Efficiently Learning Multilingual Sentence Representation for Cross-lingual Sentence Classification

Zhuoyuan Mao Chenhui Chu Sadao Kurohashi
Kyoto University
{zhuoyuanmao, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Massively multilingual sentence representation models benefit cross-lingual sentence classification tasks. However, multiple training procedures, the use of a large amount of data, or inefficient model architectures result in heavy computation to train a new model for preferred languages and domains. To address this, we introduce an approach to efficiently learn multilingual sentence representation, using cross-lingual sentence reconstruction and sentence-level contrastive learning as training objectives. Empirical results show that our model yields significantly better or comparable results on two cross-lingual classification benchmarks. We release our model, which supports 62 languages: https://github.com/Mao-KU/EMS.

1 Introduction

Cross-lingual sentence representation (CSR) models [1, 2, 3, 4, 5, 6] prove to be essential for cross-lingual transfer on sentence classification tasks without the need for initial training and monolingual model. Thus, CSR models benefit low-resource languages without sufficient training data. However, existing multilingual CSR models, LASER [1], SBERT-distill [4], and LaBSE [3], require a considerable amount of data, complicated model architectures, or monolingual/multilingual pre-trained language models, for which the efficient model training has not been explored.

In this study, we present a computationally lite and effective architecture for training CSR without relying on any large-scale pre-trained language model, which makes it computational lite to train a CSR model according to our preferred domains or language groups and may have a promising future for deploying pre-trained CSR models on memory-limited devices. In particular, we propose cross-lingual token-level reconstruction (XTR) and sentence-level contrastive learning as training objectives. XTR captures the target token distribution information, whereas the contrastive objective serves to recognize translation pairs. We claim that these two objectives lead to effective language-agnostic sentence representation for the encoder-only model without language model pre-training, and the encoder-only model results in highly efficient model training. Compared with previous work, our CSR model can be trained using significantly fewer training data and less GPU consumption.

Despite the small amount of training data and low-cost training, experimental results demonstrate that our model learned a robustly aligned multilingual sentence representation space. We evaluate the language-agnostic representation based on two classification tasks in a zero-shot manner, document genre classification based on ML-Doc [7], and sentiment classification based on Amazon Review dataset version-1 [8]. Empirical results show that our model outperforms LASER and SBERT-distill on ML-Doc and Amazon Review dataset and yields comparable performance with LaBSE on Amazon Review dataset.

2 Proposed Methods

We conduct massive multilingual CSR learning by employing the dual transformer encoder as the backbone of the training framework and jointly optimizes it with generative and contrastive objectives.

2.1 Architecture

We introduce the dual transformer sharing parameters to encode parallel sentences along with several linear layers to extract cross-lingual information and compute the generative and contrastive losses (Fig. 1).

Specifically, as shown in Fig. 1, assume that we have

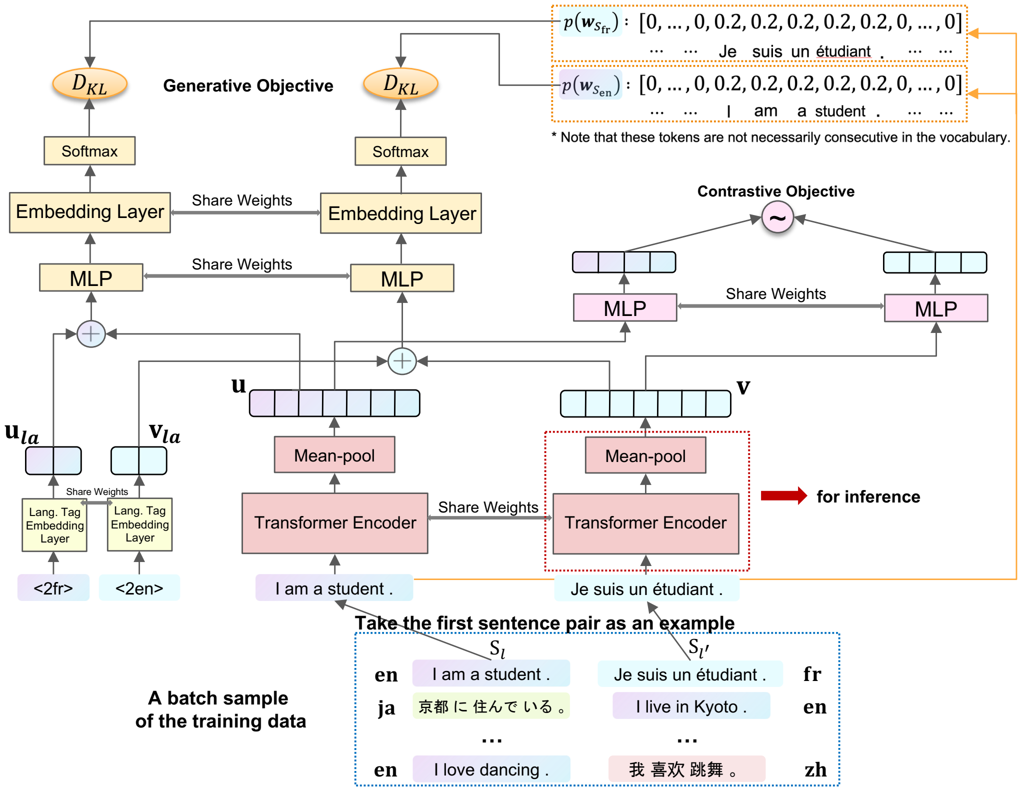

Figure 1 **Training architecture of our model.** **u** and **v** are language-agnostic sentence representations for inference, and the model components in the red dashed rectangle are used for inference. $\mathbf{u}_{la}$ and $\mathbf{u}_{la}$ are the target language token representations. ⊕ denotes the hidden vector concatenation. A batch sample of the training data is given in the blue dashed box. Orange arrows and dashed box denotes the gold token distributions within the generative objective.

a parallel corpus **C** that includes multiple languages $\{l_1, l_2, ..., l_N\}$, and each sentence pair $S = (S_l, S_{l'})$ contains a sentence in language l and its translation in language l', where $l, l' \in \{l_1, l_2, ..., l_N\}$, as shown in the blue dashed box in Fig. 1. We use the dual transformer encoder E sharing parameters to encode each sentence pair. Assume that the transformer encoder outputs of S_l are $(\mathbf{h}_1^{\mathrm{T}}, \mathbf{h}_2^{\mathrm{T}}, ..., \mathbf{h}_{\|S_l\|}^{\mathrm{T}})$, where $\|S_l\|$ indicates the length of S_l. We use the mean-pooled hidden states as the language-agnostic sentence representation **u**:

$$\mathbf{u} = \frac{1}{\parallel S_l \parallel} \sum_i \mathbf{h}_i \tag{1}$$

Similarly, we can obtain **v** for $S_{l'}$.

Using **u** and **v**, two groups of MLPs are employed to construct two training objectives. After completing the model training, given a sentence in any language, we use the transformer encoder to infer the language-agnostic sentence representation. We can implement cross-lingual downstream tasks in a zero-shot manner using **u** or **v**, as they are representations independent of the specific language.

2.2 Generative Objective

Generative objective plays an essential role for CSR learning. Inspired by LASER, we include the generative objective for the one-run model training. However, the presence of the transformer decoder in LASER increases the computational overhead. Instead, we propose XTR to improve the training efficiency while retaining the quality of sentence representation, which circumvents using a decoder.

We compute a target language representation for each sentence by employing a language embedding layer L_{la} to encode the target language token (e.g., $< 2en >$ if the target language is English) to notify the model what the target language is. For each sentence pair $S = (S_l, S_{l'})$,

$$\mathbf{u}_{la} = \mathbf{W}_{la}\mathbf{h}_{l'} \tag{2}$$

$$\mathbf{v}_{la} = \mathbf{W}_{la}\mathbf{h}_{l} \tag{3}$$

where $\mathbf{W}_{la} \in \mathbb{R}^{d_{la} \times d_{vcb}}$ denotes the parameters of L_{la}. $\mathbf{h}_l$ and $\mathbf{h}_{l'}$ respectively denote the one-hot embedding of $< 2l >$ and $< 2l' >$. d_{la} and d_{vcb} denote the dimension of the language embedding and the size of the vocabulary.

Subsequently, we concatenate the language representation with the sentence representation and use a fully connected layer L_{fc} to transform the concatenated representation for extracting the cross-lingual information. Finally, we use another linear embedding layer L_{emb} followed by Softmax to transform the representation to present two probability distributions, which are formulated as:

$$q_{S_l} = \mathrm{softmax}(\mathbf{W}_{emb}\sigma_{xtr}(\mathbf{W}_{fc}(\mathbf{u}_{la} \oplus \mathbf{u}))) \tag{4}$$

$$q_{S_{l'}} = \mathrm{softmax}(\mathbf{W}_{emb}\sigma_{xtr}(\mathbf{W}_{fc}(\mathbf{v}_{la} \oplus \mathbf{v}))) \tag{5}$$

where $\mathbf{W}_{emb} \in \mathbb{R}^{d_{vcb} \times (d_{la}+d)}$, $\mathbf{W}_{fc} \in \mathbb{R}^{(d_{la}+d) \times (d_{la}+d)}$, and d indicates the dimension of $\mathbf{u}$ (or $\mathbf{v}$). σ_{xtr} is the activation function in L_{fc}. $\oplus$ indicates concatenation.

Assume that $\mathbf{B}_i$ is a batch sampled from the training corpus $\mathbf{C}$. Then, the training loss of the XTR objective for the $\mathbf{B}_i$ is formulated as follows:

$$\mathcal{L}^{(i)}_{XTR} = \sum_{S \in \mathbf{B}_i} \Big(\mathcal{D}_{KL}\left(p_{S_{l'}}(\mathbb{W}) \parallel q_{S_l}\right) + \mathcal{D}_{KL}\left(p_{S_l}(\mathbb{W}) \parallel q_{S_{l'}}\right)\Big) \tag{6}$$

where $\mathcal{D}_{KL}$ denotes KL-divergence and $\mathbb{W}$ indicates the vocabulary set. As illustrated in the orange dashed box in Fig. 1, we use discrete uniform distribution for the tokens in S_l to define p_{S_l}. For each $w \in \mathbb{W}$, $p_{S_l}(w)$ is defined as:

$$p_{S_l}(w) = \begin{cases} \dfrac{N_w}{\|S_l\|}, & w \in S_l \\ 0, & w \notin S_l \end{cases} \tag{7}$$

where N_w indicates the number of words w in sentence S_l, and N_w is 1 in most cases. $\|S_l\|$ indicates the length of S_l. Similarly, we can obtain the definition of $p_{S_{l'}}(\mathbb{W})$.

Herein, we use the KL-divergence to measure the similarity between the token distribution of the sentence in the target language and the model output of the sentence in the source language, which helps align the language-agnostic representation space.

2.3 Contrastive Objective

We employ a sentence-level contrastive objective as an assisting objective to force the model to grasp similar information of sentences across languages. We demonstrate that the sentence-level contrastive objective is a beneficial model component to jointly assist the generative objective.

Specifically, we employ in-batch sentence-level contrastive learning by discriminating between positive and negative samples for each sentence. Given a sentence, its translation (paired sentence in another language) is deemed as a positive sample, whereas other sentences within the batch are used as the negative samples. Assume that $\mathbf{B}_i$ is a batch sampled from the training corpus $\mathbf{C}$, and the j-th sentence pair of $\mathbf{B}_i$ is $S^{(ij)} = (S^{(ij)}_l, S^{(ij)}_{l'})$. Then the sentence-level contrastive objective for $\mathbf{B}_i$ is formulated as:

$$\mathcal{L}^{(i)}_{cntrs} = -\sum_{S^{(ij)} \in \mathbf{B}_i} \Bigg(\log \frac{\exp(\mathrm{sim}(S^{(ij)}_l, S^{(ij)}_{l'})/T)}{\sum_{S^{(ik)} \in \mathbf{B}_i} \exp(\mathrm{sim}(S^{(ij)}_l, S^{(ik)}_{l'})/T)} + \log \frac{\exp(\mathrm{sim}(S^{(ij)}_l, S^{(ij)}_{l'})/T)}{\sum_{S^{(ik)} \in \mathbf{B}_i} \exp(\mathrm{sim}(S^{(ik)}_l, S^{(ij)}_{l'})/T)}\Bigg) \tag{8}$$

where T denotes a temperature hyperparameter to scale the cosine similarity. $\mathrm{sim}(S_l, S_{l'})$ is defined as:

$$\mathrm{sim}(S_l, S_{l'}) = \cos(\mathbf{h}(S_l), \mathbf{h}(S_{l'})) \tag{9}$$

$$\mathbf{h}(S_l) = \mathbf{W}_1\sigma_{cntrs}(\mathbf{W}_2\mathbf{u}) \tag{10}$$

$$\mathbf{h}(S_{l'}) = \mathbf{W}_1\sigma_{cntrs}(\mathbf{W}_2\mathbf{v}) \tag{11}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_{cntrs} \times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{d \times d}$ mean the weights of two fully-connected layers. We use ReLU [9] for σ_{cntrs}.

2.4 Joint Training

We train the model by jointly optimizing the losses of the proposed generative and contrastive objectives. Specifically, we simultaneously train each batch with:

$$\mathcal{L}^{(i)} = \frac{1}{\parallel \mathbf{B}_i \parallel}(\mathcal{L}^{(i)}_{XTR} + \mathcal{L}^{(i)}_{cntrs}) \tag{12}$$

where $\parallel \mathbf{B}_i \parallel$ denotes the number of sentence pairs within batch $\mathbf{B}_i$, namely, the batch size.

3 Evaluation

We used 143M parallel sentences to train a multilingual sentence representation model supporting 62 languages with the above training objectives. Refer to Appendix A for training details.

In this section, we evaluate the performance of our model for two cross-lingual sentence classification tasks in a zero-shot manner. Two evaluation tasks include the MLDoc benchmark [7] and cross-lingual sentiment classification on the first version of the multilingual Amazon Review corpora [8]. We compare with LASER [1], SBERT-distill [4], and LaBSE [3].[1)]

[1)] We *italicize* this model in the results as the upper bound performance on downstream tasks because a large number of parallel sentences, 6B, are used for training.

Table 1 MLDoc benchmark results (zero-shot scenario). We report the mean value of 5 runs.

Model	en-de		en-es		en-fr		en-it		en-ja		en-ru		en-zh		Avg.
	→	←	→	←	→	←	→	←	→	←	→	←	→	←	
LASER	86.3	76.7	76.2	68.1	82.1	75.7	70.3	69.8	**71.5**	59.8	64.6	68.9	**77.7**	67.3	72.5
SBERT-distill	78.5	78.7	72.7	73.3	79.7	79.6	64.4	73.0	65.7	72.0	64.2	72.7	60.3	**70.2**	71.8
ours	**87.6**	**81.1**	**82.0**	**75.5**	**82.9**	**80.6**	**70.4**	**73.6**	67.0	**72.3**	**68.5**	**77.5**	68.6	69.1	**75.5**
LaBSE	87.2	82.8	78.8	78.2	87.3	83.6	74.1	74.8	73.4	78.8	74.6	79.0	85.3	80.0	79.9

Table 2 Results of the cross-lingual sentiment classification of Amazon Review. We report the mean value of 5 runs.

Model	en-de						en-fr						en-ja						Avg.
	books		dvd		music		books		dvd		music		books		dvd		music		
	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←	
LASER	78.3	76.0	73.7	73.4	76.1	77.2	77.2	77.4	76.8	75.4	**75.8**	76.6	72.0	72.9	73.0	70.9	75.5	75.5	75.2
SBERT-distill	78.2	81.2	73.9	**77.1**	74.1	80.1	78.9	80.6	77.8	79.4	70.6	78.8	74.5	**81.9**	**76.5**	78.2	78.2	78.6	77.7
ours	**82.3**	**84.9**	**77.0**	76.7	**78.8**	**81.9**	**80.4**	**84.6**	**78.0**	**81.1**	74.7	**83.0**	**75.6**	79.5	75.4	**79.2**	**79.2**	**80.9**	**79.6**
LaBSE	82.2	79.9	77.1	77.2	79.0	80.0	83.2	82.3	81.0	80.1	77.9	80.3	78.0	80.7	77.7	77.1	81.6	79.0	79.7

3.1 MLDoc: Multilingual Document Classification

We evaluate the model performance based on the MLDoc classification task. MLDoc[2] is a benchmark to evaluate cross-lingual sentence representations, which contain datasets for eight languages [10]. Following [1], we conduct the evaluation in the zero-shot manner using 1,000 sentences in language l_1 for training, 1,000 sentences in language l_1 for validation, and 4,000 sentences in language l_2 for test. Specifically, we train a multilayer perceptron classifier based on source language representations and test the classifier for the target language. We list the average results of 5 runs for 7 language pairs and 14 directions in Table 1. We significantly observe higher accuracies of our model in most directions than those of LASER and SBERT-distill. These results demonstrate the effectiveness of the proposed training method.

3.2 CLS: Cross-lingual Sentiment Classification

Moreover, we gauge the quality of language-agnostic sentence representation based on the sentiment classification task. We use Amazon Review version-1 dataset for evaluation. The dataset [8] includes the data for English–German, English—French, and English–Japanese on "books," "dvd," and "music" domains for each language pair. For each language pair and domain, we use 2,000 sentences in language l_1 for training, 2,000 sentences in language l_1 for validation, and 2,000 sentences in language l_2 for test. Same as on MLDoc, we train a multi-layer perceptron using the language-agnostic sentence representations in language l_1 and test the classifier for another language. As listed in Table 2, our model significantly outperforms LASER and SBERT-distill, and performs comparably to LaBSE, which proves the effectiveness of our model.

3.3 Training Efficiency

We used 143M parallel data for training, which is much less than competing models. The loss nearly converged after being trained for 0.5 epochs and converged completely after 3 epochs, whereas LASER is trained for 17 epochs till convergence. Concerning the training time, SBERT-distill and LaBSE rely on large-scale pre-trained models; thus, both their pre-training and fine-tuning require heavy computation. LASER is trained with 80 V100 GPU×days, while our model requires 5 V100 GPU×days to nearly converge and 20 V100 GPU×days to converge fully, which indicates 4~16 times speedup compared with LASER.

4 Conclusion

This study presents an efficient method to train language-agnostic sentence representation for cross-lingual sentence classification. To improve training efficiency while retaining the quality of sentence representations, we propose a novel framework to jointly train "XTR" generative and sentence-level contrastive objectives. The empirical results based on two cross-lingual sentence classification tasks demonstrate the effectiveness of our model. We plan to further shrink the model architecture based on knowledge distillation for faster inference experience in the future.

[2] https://github.com/facebookresearch/MLDoc

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22J13719 and Samsung, SDS.

References

[1] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. **Trans. Assoc. Comput. Linguistics**, Vol. 7, pp. 597–610, 2019.

[2] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Multilingual universal sentence encoder for semantic retrieval. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020**, pp. 87–94. Association for Computational Linguistics, 2020.

[3] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022**, pp. 878–891. Association for Computational Linguistics, 2022.

[4] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020**, pp. 4512–4525. Association for Computational Linguistics, 2020.

[5] Zhuoyuan Mao, Prakhar Gupta, Chenhui Chu, Martin Jaggi, and Sadao Kurohashi. Lightweight cross-lingual sentence representation learning. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021**, pp. 2902–2913. Association for Computational Linguistics, 2021.

[6] Zhuoyuan Mao, Prakhar Gupta, Chenhui Chu, Martin Jaggi, and Sadao Kurohashi. Learning cross-lingual sentence representations for multilingual document classification with token-level reconstruction. 言語処理学会 第27回年次大会, pp. 1049–1053, 2021.

[7] Holger Schwenk and Xian Li. A corpus for multilingual document classification in eight languages. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018**. European Language Resources Association (ELRA), 2018.

[8] Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In **ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden**, pp. 1118–1127. The Association for Computer Linguistics, 2010.

[9] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In **Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel**, pp. 807–814. Omnipress, 2010.

[10] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. RCV1: A new benchmark collection for text categorization research. **J. Mach. Learn. Res.**, Vol. 5, pp. 361–397, 2004.

[11] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021**, pp. 1351–1361. Association for Computational Linguistics, 2021.

[12] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In **Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event**, Vol. 119 of **Proceedings of Machine Learning Research**, pp. 4411–4421. PMLR, 2020.

[13] Zeljko Agic and Ivan Vulic. JW300: A wide-coverage parallel corpus for low-resource languages. In **Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers**, pp. 3204–3210. Association for Computational Linguistics, 2019.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA**, pp. 5998–6008, 2017.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**, 2015.

A Training Details

A.1 Training Data

We collected parallel corpora for 62 languages from OPUS.[3] The 62 languages that we selected cover all the languages in [11] and the languages suggested by the cross-lingual generalization benchmark, XTREME [12]. Parallel corpora used for training include Europarl, GlobalVoices,[4] NewsCommentary,[5] OpenSubtitles,[6] Ted,[7] UNPC,[8] WikiMatrix,[9] Tatoeba.[10]

The aforementioned training data leads to a 143M parallel corpus, which is much less than LASER. Moreover, we excluded the JW300 [13] corpus and pruned OpenSubtitles and UNPC corpora and included less training data than SBERT-distill.[11] In addition, [3] used 6B parallel data to fine-tune the pre-trained mBERT, which leads to enormous computational resource consumption and is impractical to reproduce. However, the proposed model used a limited number of parallel sentences while retaining the sentence representation performance.

A.2 Training Details

We employed Transformer [14] encoder as the basic unit of the training architecture (Fig. 1). We conducted a grid search for optimal hyperparameter combinations by observing the validation loss on the WikiMatrix validation datasets (Table 3).

As a result, the dual transformer encoder sharing parameters has 6 layers, 16 attention heads, a hidden size of 1,024, and a feed-forward size of 4,096. The transformer encoder can be substituted by encoders with other structures. d, d_{vcb}, d_{la}, and d_{cntrs} are 1,024, 60,000, 128, and 128, respectively. We set 0.1 for the temperature T of the contrastive objective.

For the model training, we fed the parallel sentences into the dual transformer encoder and truncated the sentences up to 120 tokens.[12] We trained three epochs for the entire training corpus with the Adam optimizer [15], the learning rate of 0.0003 with the linear warm-up strategy of 10,000 steps, a weight decay of 0.00001, and a dropout[13] of 0.1 for the transformer encoder. We used four V100 GPUs to conduct the model training with a batch size of 152 parallel sentences.

Table 3 **Values of the hyperparameters tuned by grid search.** **Bold** denotes the best hyperparameter combination.

Hyperparameters	Values
number of the transformer layers	2, 4, **6**, 12
transformer hidden dropout	0.0, **0.1**, 0.3
transformer attention dropout	0.0, **0.1**
T	0.01, **0.1**, 0.2, 0.5, 1.0
learning rate	1e-4, **3e-4**, 5e-4, 1e-3
weight decay	0.0, **1e-5**, 1e-4, 1e-3
warm-up steps	0, 5,000, **10,000**, 20,000

A.3 Model Release

We release our model at https://github.com/Mao-KU/EMS. Our model supports the encoding of the following 62 languages: af, ar, bg, bn, ca, cs, da, de, el, en, eo, es, et, eu, fa, fi, fr, gl, gu, he, hi, hr, hu, hy, id, it, ja, jv, ka, kk, ko, ku, lt, lv, mk, ml, mn, mr, ms, my, nb, nl, pl, pt, ro, ru, sk, sl, sq, sr, sv, sw, ta, te, th, tl, tr, uk, ur, vi, yo, zh.[14]

[3] https://opus.nlpl.eu/

[4] http://casmacat.eu/corpus/global-voices.html

[5] https://statmt.org/

[6] opensubtitles.org

[7] https://opus.nlpl.eu/TED2020.php

[8] https://opus.nlpl.eu/UNPC.php

[9] https://opus.nlpl.eu/WikiMatrix.php

[10] https://tatoeba.org/

[11] [11] used JW300 and all of the entire corpora we used.

[12] Although LASER and SBERT-distill allowed much longer sentences during the training phase, we demonstrate that 120 tokens are sufficient for a single sentence with the complete semantics. For the evaluation, documents longer than 120 tokens can be separated into several sentences, which would not limit the usage of our model.

[13] the hidden and attention dropouts

[14] Refer to https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes for language codes.