoders have now been replaced by encoder-only models based on self-attention networks (Vaswani et al., 2017) for computational efficiency and improved performance in downstream tasks. Recent multilingual sentence encoders, such as mBERT (Devlin et al., 2019) and XLM-R (Conneau and Lample, 2019; Conneau et al., 2020), are single self-attention networks pre-trained on monolingual corpora in over 100 languages for the objective function of masked language modelling (Devlin et al., 2019; Liu et al., 2019). LaBSE (Feng et al., 2020) is a state-of-the-art multilingual sentence encoder for parallel text retrieval trained in both masked language modelling and translation language modelling (Conneau and Lample, 2019). LaBSE is trained using a maximum of 100 million sentence pairs in each language, with a total of 6 billion sentence pairs of bilingual corpora. We extended these SAN-based multilingual sentence encoders for unsupervised cross-lingual similarity estimation.

The multilingual version of Sentence-BERT (SBERT) (Reimers and Gurevych, 2020) was obtained by knowledge distillation from the English version of SBERT (Reimers and Gurevych, 2019). Although this model achieves the best performance in cross-lingual STS tasks, it is not fully unsupervised because SBERT is fine-tuned for STS tasks.

2.2 Unsupervised Methods for Cross-lingual Sentence Similarity Estimation

Libovický et al. (2020) extracts language-neutral embeddings (centered and projection) from pre-trained multilingual sentence encoders. The centered method subtracts the mean embedding for each language from the sentence embedding. The projection method involves bilingual projections using a parallel corpus and map embeddings in other languages into the space of English.

BERTScore (Zhang et al., 2020) estimates the semantic similarity between sentences by matching token embeddings from BERT (Devlin et al., 2019). Although the BERTScore of its original form is a reference-based automatic evaluation method, it can be applied to an unsupervised cross-lingual similarity estimation by using multilingual sentence encoders instead of BERT.

D-TP and D-Lex-Sim (Fomicheva et al., 2020b) are unsupervised QE methods; however, they use neural machine translation (NMT) systems that are the targets of QE. D-TP uses a sequence-level translation probability normalised by sentence length. D-Lex-Sim calculates the METEOR score (Banerjee and Lavie, 2005) based on the lexical variation between the translation hypotheses. These methods are useful for white-box machine translation systems; however, in general, users can access only the output sentences.

Prism (Thompson and Post, 2020) and BGT (Wieting et al., 2020) are state-of-the-art unsupervised methods for QE and STS, respectively. These are NMT models that train encoder-decoder structures of SANs on bilingual corpora. Prism uses the generation probability of force-decoding a target sentence as the QE score. BGT disentangles language-specific and language-agnostic embeddings from input sentences based on an auto-encoding mechanism. By calculating the cosine similarity between such language-agnostic embeddings, BGT estimates cross-lingual sentence similarity. The need for large-scale bilingual corpora to train NMT models limits the language pairs that these models can support. While multilingual sentence encoders cover over 100 languages, Prism covers only 39 languages. Although we extract both language-specific and language-agnostic embeddings, the decoder-free architecture of our model supports to support low-resource language pairs. In other words, our method is sufficiently efficient to support the massively multilingual scenario.

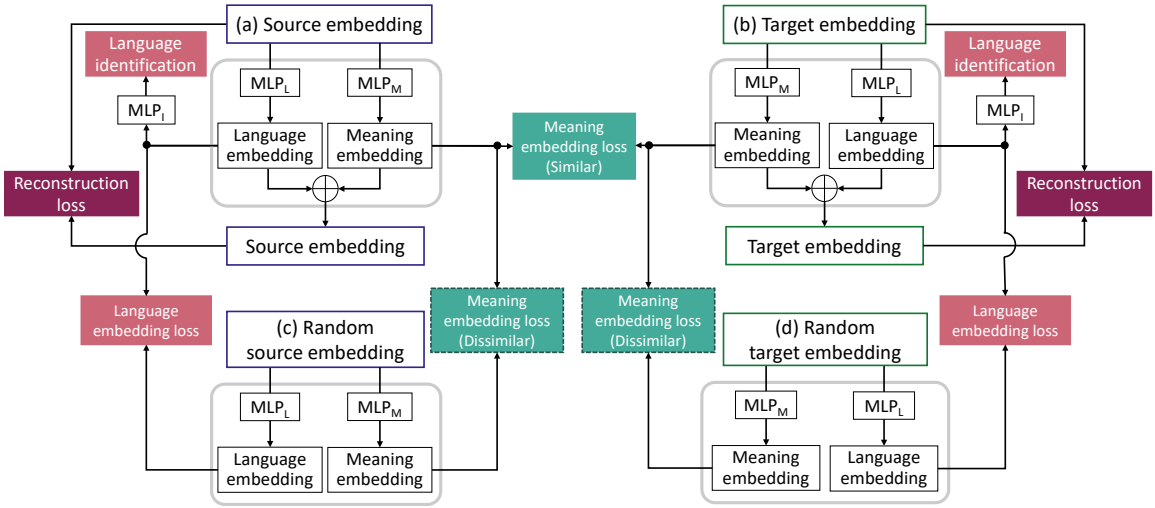

Figure 2: Multitask training for distilling meaning embeddings from multilingual sentence embeddings

3 Proposed Method

Although multilingual sentence encoders are useful for cross-lingual NLU, their embeddings are highly biased by language-specific information, which separates sentence embeddings into multiple languages, as shown in Figure 1. We distil language-agnostic meaning embedding from multilingual sentence embeddings to estimate cross-lingual sentence similarity in an unsupervised manner. By training with bilingual corpora, we unite embeddings of semantically similar sentences from pre-trained multilingual sentence encoders.

Our model is an autoencoder comprising two multi-layer perceptrons (MLPs), MLP_M and MLP_L, as shown in Figure 3. The former is responsible for extracting meaning, while the latter extracts language-specific information, and then, these outputs are summed to reconstruct the input sentence embedding.

We train these MLPs using multilingual and multitask learning using three loss functions:

$$L = L_R + L_M + L_L, \quad (1)$$

where L_R for reconstruction (Section 3.1), L_M is used for extracting the meaning (Section 3.2), and L_L for extracting language information (Section 3.3).

Figure 2 presents an overview of our multitask training, for which we input a pair of bilingual sentences, (a) and (b), as well as randomly selected sentences of each language, (c) and (d). Sentences (a) and (c) are from the same language, as are (b) and (d). The constraints in L_M make the meaning embeddings derived from (a) and (b) to come closer, while the meaning embeddings derived from (a) and (c) (also (b) and (d)) become distant. In contrast, the constraints of L_L make the language embeddings derived from (a) and (c) as well as (b) and (d) to come closer, respectively. In addition, it further acts as a constraint on how language embeddings retain language-specific information using language identification.

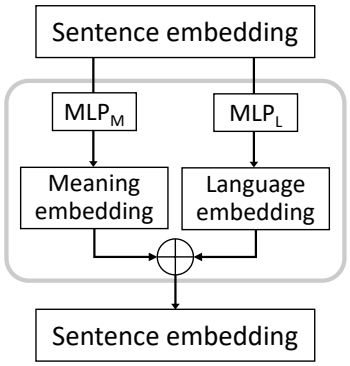

Figure 3: Design of our autoencoder

We perform multitask learning in a multilingual manner, that is, by mixing all languages to support the target task. All parameters of MLP_M are shared (same for MLP_L). Obviously, our model is trained using only parallel corpora without any human annotations, such as QE labels.

3.1 Reconstruction Loss

The reconstruction loss L_R in Equation (1) is the basis of the autoencoder training, which ensures that meaning and language embeddings, $\hat{\boldsymbol{e}}_M \in \mathbb{R}^d$ and $\hat{\boldsymbol{e}}_L \in \mathbb{R}^d$, respectively, can reconstruct the input sentence embedding $\boldsymbol{e} \in \mathbb{R}^d$ (d is the

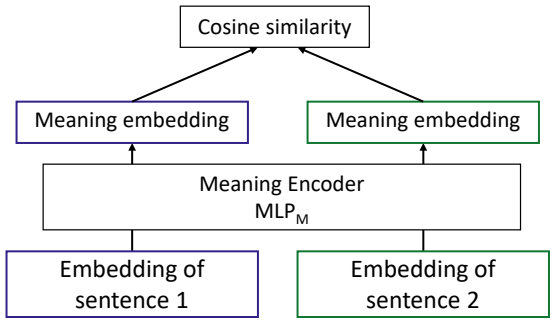

Figure 4: Evaluation Method

dimension of the sentence embedding). We define reconstruction loss as:

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

The embeddings of $\hat{\boldsymbol{e}}_M$ and $\hat{\boldsymbol{e}}_L$ are derived from $\boldsymbol{e}$ using the meaning encoder $\mathrm{MLP}_M(\cdot)$ and the language encoder $\mathrm{MLP}_L(\cdot)$ as follows:

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

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

3.2 Meaning Embedding Loss

The constraint of L_M in Equation (1) is such that $\mathrm{MLP}_M(\cdot)$ extracts language-agnostic meaning representation as $\hat{\boldsymbol{e}}_M$. To achieve this, L_M considers a pair of parallel sentences ((a) and (b) in Figure 2) and random sentences of each language ((c) and (d) in Figure 2). The meaning embeddings of the former should be closer, while those of the latter should be distant, which is achieved by losses of L_M^x and L_M^m, respectively:

$$L_M = L_M^x + L_M^m. \quad (5)$$

L_M^x takes the meaning embeddings of parallel sentences, that is, an embedding of a source sentence $\hat{\boldsymbol{s}}_M \in \mathbb{R}^d$ and an embedding of a target sentence $\hat{\boldsymbol{t}}_M \in \mathbb{R}^d$, and computes the cosine distance.

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

where $\phi(\cdot)$ computes cosine similarity.

In contrast, L_M^m takes the meaning embeddings of the same language $\hat{\boldsymbol{s}}_M$ and $\hat{\boldsymbol{s}}'_M \in \mathbb{R}^d$. Because these sentences are randomly paired, their meaning embeddings should be distant. The same constraint applies to the meaning embeddings of the other languages, $\hat{\boldsymbol{t}}_M$ and $\hat{\boldsymbol{t}}'_M \in \mathbb{R}^d$. We define L_M^m as:

$$L_M^m = \max(0, \phi(\hat{\boldsymbol{s}}_M, \hat{\boldsymbol{s}}'_M)) + \max(0, \phi(\hat{\boldsymbol{t}}_M, \hat{\boldsymbol{t}}'_M)). \quad (7)$$

Language Pair	Number of sentence pairs
en-de	1,172,003
en-zh	1,015,264
ro-en	195,082
et-en	43,997
ne-en	24,914
si-en	32,340

Table 1: The number of parallel sentence pairs in each language used to train our model for the QE task

3.3 Language Embedding Loss

The constraint of L_L in Equation (1) is such that $\mathrm{MLP}_L(\cdot)$ extracts language-specific information as $\hat{\boldsymbol{e}}_L$. To achieve this, L_L consists of two sub-loss functions: language embedding loss L_L^m and language identification loss L_L^i:

$$L_L = L_L^m + L_L^i. \quad (8)$$

The constraint of L_L^m is such that language embeddings of the same language come closer. In addition, the constraint of L_L^i is such that language embeddings become useful for language identification and for avoiding collection of random noises.

L_L^m takes language embeddings of the same language: for one language $\hat{\boldsymbol{s}}_L$ and $\hat{\boldsymbol{s}}'_L$, and another $\hat{\boldsymbol{t}}_L$ and $\hat{\boldsymbol{t}}'_L$. Then, L_L^m computes the cosine distances of each pair of language embeddings:

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

By minimising the distance between language embeddings of non-parallel sentences, the indirect constraint of L_L^m is such that meaning and language-specific information are clearly separated. Such non-parallel sentences are written in the same language, but their meanings are different. The constraint of our meaning embedding loss makes these non-parallel sentences distant, while the constraint of our language embedding loss makes language embeddings come closer. In other words, meaning and language embedding losses operate in opposite directions for non-parallel sentences. We expect that this training helps clearly separate the meaning and language embeddings. In contrast, language-specific embeddings in BGT (Wieting et al., 2020) are trained with only parallel sentences, which may allow meaning information to leak into the language-specific embeddings.

L_L^i computes the loss for language identification. We conduct language identification using an MLP:

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\mathrm{MLP}_I(\hat{\boldsymbol{e}}_L)), \quad (10)$$

	High Resource		Medium Resource		Low Resource		
Model	**en-de**	**en-zh**	**ro-en**	**et-en**	**ne-en**	**si-en**	**Avg.**
mBERT	0.071	0.010	0.182	0.009	0.025	-	0.056
mBERT (Meaning)	**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 (Meaning)	**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 (Meaning)	**0.151**	**0.156**	**0.711**	0.549	**0.627**	**0.552**	**0.458**
mBERT (centered)	0.088	0.079	0.592	0.285	0.430	-	0.295
mBERT (projection)	0.105	0.054	0.468	0.187	0.170	-	0.197
LASER	0.105	0.106	0.705	0.463	0.182	0.325	0.314
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
D-Lex-Sim	0.172	0.313	0.669	0.612	0.600	0.513	0.480
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 2: Pearson correlation coefficients evaluated on WMT20 QE task (For reference, the last set of rows shows the state-of-the-art models, which are not directly comparable to ours because the settings are different.)

where $\hat{\boldsymbol{e}}_L$ is either $\hat{\boldsymbol{s}}_L$ or $\hat{\boldsymbol{t}}_L$ and $\mathrm{softmax}(\cdot)$ is a softmax function. L_L^i computes the multiclass cross-entropy loss as:

$$L_L^i = -\sum_j \boldsymbol{y}_j \log \hat{\boldsymbol{y}}_j. \quad (11)$$

3.4 Training Details

All the MLPs in our model, MLP_M, MLP_L, and MLP_I, are a single-layer feedforward networks. We used mBERT[2] (Devlin et al., 2019), XLM-R[3] (Conneau et al., 2020), and LaBSE[4] (Feng et al., 2020), which are state-of-the-art pre-trained multilingual sentence encoders (Wolf et al., 2020). We froze the parameters of these multilingual sentence encoders and trained only the MLPs using parallel corpora. We used the output embedding of the [CLS] token for sentence embedding.

We trained our model with a batch size of 512. As an optimiser, we used Adam (Kingma and Ba, 2015) with a learning rate of $1e-4$ for all the models. We employed early stopping for training with a patience of 15 using a validation loss. The validation set was created by randomly sampling 10% from the training set.

[2] https://huggingface.co/bert-base-multilingual-cased

[3] https://huggingface.co/xlm-roberta-large

[4] https://huggingface.co/sentence-transformers/LaBSE

4 Evaluation

We evaluated the effectiveness of the proposed method in two regression tasks: the WMT20 QE task (Specia et al., 2020) and SemEval-2017 cross-lingual STS task (Cer et al., 2017). As shown in Figure 4, the meaning embeddings of each input sentence are extracted using our meaning encoder. In this experiment, we evaluated the correlation between the cosine similarity of meaning embeddings and human labels. Following the official evaluation metrics, we used Pearson correlation for both tasks implemented in the SciPy[5] package.

4.1 WMT20 Quality Estimation Task

4.1.1 Setting

In this task, we trained our model on the publicly available bilingual corpora[6] that were used to train the target machine translation systems[7] (Ott et al., 2019) for QE. The dataset contains sentence pairs for English-German (en-de), English-Chinese (en-zh), Romanian-English (ro-en), Estonian-English (et-en), Nepalese-English (ne-en), and Sinhala-English (si-en).[8] To train our model, we randomly sampled 5% of parallel sentence pairs for each lan-

[5] https://docs.scipy.org/doc/scipy/reference/index.html

[6] http://www.statmt.org/wmt20/quality-estimation-task.html

[7] https://github.com/pytorch/fairseq

[8] We excluded the Russian-English language pair in this experiment as we did not have access to the dataset.

guage pair.[9] Table 1 lists the numbers of parallel sentence pairs used in this experiment.

We compared previous unsupervised QE methods based on pre-trained multilingual sentence encoders. The method proposed by Libovický et al. (2020) obtains language-neutral embeddings from mBERT, denoted as mBERT (centered) and mBERT (projection).[10] Owing to the lack of a development set to determine which layer to use in these methods, we used the 8th layer, which was reported to perform consistently well in the original paper. LASER[11] (Artetxe and Schwenk, 2019a,b) is a multilingual encoder-decoder model. BERTScore[12] (Zhang et al., 2020) is a method for estimating sentence similarity by matching token embeddings from a sentence encoder. In this experiment, we used BERTScore with `xlm-roberta-large`, which has been reported to have the highest performance.

4.1.2 Result

Table 2 shows the Pearson correlation coefficients of the models compared. The first set of rows shows the scores of the original mBERT, XLM-R, and LaBSE, and their meaning embeddings using our method. Our method consistently improved the QE performance on all of the multilingual sentence encoders. Among them, meaning embeddings of LaBSE achieved the best performance. While the meaning embeddings of mBERT and XLM-R are inferior to those of LaBSE, improvements over the original models are noticeable.

Our method outperformed the original LaBSE, even though the bilingual corpora we used are three orders of magnitude smaller than those used to train the LaBSE. This implies that the benefits of our method are not simply due to the use of bilingual corpora, but also due to the effectiveness of the distillation method.

The second set of rows shows the performance of previous methods. The meaning embeddings of LaBSE outperformed these methods for both high- and low-resource language pairs.

The last set of rows shows the performance of other QE models that do not use pre-trained multilingual sentence encoders. These methods achieve higher performance than ours in high-resource language pairs, but not in low-resource language pairs. D-TP and D-Lex-Sim (Fomicheva et al., 2020b) are unsupervised QE methods using the NMT models that are the targets of QE. It is unlikely that we will always be allowed to use these NMT parameters in practice, while our method can conduct QE for black-box NMT systems. Prism[13] (Thompson and Post, 2020) is the current state-of-the-art unsupervised QE method, which is based on an encoder-decoder model trained on large-scale bilingual corpora. In contrast, our meaning embeddings of LaBSE efficiently support low-resource language pairs.

The last row shows the performance of the Predictor-Estimator (Kim et al., 2017), which is the *supervised* QE model.[14] Predictor-Estimator is regarded as the strong baseline for supervised QE tasks. It is notable that the meaning embeddings of LaBSE outperformed the supervised Predictor-Estimator in both medium- and low-resource language pairs.

Language Pair	Number of sentence pairs
en-ar	27,593
en-de	299,766
en-es	207,514
en-fr	262,075
en-it	482,945
en-nl	72,388
en-tr	668,260

Table 3: The number of parallel sentence pairs in each language used to train our model for the STS task

4.2 SemEval-2017 Cross-lingual STS Task

4.2.1 Setting

In this experiment, we evaluated the effects of our method on a cross-lingual STS task (Cer et al., 2017; Reimers and Gurevych, 2020).[15] The dataset provides 7 cross-lingual sentence pairs of English-Arabic (en-ar), English-German (en-de), English-Turkish (en-tr), English-Spanish (en-es), English-French (en-fr), English-Italian (en-it), English-

[9] As mBERT does not support Sinhala, mBERT-based models were trained on only for language pairs other than si-en.

[10] https://github.com/jlibovicky/assess-multilingual-bert

[11] https://github.com/facebookresearch/LASER

[12] https://github.com/Tiiiger/bert_score

[13] https://github.com/thompsonb/prism

[14] We used OpenKiwi (https://github.com/Unbabel/OpenKiwi) (Kepler et al., 2019) to implement the Predictor-Estimator model.

[15] https://public.ukp.informatik.tu-darmstadt.de/reimers/sentence-transformers/datasets/STS2017-extended.zip

Model	en-ar	en-de	en-tr	en-es	en-fr	en-it	en-nl	Avg.
mBERT	0.048	0.068	0.068	0.048	0.047	0.070	0.058	0.058
mBERT (Meaning)	**0.156**	**0.234**	**0.172**	**0.092**	**0.313**	**0.342**	**0.353**	**0.237**
XLM-R	0.089	0.095	0.153	0.113	0.029	0.080	0.073	0.090
XLM-R (Meaning)	**0.233**	**0.281**	**0.337**	**0.164**	**0.323**	**0.277**	**0.380**	**0.285**
LaBSE	0.705	0.721	0.748	**0.692**	0.759	0.760	0.755	0.734
LaBSE (Meaning)	**0.730**	**0.746**	**0.753**	0.688	**0.782**	**0.781**	**0.776**	**0.751**
mBERT (centered)	0.168	0.204	0.115	0.168	0.282	0.285	0.306	0.218
mBERT (projection)	0.159	0.096	−0.016	−0.016	0.204	0.216	0.153	0.114
LASER	0.656	0.659	0.721	0.599	0.694	0.717	0.689	0.676
BERTScore	0.451	0.452	0.441	0.376	0.479	0.479	0.531	0.458
Multilingual SBERT	0.745	0.766	0.755	0.757	0.767	0.783	0.762	0.762
BGT	0.735	-	0.749	0.756	-	-	-	0.747

Table 4: Pearson correlation coefficient evaluated on cross-lingual STS task (For reference, the last set of rows shows the state-of-the-art models, which are not directly comparable to ours because their settings are different.)

Model	ar-ar	en-en	es-es	Avg.
mBERT	0.306	0.250	0.294	0.283
mBERT (Meaning)	**0.391**	**0.408**	**0.403**	**0.401**
XLM-R	0.084	0.141	0.149	0.125
XLM-R (Meaning)	**0.321**	**0.425**	**0.344**	**0.363**
LaBSE	0.705	0.759	**0.823**	0.762
LaBSE (Meaning)	**0.709**	**0.791**	0.817	**0.772**
mBERT (centered)	0.451	0.416	0.365	0.410
mBERT (projection)	−0.080	0.123	0.105	0.049
LASER	0.693	0.773	0.797	0.754
BERTScore	0.579	0.605	0.589	0.591
Multilingual SBERT	0.757	0.806	0.809	0.791
BGT	0.749	-	0.857	0.803

Table 5: Pearson correlation coefficient evaluated on monolingual STS task (For reference, the last set of rows shows the state-of-the-art models, which are not directly comparable to ours because their settings are different.)

Dutch (en-nl), and 3 monolingual sentence pairs of Arabic (ar-ar), English (en-en), and Spanish (es-es). Following Reimers and Gurevych (2020), we trained our model using parallel corpora from Tatoeba.[16] Table 3 lists the numbers of parallel sentence pairs used to train our models. Again, we compared mBERT (centered), mBERT (projection), LASER, and BERTScore as baselines.

4.2.2 Result

The first and second sets of rows in Tables 4 and 5 show the Pearson correlation coefficients of the original multilingual sentence encoders, the meaning embeddings by our model, and baselines, measured on the cross-lingual STS task, respectively. Similar to the evaluations of the QE task, our method consistently improved the performance of mBERT, XLM-R, and LaBSE for languages other than Spanish on LaBSE. Our method substantially improved STS performance not only of language pairs for which large-scale training data are available in Table 3, but also of language pairs with fewer data, such as Arabic (en-ar) and Dutch (en-nl). Table 5 implies that our method improves performance not only for cross-lingual but also for monolingual tasks.

The last sets of rows in Tables 4 and 5 show the performance of state-of-the-art models: the multilingual version of SBERT (Reimers and Gurevych, 2020). It uses knowledge distillation by setting SBERT trained with AllNLI (SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018)), and STSB (Cer et al., 2017) as a teacher and training

[16] https://tatoeba.org

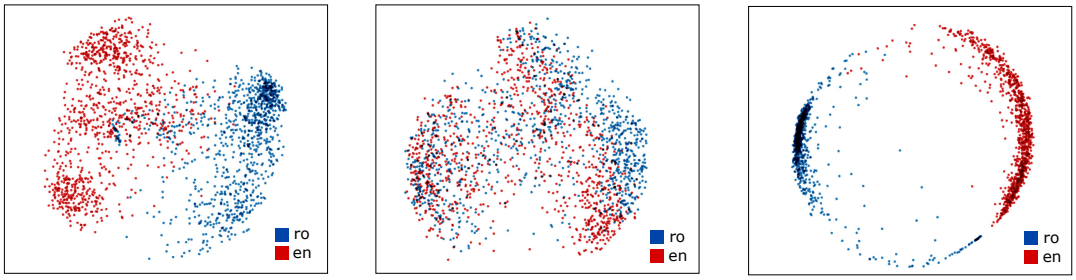

Figure 5: Visualisation of embeddings from 1,000 sentence pairs in ro-en parallel corpus.

	mBERT	XLM-R	LaBSE
Our Method	0.335	0.289	0.458
w/o Language Loss	0.322	0.286	0.449
w/o Meaning Loss	0.008	0.026	0.274

Table 6: Pearson correlation coefficients of WMT20 QE task on the ablation study

XLM-R as a student.[17] As the teacher model is exposed to the training of the STS task, it is expected that this model will achieve higher performance. Nonetheless, our meaning embeddings of LaBSE showed competitive performance without any supervision of STS. The last row shows the performance of BGT (Wieting et al., 2020) that disentangles language-agnostic and language-specific representations using an encoder-decoder model. In contrast to this model, which trains a decoder using a large-scale bilingual corpus, our method achieves higher performance in low-resource language pairs.

5 Analysis

We further analyse our method through an ablation study and visualisation of sentence embeddings.

5.1 Ablation Study

Table 6 shows the performance in the QE task when each meaning loss (Section 3.2) and language loss (Section 3.3) is removed from our method. We observe that the model's performance tends to worsen without either constraint. In particular, removing meaning loss has a serious impact on QE performance. We conjecture that this is because meaning loss allows learning semantic equivalence and inequivalence, which is useful for conducting QE.

5.2 Visualisation

Figure 5 shows the sentence embeddings from mBERT for randomly sampled 1,000 parallel sentences in English and Romanian, where dimensions were reduced by principal component analysis (Maćkiewicz and Ratajczak, 1993). Despite these parallel sentence pairs representing the same meaning, their embeddings from the original mBERT (left) form clusters by language rather than by meaning, as shown in Figure 1. By applying our method, the meaning embeddings (center) became language-agnostic. Besides, the language embeddings (right) are more clearly divided. Similar analyses in other languages are shown in Figure 6. The same tendency can be observed regardless of the language pair.

6 Conclusion

To achieve unsupervised language-agnostic sentence similarity estimation, we distilled the meaning embeddings using pre-trained multilingual sentence encoders. We trained the autoencoder consisting of two MLPs, that is, meaning encoder and language encoder, in a multitask and multilingual manner. Our method successfully distils language-agnostic (*i.e.*, meaning embedding) information by removing language-specific (*i.e.*, language embedding) information from the original sentence embedding.

Our method has following advantages: (1) It can be trained using only parallel corpora without any human annotations. (2) Based on pre-trained multilingual sentence encoders, our single model can cover more than 100 languages.

[17] Because Reimers and Gurevych (2020) measured the performance using Spearman's rank correlation coefficient, we re-evaluated the performance of `xlm-r-bert-base-nli-stsb-mean-tokens` model (available at Hugging Face) using Pearson correlation coefficient.

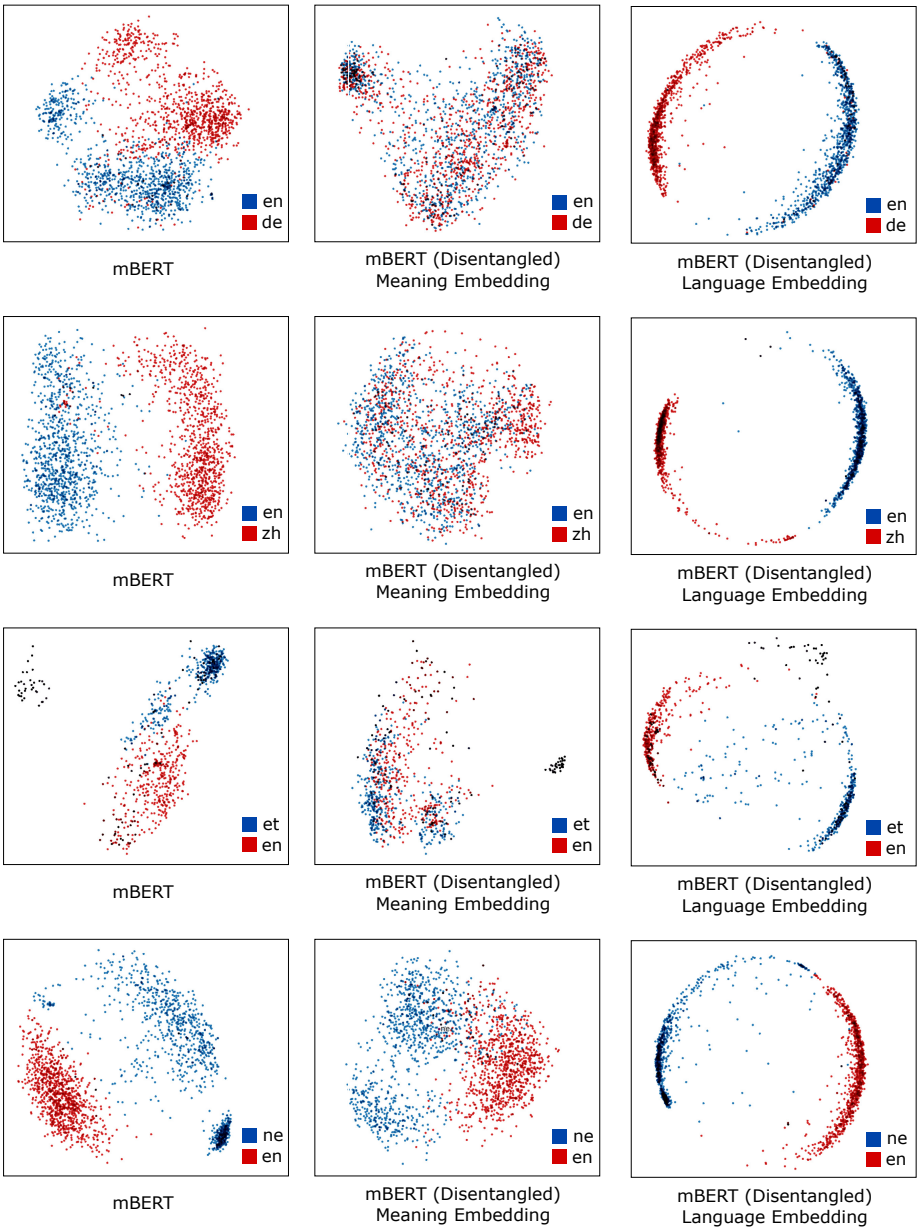

Figure 6: Visualisation of embeddings from $1,000$ sentence pairs in en-de, en-zh, et-en, and ne-en parallel corpora.

Experimental results in both the QE and cross-lingual STS tasks revealed that our method consistently improves the performance of original multilingual sentence encoders, such as mBERT, XLM-R, and LaBSE. Substantial improvements were obtained even from tens of thousands of parallel sentence pairs, achieving the highest performance in QE for low-resource language pairs.

Acknowledgments

This project was supported by Microsoft Research Asia and JSPS KAKENHI Grant Number JP20K19861.

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.

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 Ex-*

trinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 1–14.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, pages 7059–7069.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT Sentence Embedding. *arXiv:2007.01852*, pages 1–13.

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 5th 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.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *Proceedings of the 8th International Conference on Learning Representations*, pages 1–12.

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

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Proceedings of the 29th Conference on Neural Information Processing Systems*, pages 3294–3302.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proceedings of the 8th International Conference on Learning Representations*, pages 1–17.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the Language Neutrality of Pre-trained Multilingual Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*, pages 1–13.

Lajanugen Logeswaran and Honglak Lee. 2018. An Efficient Framework for Learning Sentence Representations. In *Proceedings of the 6th International Conference on Learning Representations*, pages 1–16.

Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal Components Analysis (PCA). *Computers & Geosciences*, 19(3):303 – 342.

Akifumi Nakamachi, Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. TMUOU Submission for WMT20 Quality Estimation Shared

Task. In *Proceedings of the 5th 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.

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. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992.

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*, pages 4512–4525.

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 5th Conference on Machine Translation*, pages 743–764.

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*, pages 90–121.

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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

John Wieting, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A Bilingual Generative Transformer for Semantic Sentence Embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1581–1594.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1112–1122.

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.