Learning Cross-lingual Sentence Representations for Multilingual Document Classification with Token-level Reconstruction

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

1 Introduction

Fixed-dimensional cross-lingual sentence representations [1, 2, 3, 4, 5, 6, 7] are widely used to initialize another light neural network for cross-lingual sentence understanding downstream tasks. When learning such representations, efficiency and robustness form the core. In this work, we explore different training tasks for learning fixed-dimensional cross-lingual sentence representations in a 2-layer dual-transformer framework aiming to improve the efficiency and retain the robustness. We 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 **Single-word Masked Language Model (SMLM)** with **Token-level Reconstruction (TR)** task. Our comparisons with competing models on multilingual document classification verify the effectiveness of the newly proposed training tasks for a shallow model.

2 Related Work

Training tasks for cross-lingual sentence representation learning can be classified into 2 groups: generative or contrastive. As our proposed method focuses on improving previous generative methods, we mainly revisit related generative methods in this section. Generative tasks measure a generative probability between predicted tokens and real tokens by training a language model. BERT-style masked language models [8] mask and predict contextualized tokens within a given sentence. For the cross-lingual scenario, cross-lingual supervision is implemented by shared cognates and joint training (mBERT [8]), concatenating source sentences in multiple languages [9, 10] or explicitly predicting the translated token [11].[1] [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 [12]. Sequence to sequence methods [1, 3, 6] autoregressively reconstruct the translation of the source sentence. The intermediate state between the encoder and the decoder are extracted as sentence representations. Particularly, the outstanding cross-lingual sentence representation quality of LASER [3] benefits from a massive multilingual machine translation task covering 93 languages. In our work, we revisit the BERT-style training tasks and introduce a novel proper generative loss for the tiny dual-transformer framework.

3 Proposed Methods

We perform cross-lingual sentence representation learning by a tiny dual-transformer framework. With regard to the training tasks, we propose a novel cross-lingual language model which combines SMLM and TR.

3.1 Architecture

We employ the dual transformer sharing parameters with each other without any decoder as the basic unit to encode parallel sentences respectively and avoid the loss in efficiency caused by decoder. First, in order to design a tiny model, we do not additionally employ any decoder to reconstruct sentences. Second, unlike TLM [9], we utilize dual model architecture rather than a single transformer to encode sentence pairs since it can force the encoder to capture more cross-lingual characteristics [5, 12]. Moreover, to establish a tiny model, we decrease the number of layers and embedding dimension to accelerate the training phase.

[1] https://github.com/google-research/bert/blob/master/multilingual.md

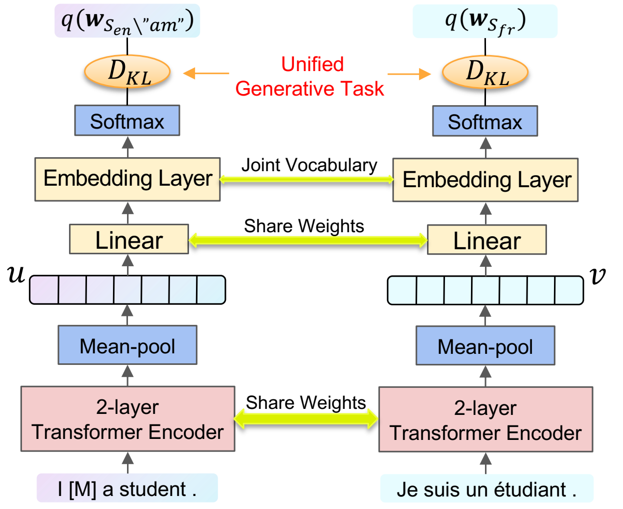

Figure 1 Model architecture

The architecture of the proposed method is illustrated in Figure 1. We build sentence representations on the top of 2-layer transformer [13] 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 u and v. Moreover, we add a fully-connected layer before computing the loss of the cross-lingual language model. This linear layer can enhance our tiny model by a nontrivial margin for that the hidden state for computing loss for the generative task is far different from the sentence presentation we aim to train.[2)] Two transformer encoders and linear layers share parameters with each other which has been proved effective and necessary for cross-lingual representation learning [15].

3.2 Training Task

Single-word Masked Language Model (SMLM). SMLM is a single-word masked language model proposed by Sabet et al. [16]. Following them, we implement their task by a dual transformer architecture, where the transformer encoder for language l_1 predicts a masked token in a sentence in l_1 as the monolingual loss, while 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. 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 l_2 encoder $\boldsymbol{E}_{l_2}$ is formulated as:

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

where w_t is the word to be predicted, $S_{l_1\backslash\{w_t\}}$ is a sentence in which w_t is corrupted, $S = (S_{l_1}, S_{l_2})$ denotes a parallel sentence pair, θ 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.

We employ SMLM which is a variant of the standard Masked Language Model (MLM) in BERT to enforce the monolingual performance since MLM is a quite complicated task for a shallow transformer encoder to learn. A detailed comparison between SMLM and MLM under our tiny model setting is conducted (see Section 5.3).

Token-level Reconstruction (TR). As demonstrated above, MLM suffices not to capture the cross-lingual features for a tiny transformer encoder, which indicates that it is essential to introduce the reconstruction loss like that in LASER for cross-lingual representation learning. However, introducing a decoder to implement the translation loss will increase the training time by a large margin which contradicts with our aiming to design a computationally-lite model architecture.

To implement the reconstruction loss with just the encoder, we introduce a TR loss by which we jointly enforce the encoder to reconstruct the word distribution of corresponding target sentence as shown by q in Figure 1. This loss is inspired from BiVec loss formulation [17] in which cross-lingual tokens are reconstructed on the basis of Skip-gram language model pattern [18]. Specifically, we utilize the following KL-Divergence based formulation as the training loss:

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

where $\mathscr{D}_{KL}$ denotes KL-Divergence based loss and $p\left(\mathbf{h}_{S_l};\theta\right)$ represents the hidden state on the top of encoder $\boldsymbol{E}_l$ as shown in Figure 1 under the input S_l. We utilize discrete uniform distribution for the tokens in target language

[2)] Adding a linear layer is similarly conducted in Chen et al. [14]. We can hardly obtain any accuracy on downstream tasks without this linear layer.

to define q. Specifically, $q\left(\mathbf{w}_{S_l}\right)$ is defined as:

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

where N_{w_i} means the number of word w_i in sentence $S_{l'}$ and $\left\|S_{l'}\right\|$ means the length of $S_{l'}$.

Unified Generative Task. Finally, we unify SMLM (Eq. (1)) and TR (Eq. (2)) by redefine the label distribution $q\left(\mathbf{w}_{S_l}\right)$ for KL-Divergence based loss. It is formulated the same as Eq. (3) if the token w_t is corrupted from $S_{l'}$, else if w_t is corrupted within S_l:

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

4 Training Details

We build our PyTorch implementation on the top of HuggingFace's Transformers.[3] Training data is composed of ParaCrawl v5.0 datasets for each language pair.[4] We experiment on English–French, English–German, English–Spanish and English–Italian. We filter the parallel corpus for each language pair by removing sentences which cover tokens out of 2 languages. Raw and filtered number of the parallel sentences for each pair are shown in Table 1. We tokenize sentences in all the languages by SentencePiece and build a shared vocabulary with the size of 50k for each language pair.[5]

Table 1 Training data overview

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

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 loss for our proposed cross-lingual language model (specifically, **Unified Generative Task** in 3.2). We train 12 epochs for each language pair (30 epochs for English-Italian because of nearly half number of parallel sentences) with an 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 by 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.

5 Evaluation

We evaluate our cross-lingual sentence representation models by cross-lingual document classification. We select MLDoc [19] to evaluate the classifier transfer ability of the cross-lingual sentence representation model.

5.1 MLDoc: Multilingual 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 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.[6]

5.2 Results

As shown in Table 2, our tiny transformer model obtains the best results for most language pairs compared with MLDoc baseline [19], previous fixed-dimensional word [16] and sentence [2, 3, 6] representation learning methods. This demonstrates that our proposed method can obtain state-of-the-art results on MLDoc benchmark among fixed-dimensional representation methods with an extremely light and efficient training framework.

On the other hand, our method yield only slightly worse performance even when compared with the state-of-the-art *pre-training* + *fine-tuning* style methods on this task which are referred as *reference* in Table 2.[7] This is because the entire model will be updated in the fine-tuning phase

[3] https://github.com/huggingface/transformers

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

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

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

[7] *pre-training* + *fine-tuning* style methods first learn a cross-lingual language model, then conduct a deliberate task-specific fine-tuning on downstream task [8, 20]. Sentence representations trained from this kind of methods are usually of higher quality but lower fine-tuning efficiency.

Table 2 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 pre-training + fine-tuning style methods here for reference.

Method	en-fr		en-de		en-es		en-it		Avg.
	→	←	→	←	→	←	→	←	
baseline									
MultiCCA + CNN [19]	72.4	64.8	81.2	56.0	72.5	74.0	69.4	53.7	68.0
fixed-dimensional word representation methods									
Bi-Sent2Vec [16]	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. [2]	80.8	81.0	80.2	77.1	74.1	74.1	70.8	74.8	76.6
LASER [3]	78.0	80.1	86.3	**80.8**	79.3	69.6	70.2	74.2	77.3
T-LASER [6]	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: pre-training + fine-tuning style methods									
mBERT [8]	83.0	-	82.4	-	75.0	-	68.3	-	-
MultiFit [20]	89.4	-	91.6	-	79.1	-	76.0	-	-

Table 3 Different Generative Tasks

Tasks	en→fr	fr→en
MLM	78.5	77.6
SMLM	75.0	78.7
TR	84.2	81.2
MLM + TR	82.2	78.2
SMLM + TR	**85.1**	**82.4**

while merely an additional dense layer will be trained for fixed-dimensional methods, which leads to their higher efficiency.

5.3 Different Generative Tasks

We report the results with different generative tasks in Table 3. We observe that proposed TR outperforms other generative tasks by a significant margin on MLDoc benchmark. TR will yield further improvements if unified with SMLM, which is introduced as **Unified Generative Task** above in 3.2. This demonstrates the necessity of a well-designed generative task for the tiny dual-transformer architecture. Moreover, by comparing the en-fr and fr-en results of LASER in Table 1 and *MLM tasks in Table 3, we observe that the reconstruction of the original sentence in other languages does benefit the representation quality. This reveals why our proposed TR can achieve promising results whereas TR does not reconstruct the translation directly.

6 Conclusion

In this work, we proposed a novel cross-lingual language model, which combines the SMLM with TR task for cross-lingual sentence representations learning in the light training framework. In spite of the tiny model capacity, we obtained substantial improvements on the MLDoc benchmark compared to fixed-dimensional representation methods. In the future, we plan to verify the performance of the newly proposed losses in large model architectures and extend this bilingual based method to multilingual settings.

Acknowledgments

This work was supported by Grant-in-Aid for Young Scientists #19K20343, JSPS.

References

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

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

[3] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 597–610, March 2019.

[4] Yunsu Kim, Hendrik Rosendahl, Nick Rossenbach, Jan Rosendahl, Shahram Khadivi, and Hermann Ney. 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)*, pp. 61–71, Florence, Italy, August 2019. Association for Computational Linguistics.

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

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

[7] Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 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)*, pp. 250–259, Florence, Italy, August 2019. Association for Computational Linguistics.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 7057–7067, 2019.

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

[11] Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. 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)*, pp. 770–779, Hong Kong, China, November 2019. Association for Computational Linguistics.

[12] Nils Reimers and Iryna Gurevych. 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)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 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*, pp. 1597–1607, 2020.

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

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

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

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[19] 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 2018. European Languages Resources Association (ELRA).

[20] Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kardas, Sylvain Gugger, and Jeremy Howard. Multifit: Efficient multi-lingual language model fine-tuning. *CoRR*, Vol. abs/1909.04761, , 2019.