Lightweight Cross-Lingual Sentence Representation Learning

Zhuoyuan Mao[♠] **Prakhar Gupta**[♣]
Chenhui Chu[♠] **Martin Jaggi**[♣] **Sadao Kurohashi**[♠]
[♠]Kyoto University, Japan [♣]EPFL, Switzerland
{zhuoyuanmao, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp
{prakhar.gupta, martin.jaggi}@epfl.ch

Abstract

Large-scale models for learning fixed-dimensional cross-lingual sentence representations like LASER (Artetxe and Schwenk, 2019b) lead to significant improvement in performance on downstream tasks. However, further increases and modifications based on such large-scale models are usually impractical due to memory limitations. In this work, we introduce a lightweight dual-transformer architecture with just 2 layers for generating memory-efficient cross-lingual sentence representations. We explore different training tasks and observe that current cross-lingual training tasks leave a lot to be desired for this shallow architecture. To ameliorate this, we propose a novel cross-lingual language model, which combines the existing single-word masked language model with the newly proposed cross-lingual token-level reconstruction task. We further augment the training task by the introduction of two computationally-lite sentence-level contrastive learning tasks to enhance the alignment of cross-lingual sentence representation space, which compensates for the learning bottleneck of the lightweight transformer for generative tasks. Our comparisons with competing models on cross-lingual sentence retrieval and multilingual document classification confirm the effectiveness of the newly proposed training tasks for a shallow model. [1]

1 Introduction

Cross-lingual sentence representation models (Schwenk and Douze, 2017; España-Bonet et al., 2017; Yu et al., 2018; Devlin et al., 2019; Chidambaram et al., 2019; Artetxe and Schwenk, 2019b; Kim et al., 2019; Sabet et al., 2019; Conneau and Lample, 2019; Feng et al., 2020; Li and Mak, 2020) learn language-agnostic representations facilitating tasks like cross-lingual sentence retrieval (XSR) and cross-lingual knowledge transfer on downstream tasks without the need for training a new monolingual representation model from scratch. Thus, such models benefit from an increased amount of data during training and lead to improved performances for low-resource languages.

The above-mentioned models can be categorized into two classes. On one hand, *global fine-tuning* methods like mBERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019) require being fine-tuned globally which results in a significant overhead of its own. On the other hand, *fixed-dimensional* methods like LASER (Artetxe and Schwenk, 2019b) fix the sentence representations during the pre-training phase, and subsequently the fine-tuning for specific downstream tasks without back-propagating to the pre-trained model will be extremely computationally-lite. Lightweight models have been sufficiently explored for the former group by either shrinking the model (Lan et al., 2020) or training a student model (Sanh et al., 2019; Jiao et al., 2020; Reimers and Gurevych, 2020; Sun et al., 2020). However, the lightweight models for the latter group have not been explored before, which may have a more promising future for deploying task-specific fine-tuning onto edge devices.

In this work, we propose a variety of training tasks for a lightweight cross-lingual sentence model while retaining the robustness. To improve the computational efficiency, we utilize a lightweight dual-transformer architecture with just 2 layers, significantly decreasing the memory consumption and accelerating the training to further improve the efficiency. Our model uses significantly less number of parameters compared to both global fine-tuning methods like mBERT, and fixed-dimensional representation methods like LASER,

[1] https://github.com/Mao-KU/lightweight-crosslingual-sent2vec

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2902–2913
August 1–6, 2021.

Method	Architecture	d_h	d_{fc}	$attn_h$	$Enc.$	$Dec.$	$Params.$
mBERT (Devlin et al., 2019)	Transformer	768	3,072	12	12	N/A	110M
LASER (Artetxe and Schwenk, 2019b)	Bi-LSTM	512×2	N/A	N/A	5	5	154M
T-LASER (Li and Mak, 2020)	Transformer	1,024	4,096	16	6	1	246M
Ours	Transformer	**512**	**1,024**	**8**	**2**	N/A	**30M**

Table 1: **Model sizes of related work and ours.** Our work mainly focuses on the comparison with previous fixed-dimensional methods like LASER, T-LASER, etc. d_h, d_{fc}, $attn_h$, $Enc.$, $Dec.$, $Params.$ denote dimension of the hidden state, dimension of the feed-forward hidden state, number of the attention heads, number of the encoder layers, number of the decoder layers, and number of the parameters respectively.

and T-LASER (Li and Mak, 2020) (see Table 1).

Given a fixed training-set and model architecture, the robustness of the sentence representation is dependent on the training task. It is much more difficult for a lightweight model to learn robust representations merely with existing generative tasks (see Section 2 and Section 4.5), which could be attributed to its smaller size. In order to ameliorate this problem, we redesign a cross-lingual language model by combining the single-word masked language model (SMLM) with cross-lingual token-level reconstruction (XTR). Furthermore, we introduce two contrastive learning methods as auxiliary tasks to compensate for the learning bottleneck of lightweight transformer for generative tasks. Following the state-of-the-art fixed-dimensional model LASER, we proceed to learn cross-lingual sentence representations from parallel sentences, where we employ 2-layer dual-transformer encoders to shrink the model architecture. By introducing the above-stated training tasks, we establish a computationally-lite framework for training cross-lingual sentence models.

We evaluate the learned sentence representations on cross-lingual tasks including multilingual document classification (MLDoc) (Schwenk and Li, 2018) and XSR. Our results confirm the ability of our lightweight model to yield robust sentence representations. We also do a systematic study on the performance of our model in an ablative manner. The contributions of this work can be summarized as follows:

- We implement fixed-dimensional cross-lingual sentence representation learning in a lightweight model, achieving improved training efficiency and competitive performance of the learned sentence representations.
- Our proposed novel generative and contrastive tasks allow cross-lingual sentence representation efficiently trainable by the lightweight model. The contribution from each task is empirically analyzed.

2 Related Work

A majority of training tasks for learning fixed-dimensional cross-lingual sentence representations can be ascribed to one of the following 2 categories: generative or contrastive. In this section, we revisit the previous work in these 2 categories, which is crucial for designing a cross-lingual representation model.

Generative Tasks. Generative tasks measure a generative probability between predicted tokens and real tokens by training a language model. BERT-style MLM (Devlin et al., 2019) masks and predicts contextualized tokens within a given sentence. For the cross-lingual scenario, cross-lingual supervision is implemented by shared cognates and joint training (Devlin et al., 2019), concatenating source sentences in multiple languages (Conneau and Lample, 2019; Conneau et al., 2020a) or explicitly predicting the translated token (Ren et al., 2019). The [CLS] embedding or pooled embedding of all the tokens is introduced as the classifier embedding, which can be used as sentence embedding for sentence-level tasks (Reimers and Gurevych, 2019). Sequence to sequence methods (Schwenk and Douze, 2017; España-Bonet et al., 2017; Artetxe and Schwenk, 2019b; Li and Mak, 2020) autoregressively reconstruct the translation of the source sentence. The intermediate state between the encoder and the decoder are extracted as sentence representations. Particularly, the cross-lingual sentence representation quality of LASER (Artetxe and Schwenk, 2019b) benefits from a massively multilingual machine translation task covering 93 languages. In our work, we revisit the BERT-style training tasks and introduce a novel

generative loss enhanced by KL-Divergence based token distribution prediction. Our proposed generative task performs effectively for the lightweight dual-transformer framework while other generative tasks should be implemented via a large-capacity model.

Contrastive Tasks. Contrastive tasks measure (contrast) the similarities of sample pairs in the representation space. Negative sampling, which is a typical feature of the contrastive methods is first introduced in the work of word representation learning (Mikolov et al., 2013). Subsequently, contrastive tasks gradually emerged in many NLP tasks in various ways: negative sampling in knowledge graph embedding learning (Bordes et al., 2013; Wang et al., 2014), next sentence prediction in BERT (Devlin et al., 2019), token-level discrimination in ELECTRA (Clark et al., 2020), sentence-level discrimination in DeCLUTR (Giorgi et al., 2020), and hierarchical contrastive learning in HICTL (Wei et al., 2020). For the cross-lingual sentence representation training, typical ones include using correct and wrong translation pairs introduced by Guo et al. (2018); Yang et al. (2019); Chidambaram et al. (2019); Feng et al. (2020) or utilizing similarities between sentence pairs by introducing a regularization term (Yu et al., 2018). As another advantage, contrastive methods have proven to be more efficient than generative methods (Clark et al., 2020). Inspired by previous work, for our lightweight model, we propose a robust sentence-level contrastive task by leveraging similarity relationships arising from translation pairs.

3 Methodology

We perform cross-lingual sentence representation learning by a lightweight dual-transformer framework. Concerning the training tasks, we propose a novel cross-lingual language model, which combines SMLM and XTR. Moreover, we introduce two sentence-level self-supervised learning tasks (sentence alignment and sentence similarity losses) to leverage robust parallel level supervision to better conduct the cross-lingual sentence representation space alignment.

3.1 Architecture

We employ the dual transformer sharing parameters without any decoder as the basic unit to encode parallel sentences respectively, to avoid the loss in efficiency caused by the presence of a decoder. Unlike XLM (Conneau and Lample, 2019), we utilize a dual model architecture rather than a single transformer to encode sentence pairs, because it can force the encoder to capture more cross-lingual characteristics (Reimers and Gurevych, 2019; Feng et al., 2020). Moreover, we decrease the number of layers and embedding dimension to accelerate the training phase, as shown in Table 1.

The architecture of the proposed method is illustrated in Figure 1 (left). We build sentence representations on the top of 2-layer transformer (Vaswani et al., 2017) encoders by a mean-pooling operation from the final states of all the positions within a sentence. Pre-trained sentence representations for downstream tasks are denoted by $\boldsymbol{u}$ and $\boldsymbol{v}$, which are used to compute the loss for the sentence-level contrastive task. Moreover, we add a fully-connected layer before computing the loss of the cross-lingual language model inspired by Chen et al. (2020). This linear layer can enhance our lightweight model by a nontrivial margin, because the hidden state for computing loss for the generative task is far different from the sentence presentation we aim to train. Two transformer encoders and linear layers share parameters, which has been proved effective and necessary for cross-lingual representation learning (Conneau et al., 2020b).

3.2 Generative Task

SMLM. SMLM is proposed by Sabet et al. (2019), which is a variant of the standard MLM in BERT (Devlin et al., 2019). SMLM can enforce the monolingual performance, because the prediction of a number of masked tokens in MLM is too complicated for the shallow transformer encoder to learn.[2] Inspired by this, we implement SMLM by a dual transformer architecture. The transformer encoder for language l_1 predicts a masked token in a sentence in l_1 as the monolingual loss. The language l_2 encoder sharing all the parameters with l_1 encoder predicts the same masked token by the corresponding sentence (translation in l_2) as the cross-lingual loss, as shown in Figure 1 (top right). Specifically, for a parallel corpus $\boldsymbol{C}$ and language l_1 and l_2, the loss of SMLM computed from l_1 encoder $\boldsymbol{E_{l_1}}$ and

[2] A detailed comparison between SMLM and MLM under our lightweight model setting is conducted (see Section 4.5).

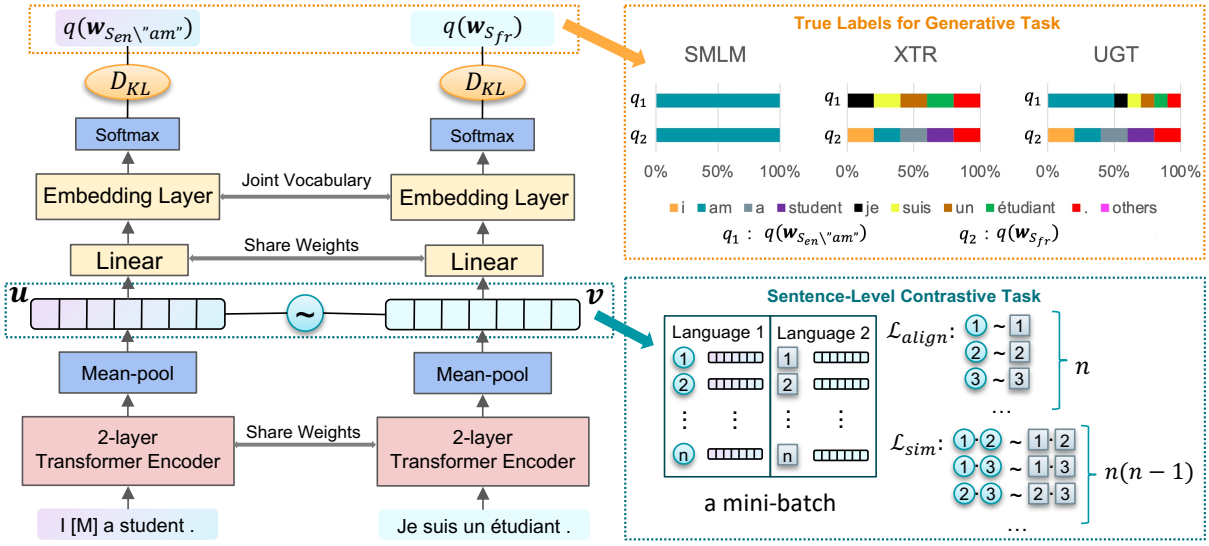

Figure 1: **Architecture of the proposed model (left), proposed unified generative task (top right), and proposed sentence-level contrastive task (bottom right).** In the left sub-figure, [M] denotes the masked token introduced by SMLM. Hidden states u and v are 512-dimensional sentence representations for the sentence-level contrastive task and for downstream tasks. In the top right sub-figure, SMLM is inspired by Sabet et al. (2019); XTR and UGT are our proposed methods. q_1 and q_2 respectively denote 2 distributions at the top of left sub-figure, the token distributions that we introduce as labels for the model to learn. In the bottom right sub-figure, n denotes the size of a mini-batch. ◯ and □ represent language l_1 and l_2, respectively. i in ◯ indicates the sentence representation of the i-th l_1 sentence in the mini-batch and same for j in □.

l_2 encoder $\boldsymbol{E_{l_2}}$ is formulated as:

$$\mathcal{L}_{SMLM} = \sum_{\substack{S\in\boldsymbol{C} \\ l,l'\in\{l_1,l_2\} \\ l\neq l'}} \{ -\log\left(P(w_t|S_{l\backslash\{w_t\}};\boldsymbol{\theta})\right) - \log\left(P(w_t|S_{l'};\boldsymbol{\theta})\right) \} \quad (1)$$

where w_t is the word to be predicted, $S_{l_1\backslash\{w_t\}}$ is a sentence in which w_t is masked, $S = (S_{l_1}, S_{l_2})$ denotes a parallel sentence pair, $\boldsymbol{\theta}$ represents the parameters to be trained in $\boldsymbol{E_{l_1}}$ and $\boldsymbol{E_{l_2}}$, and the classification probability P is computed by Softmax on the top of the embedding layer.

XTR. Inspired by LASER, we also use a reconstruction loss. However, introducing a decoder to implement the translation loss like LASER will increase the computational overhead associated with our model, which contradicts with our objective to design a computationally-lite model architecture.

To implement the reconstruction loss with just the encoder, we propose a XTR loss by which we jointly enforce the encoder to reconstruct the word distribution of corresponding target sentence as shown by q in Figure 1 (top right). Specifically, we utilize the following KL-Divergence based formulation as the training loss:

$$\mathcal{L}_{XMLM} = \sum_{\substack{S\in\boldsymbol{C} \\ l,l'\in\{l_1,l_2\} \\ l\neq l'}} \{ -\mathcal{D}_{KL}\left(p\left(\mathbf{h}_{S_l};\boldsymbol{\theta}\right) \parallel q\left(\mathbf{w}_{S_{l'}}\right)\right) -\mathcal{D}_{KL}\left(p\left(\mathbf{h}_{S_{l'}};\boldsymbol{\theta}\right) \parallel q\left(\mathbf{w}_{S_l}\right)\right) \} \quad (2)$$

where $\mathcal{D}_{KL}$ denotes KL-Divergence based loss, $p\left(\mathbf{h}_{S_l};\boldsymbol{\theta}\right)$ represents the hidden state on the top of encoder $\boldsymbol{E_l}$ as shown in Figure 1 (left) under the input S_l, and $\mathbf{w}_{S_l}$ indicates the set that contains all the tokens in S_l. We utilize discrete uniform distribution for the tokens in target language to define q for $\mathbf{w}_{S_l}$. Specifically, $q\left(\mathbf{w}_{S_l}\right)$ is defined as:

$$q\left(w_i\right) = \begin{cases} \dfrac{N_{w_i}}{\|S_{l'}\|}, & w_i \in S_{l'} \\ 0, & w_i \notin S_{l'} \end{cases} \quad (3)$$

where N_{w_i} indicates the number of words w_i in sentence $S_{l'}$ and $\|S_{l'}\|$ indicates the length of $S_{l'}$.[3]

Unified Generative Task (UGT). Finally, we unify SMLM (Eq. (1)) and XTR (Eq. (2)) by redefining the label distribution $q\left(\mathbf{w}_{S_l}\right)$ for KL-Divergence based loss. As shown in Figure 1 (top

[3]We set all the N_{w_i} to be 1 in the current implementation. Word frequency will be taken into consideration for the generative task in future work.

right), the model is forced to learn under the supervision of a biased cross-lingual probability distribution of tokens. It is formulated the same as Eq. (3) if the token w_t is masked from $S_{l'}$, else if w_t is masked within S_l:

$$q(w_i) = \begin{cases} \frac{N_{w_i}}{2\,\|S_{l'}\|}, & w_i \in S_{l'} \\ 1/2, & w_i = w_t \\ 0, & others \end{cases} \tag{4}$$

3.3 Sentence-Level Contrastive Task

Meanwhile, as shown in Figure 1 (bottom right), we introduce two auxiliary similarity-based training tasks to strengthen sentence-level supervision. We construct these two assisting tasks on the basis of mean pooled sentence representations, aiming to capture sentence similarity information across languages.

Inspired by Guo et al. (2018); Yang et al. (2019); Feng et al. (2020), we propose a sentence alignment loss. The sentence alignment loss aims to force the transformer model to recognize the sentence pair, where one sentence is the translation of the other. One positive and other negative samples contribute to the gradient update in a single batch, which provides contrastive training patterns for the model training. For contrastively discriminating positive and negative samples, we use $(batchsize - 1) \times 2$ negative samples.[4] This indicates all the sentences within a batch except the positive one will be negative samples.

More precisely, assuming the mean pooled sentence representations of S_{l_1} and S_{l_2} are $\boldsymbol{u}(S_{l_1})$ and $\boldsymbol{v}(S_{l_2})$. Assume that $\boldsymbol{B}_i$ is a specific batch of several paired sentences, $\boldsymbol{u}_{ij}$ and $\boldsymbol{v}_{ij}$ respectively indicate the representation of j-th sentence $S^{(j)} = (S^{(j)}_{l_1}, S^{(j)}_{l_2})$ in language l_1 and l_2 within batch $\boldsymbol{B}_i$. Note that the masked token w_t is omitted in the following equations. The above-proposed in-batch sentence alignment loss to align sentence pairs is defined as:

$$\mathcal{L}_{align} = -\sum_i \sum_j (\log \frac{\exp(\boldsymbol{u}_{ij}^\top \boldsymbol{v}_{ij})}{\sum_k \exp(\boldsymbol{u}_{ij}^\top \boldsymbol{v}_{ik})} + \log \frac{\exp(\boldsymbol{u}_{ij}^\top \boldsymbol{v}_{ij})}{\sum_k \exp(\boldsymbol{u}_{ik}^\top \boldsymbol{v}_{ij})}) \tag{5}$$

where $S^{(k)}, S^{(j)} \in \boldsymbol{B}_i$.

We further introduce a sentence similarity loss to better align similarities for all the sentence pairs throughout a batch. By constructing these similarity-based sentence-level contrastive tasks, we hope that it can force the sentence representations to be competent for sentence-level alignment downstream tasks. Specifically, in-batch sentence similarity loss, $\mathcal{L}_{sim}$ is formulated as:

$$\mathcal{L}_{sim} = -\sum_i \sum_j \log\cos\left\{\frac{\pi}{2}\left(\frac{\exp(\boldsymbol{u}_{ij_1}^\top \boldsymbol{u}_{ij_2})}{\sum_k \exp(\boldsymbol{u}_{ij_1}^\top \boldsymbol{u}_{ik})} - \frac{\exp(\boldsymbol{v}_{ij_1}^\top \boldsymbol{v}_{ij_2})}{\sum_k \exp(\boldsymbol{v}_{ij_1}^\top \boldsymbol{v}_{ik})}\right)\right\} \tag{6}$$

where $S^{(k)}, S^{(j)} \in \boldsymbol{B}_i$.[5]

In summary, Eq. (5) optimizes a loss for the contrastive task by discriminating correct translation from others for a given sentence, as shown in Figure 1 ($\mathcal{L}_{align}$ in bottom right). Eq. (6) aligns the cross similarities between every sentence pairs within a batch, as shown in Figure 1 ($\mathcal{L}_{sim}$ in bottom right). The similarity score matrix generated by the inner product between sentence pairs in a batch will be trained to be a symmetrical matrix with diagonal elements approximate to 1 after the Softmax operation.

3.4 Weighted Loss for Generative and Contrastive Tasks

We jointly minimize the loss of the generative task and two auxiliary contrastive tasks with the weight combination of $(1, 2, 2)$:[6]

$$\mathcal{L}(\omega_0, \omega_1, \omega_2) = \mathcal{L}_{XMLM} + 2\mathcal{L}_{align} + 2\mathcal{L}_{sim} \tag{7}$$

where $\mathcal{L}_{XMLM}$ denotes the loss of Eq. (2) and the label distribution for KL-Divergence based loss is the unified reconstruction distribution formulated by Eq. (4). $\mathcal{L}_{align}$ and $\mathcal{L}_{sim}$ represent the losses in Eq. (5) and Eq. (6), respectively.

[4] For each language, there are $batchsize - 1$ negative samples. Note that this contrastive task is different from those in Yang et al. (2019) and Feng et al. (2020), where they utilize cosine similarity while we directly use the inner product to accelerate the model.

[5] With regard to Eq. 6, log cos is employed for implementing a regression loss because we focused on the hidden states after Softmax that indicate the probabilities. We will consider using MSE loss on the states before Softmax in future exploration.

[6] We assign a bigger weight for contrastive tasks according to the task discrepency between the generative task and contrastive tasks introduced by sentence pair similarities.

4 Experiments

We evaluate our cross-lingual sentence representation models by cross-lingual document classification and bitext mining for these 2 main downstream tasks belong to 2 groups: unrelated and related to the training task. For the former, we select MLDoc (Schwenk and Li, 2018) to evaluate the classifier transfer ability of the cross-lingual model, while for the latter we conduct sentence retrieval on another parallel dataset Europarl[7] to evaluate the performance of our models.

4.1 Configuration Details

Language Pair	en-fr	en-de	en-es	en-it
Raw	51.3M	36.9M	39.0M	22.1M
Filtered	37.8M	29.6M	32.8M	17.3M

Table 2: **Training data overview.** Number of raw and filtered parallel sentences from ParaCrawl v5.0.

We build our PyTorch implementation on top of HuggingFace's Transformers library (Wolf et al., 2020). Training data is composed of the ParaCrawl[8] (Bañón et al., 2020) v5.0 datasets for each language pair. We experiment on English–French, English–German, English–Spanish and English–Italian. We filter the parallel corpus for each language pair by removing sentences that cover tokens out of 2 languages. Raw and filtered number of the parallel sentences for each pair are shown in Table 2. 10,000 sentences are selected for validation on each language pair. We tokenize sentences by SentencePiece[9] (Kudo, 2018) and build a shared vocabulary with the size of 50k for each language pair.

For each encoder, we use the transformer architecture with 2 hidden layers, 8 attention heads, hidden size of 512 and filter size of 1,024, and the parameters of two encoders are shared with each other. The sentence representations generated are 512 dimensional. For the training phase, it minimizes the weighted losses for our proposed cross-lingual language model jointly with 2 auxiliary tasks. We train 12 epochs for each language pair (30 epochs for English-Italian because of nearly half number of parallel sentences) with the Adam optimizer, learning rate of 0.001 with warm-up strategy for 3 epochs (6 epochs for English-Italian) and dropout-probability of 0.1 on a single TITAN X Pascal GPU with the batch size of 128 paired sentences. Training loss for each language pair can converge within 10 GPU (12GB)×days, which is far more efficient than most cross-lingual sentence representation learning methods.[10]

4.2 Baselines

For evaluation on the MLDoc benchmark, we use the state-of-the-art fixed-dimensional word representation methods MultiCCA+CNN method (Schwenk and Li, 2018) and Bi-Sent2Vec (Sanh et al., 2019), the representative fixed-dimensional sentence representation methods (Yu et al., 2018), LASER (Artetxe and Schwenk, 2019b), and T-LASER (Li and Mak, 2020) as baselines. In addition, as reference only, we present the results of the global fine-tuning methods, mBERT (Devlin et al., 2019) and the state-of-the-art BERT-based variant, MultiFit (Eisenschlos et al., 2019).

For the XSR task, bilingual fixed-dimensional methods, Bi-Vec (Luong et al., 2015) & Bi-Sent2Vec (Sabet et al., 2019), and multilingual fixed-dimensional methods, TransGram (Coulmance et al., 2015) & LASER (Artetxe and Schwenk, 2019b) are used as baselines.

Note that T-LASER and LASER are trained on 223M parallel sentences on 93 languages, which uses significantly more training data than ours.

We also show the results by comparing with (Reimers and Gurevych, 2020) in Appendix A, which is a recent work using global fine-tuning methods to generate multilingual sentence representations.

4.3 MLDoc: Zero-shot Cross-lingual Document Classification

The MLDoc task, which consists of news documents given in 8 different languages, is a benchmark to evaluate cross-lingual sentence representations. We conduct our evaluations in a zero-shot scenario: we train and validate a new linear

[7] https://www.statmt.org/europarl/

[8] http://opus.nlpl.eu/ParaCrawl-v5.php

[9] https://github.com/google/sentencepiece

[10] Note that it is impractical to compare the efficiency with LASER, which is trained by 80 V100 GPU×days due to different training data settings. However, it is obvious that our lightweight model is significantly more efficient than the 5-layer LSTM-based encoder-decoder model structure of LASER, because of the parallel computing nature of the transformer encoder (Vaswani et al., 2017) of our model without any decoder.

Method	en-fr		en-de		en-es		en-it		Avg.
	→	←	→	←	→	←	→	←	
fixed-dimensional word representation methods									
MultiCCA + CNN (Schwenk and Li, 2018)	72.4	64.8	81.2	56.0	72.5	74.0	69.4	53.7	68.0
Bi-Sent2Vec (Sabet et al., 2019)	81.6	82.2	86.5	79.2	74.0	71.5	**75.0**	72.6	77.8
fixed-dimensional sentence representation methods									
Yu et al. (2018)	80.8	81.0	80.2	77.1	74.1	74.1	70.8	74.8	76.6
LASER (Artetxe and Schwenk, 2019b)	78.0	80.1	86.3	**80.8**	79.3	69.6	70.2	74.2	77.3
T-LASER (Li and Mak, 2020)	70.7	78.2	86.8	79.0	71.4	74.5	68.7	76.0	75.7
Ours	**85.1**	**82.4**	**88.8**	**80.8**	**80.8**	**79.2**	74.3	**79.9**	**81.4**
reference: global fine-tuning style methods									
mBERT (Devlin et al., 2019)	83.0	-	82.4	-	75.0	-	68.3	-	-
MultiFit (Eisenschlos et al., 2019)	89.4	-	91.6	-	79.1	-	76.0	-	-

Table 3: **MLDoc benchmark results (zero-shot scenario).** We compare our models primarily with fixed-dimensional models in which Bi-Sent2vec and LASER are state-of-the-art bag-of-words based and contextual sentence representation models, respectively. We also compare with global fine-tuning style methods here for reference. Each result is the mean value of 5 runs.

Method	en-fr		en-de		en-es		en-it		Avg.
	→	←	→	←	→	←	→	←	
bilingual representation methods									
Bi-Vec (Luong et al., 2015)	81.6	83.4	71.6	68.1	81.6	83.4	74.2	72.4	77.0
Bi-Sent2Vec (Sabet et al., 2019)	87.4	87.8	84.0	84.2	89.6	89.7	**87.6**	**87.9**	87.3
Ours	**90.2**	**90.8**	**86.3**	**86.9**	**90.7**	**91.2**	86.9	87.6	**88.8**
multilingual representation methods									
TransGram (Coulmance et al., 2015)	80.4	81.6	72.7	69.1	83.8	82.7	77.9	77.2	78.2
LASER (Artetxe and Schwenk, 2019b)	95.3	94.7	94.6	94.3	94.5	94.1	95.6	95.6	94.8

Table 4: **Cross-lingual sentence retrieval results.** We report P@1 scores of 2,000 source queries when searching among 200k sentences in the target language. Here *global fine-tuning style methods* are not considered, because they require training data to be fine-tuned. Best performances among bilingual representation methods are in bold.

classifier on the top of the pre-trained sentence representations in the source language, and then evaluate the classifier on the test set for the target language. We implement the evaluation by facebook's MLDoc library.[11] As shown in Table 3, our lightweight transformer model obtains the best results for most language pairs compared with previous fixed-dimensional word and sentence representation learning methods. Our methods yield only slightly worse performance even when compared with the state-of-the-art global fine-tuning style method, MultiFit (Eisenschlos et al., 2019), on this task. This is because the entire model will be updated in the fine-tuning phase, which indicates more parameters will be task-specific after fine-tuning. For fixed-dimensional methods, just an additional dense layer will be trained, which leads to their higher efficiency.

4.4 XSR: Cross-lingual Sentence Retrieval

We also conduct an evaluation to gauge the quality of our cross-lingual sentence representations on the bitext mining task, which is identical to some components of the training task. Specifically, given 2,000 sentences in the source language, we conduct the corresponding sentence retrieval from 200K sentences in the target language. P@1 scores of our lightweight models and previous bilingual representation methods calculated by Artetxe and Schwenk (2019a) are reported. As shown in Table 4, we observe that our lightweight models outperform the bilingual pooling-based representation learning methods by a significant margin, which reflects the basic ability of the contextualized rep-

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

N	M	T	MLDoc				XSR			
			en→fr	fr→en	en→es	es→en	en→fr	fr→en	en→es	es→en
1	7,135	19	81.7	79.4	75.5	74.9	89.4	90.0	86.4	87.7
2	**11,607**	**24**	**85.1**	**82.4**	80.8	**79.2**	90.2	90.8	90.7	91.2
3	16,804	29	84.2	81.9	**81.2**	78.1	90.9	91.5	91.1	92.0
4	21,923	34	84.2	82.0	81.1	78.7	**91.4**	91.5	91.5	**92.2**
6	28,024	44	83.0	80.8	79.8	78.3	91.2	**92.0**	**91.7**	91.9

Table 5: **Training efficiencies with different numbers of layers.** N denotes number of layers within the transformer encoder; M and T indicate memory overhead (MB) and training time (min), respectively. Memory overhead changes for different languages and here we report the numbers on English–French. Training time is measured every 10,000 training steps. The results are reported by using a single V100 GPU card with the batch size of 128 sentences. 2-layer is the default setting for our lightweight model.

Tasks	MLDoc				XSR			
	en→fr	fr→en	en→es	es→en	en→fr	fr→en	en→es	es→en
MLM	78.5	77.6	74.6	75.9	19.6	25.4	11.2	28.5
SMLM	75.0	78.7	75.3	74.0	85.0	85.3	86.4	87.1
XTR	84.2	81.2	79.9	77.6	89.5	**90.8**	**90.3**	89.5
MLM $\oplus$ XTR	82.2	78.2	78.4	76.7	84.1	85.0	87.6	88.9
UGT (SMLM $\oplus$ XTR)	**85.1**	**82.4**	**80.8**	**79.2**	**89.8**	90.6	89.4	**89.6**

Table 6: **Effectiveness of different generative tasks.** UGT indicates "SMLM $\oplus$ XTR", which indicates the training task combining SMLM and XTR. MLM $\oplus$ XTR denotes the unified training task combining MLM and XTR.

resentations generated by our lightweight models. However, our lightweight models underperform LASER, which can be attributed to our lightweight capacities and bilingual settings. Note that LASER uses significantly larger multilingual training data (see Section 4.2).

4.5 Analyses

We perform ablation experiments to confirm the efficiency and the effectiveness of each training task for our models. Analyses for other hyperparameter configurations of batch size, sentence representation dimension, and training corpus size are presented in Appendix A.

Relation among Number of Layers, Efficiency, and Performances. We report the efficiency statistics and performances of our proposed methods trained by different layer number settings. As shown in Table 5, we observe a linear increase of memory occupation and training time per 10,000 training steps by increasing the number of transformer encoder layers. Specifically, a 6-layer transformer encoder occupies nearly 2.5 times memory and costs 1.8 times training time compared to our 2-layer model. Therefore, given the same memory occupation (by adjusting the batch size), theoretically our lightweight model can be implemented over 4 times ($\approx 2.5 \times 1.8$) faster than the 6-layer model. Concerning the respective performances on MLDoc and XSR, we see that lightweight model with 2 transformer layers obtains the peak performance on MLdoc, and the performances decrease when we add more layers. This indicates that the 2-layer transformer encoder is an ideal structure for our proposed training tasks on the document classification task. On the other hand, performances on XSR keep increasing gradually with more layers, where the 1-layer model can even yield decent performance on this task.

Our proposed training tasks perform well from the 2-layer model, while 6 layers are required for standard MLM and 5 LSTM layers are required for LASER. This is why we use 2-layer as the basic unit for our model.

Effectiveness of Different Generative Tasks. We report the results with different generative tasks in Table 6. We observe that XTR outperforms other generative tasks by a significant margin on both

Tasks	MLDoc				XSR			
	en→fr	fr→en	en→es	es→en	en→fr	fr→en	en→es	es→en
UGT	**85.1**	**82.4**	**80.8**	**79.2**	89.8	90.6	89.4	89.6
+ align	84.1	81.9	78.7	77.9	89.9	90.4	89.8	90.9
+ align + sim	82.3	80.3	77.6	76.2	**90.2**	**90.8**	**90.7**	**91.2**

Table 7: **Effectiveness of the contrastive tasks.** UGT indicates the training without any sentence-level contrastive tasks.

MLDoc and XSR downstream tasks. XTR yields further improvements when unified with SMLM, which is introduced as the generative task in our model. This demonstrates the necessity of a well-designed generative task for the lightweight dual-transformer architecture.

Effectiveness of the Contrastive Tasks. In Table 7, we study the contribution of the sentence-level contrastive tasks. We observe that a higher performance on MLDoc is yielded by the vanilla model while more sentence-level contrastive tasks improve the performance on XSR. This can be attributed to the similar nature between the supervision provided by sentence-level contrastive tasks and XSR process. In other words, contrastive-style tasks have a detrimental effect on the document classification downstream task. In future work, we will explore how to train a balanced sentence representation model with contrastive tasks.

5 Conclusion

In this paper, we presented a lightweight dual-transformer based cross-lingual sentence representation learning method. For the fixed 2-layer dual-transformer framework, we explored several generative and contrastive tasks to ensure the sentence representation quality and facilitate the improvement of the training efficiency. In spite of the lightweight model capacity, we reported substantial improvements on MLDoc compared to fixed-dimensional representation methods and we obtained comparable results on XSR. In the future, we plan to verify whether our proposed methods can be combined with knowledge distillation.

Acknowledgements

We would like to thank all the reviewers for their valuable comments and suggestions to improve this paper. This work was partially supported by Grant-in-Aid for Young Scientists #19K20343, JSPS.

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, Florence, Italy. Association for Computational Linguistics.

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.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 1597–1607.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259, Florence, Italy. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1113, Lisbon, Portugal. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5702–5707, Hong Kong, China. Association for Computational Linguistics.

Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. 2017. An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification. *IEEE J. Sel. Top. Signal Process.*, 11(8):1340–1350.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

John M. Giorgi, Osvald Nitski, Gary D. Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *CoRR*, abs/2006.03659.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Yunsu Kim, Hendrik Rosendahl, Nick Rossenbach, Jan Rosendahl, Shahram Khadivi, and Hermann Ney. 2019. Learning bilingual sentence embeddings via autoencoding and computing similarities with a multilayer perceptron. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 61–71, Florence, Italy. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Wei Li and Brian Mak. 2020. Transformer based multilingual document embedding model. *CoRR*, abs/2008.08567.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

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 (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Explicit cross-lingual pre-training for unsupervised machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 770–779, Hong Kong, China. Association for Computational Linguistics.

Ali Sabet, Prakhar Gupta, Jean-Baptiste Cordonnier, Robert West, and Martin Jaggi. 2019. Robust cross-lingual embeddings from parallel sentences. *CoRR*, abs/1912.12481.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. 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. European Language Resources Association (ELRA).

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online. Association for Computational Linguistics.

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

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pages 1112–1119.

Xiangpeng Wei, Yue Hu, Rongxiang Weng, Luxi Xing, Heng Yu, and Weihua Luo. 2020. On learning universal representations across languages. *CoRR*, abs/2007.15960.

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, Online. Association for Computational Linguistics.

Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5370–5378. ijcai.org.

Katherine Yu, Haoran Li, and Barlas Oguz. 2018. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 175–179, Melbourne, Australia. Association for Computational Linguistics.

A Appendices

Comparisons with Reimers and Gurevych (2020)**.** Reimers and Gurevych (2020) use the knowledge distillation to train multilingual sentence embeddings where pre-trained encoders are utilized to initialize the teacher and student model, which is a kind of the *global fine-tuning style methods* fine-tuned by parallel sentences. As the results on MLDoc and XSR shown in Table 8, their multilingual

Method	en-fr →	en-fr ←	en-de →	en-de ←	en-es →	en-es ←	en-it →	en-it ←	Avg.
MLDoc									
Reimers and Gurevych (2020)	68.0	78.5	77.6	79.2	72.7	72.2	68.5	74.2	73.9
Ours	85.1	82.4	88.8	80.8	80.8	79.2	74.3	79.9	81.4
XSR									
Reimers and Gurevych (2020)	93.0	92.3	89.9	89.2	93.9	92.9	91.7	91.4	91.8
Ours	90.2	90.8	86.3	86.9	90.7	91.2	86.9	87.6	88.8

Table 8: Comparisons with Reimers and Gurevych (2020) on MLDoc and XSR.

Batch Size	MLDoc en→fr	MLDoc fr→en	XSR en→fr	XSR fr→en
64	82.9	**82.6**	89.6	90.3
128	**84.1**	81.9	**90.2**	**90.8**
256	82.9	81.1	**90.2**	90.7

Table 9: Effect of the batch size.

Corpus Size	MLDoc en→fr	MLDoc fr→en	XSR en→fr	XSR fr→en
12.5%	82.5	80.7	**90.8**	90.5
25%	82.5	80.3	90.5	**91.2**
50%	83.0	81.5	90.2	91.0
100%	**85.1**	**82.4**	90.2	90.8

Table 10: Impact of the corpus size.

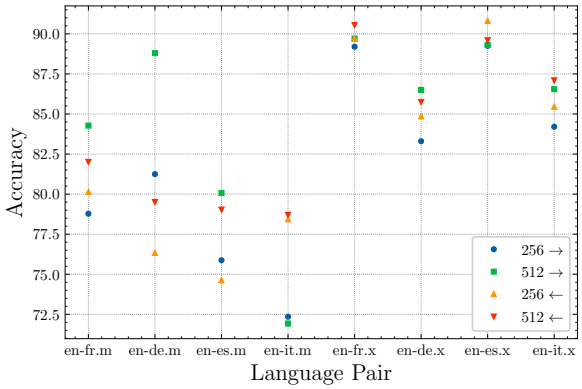

Figure 2: **Performance of different representation dimensions on MLDoc (.m) and XSR (.x).** Arrows denotes direction of zero-shot setting.

representations yield good performance on bitext mining but perform poorly on classification tasks. This demonstrates the importance of exploring task-agnostic multilingual sentence representations like LASER and ours.

Batch Size. We investigate the effect of the batch size for contrastive tasks, where different batch sizes indicate the discrepancy of the negative sample numbers. As shown in Table 9, larger batch harms the lightweight model based sentence representation learning and 128 is reported as the best batch size setting for our lightweight model. Furthermore, batch size of 128 allows the training to be assigned on 12GB GPU card while a larger batch size requires more GPU memory.

Corpus Size. We show the impact of the size of the parallel corpus on English-French in Table 10. For MLDoc, we observe higher accuracy on larger corpus while for XSR, a small fraction of the large corpus suffices to yield effective results. This indicates that more parallel data improves the performance on MLDoc.

Sentence Representation Dimension. In Figure 2, we present the effect of the sentence representation dimension. 512-dimensional sentence representations significantly outperform 256-dimensional ones in our lightweight model. Moreover, representation size of 512 yields better performance without increasing the training time.