Learning Multilingual Sentence Representations with Cross-lingual Consistency Regularization

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang
Baidu Inc. No. 10, Shangdi 10th Street, Beijing, 100085, China
{gaopengzhi,zhangliwen04,hezhongjun,wu_hua,wanghaifeng}@baidu.com

Abstract

Multilingual sentence representations are the foundation for similarity-based bitext mining, which is crucial for scaling multilingual neural machine translation (NMT) system to more languages. In this paper, we introduce MuSR: a one-for-all **Mu**ltilingual **S**entence **R**epresentation model that supports 223 languages. Leveraging billions of English-centric parallel corpora, we train a multilingual Transformer encoder, coupled with an auxiliary Transformer decoder, by adopting a multilingual NMT framework with CrossConST, a cross-lingual consistency regularization technique proposed in Gao et al. (2023). Experimental results on multilingual similarity search and bitext mining tasks show the effectiveness of our approach. Specifically, MuSR achieves superior performance over LASER3[1] (Heffernan et al., 2022) which consists of 148 independent multilingual sentence encoders.[2]

1 Introduction

Multilingual sentence representation models (Artetxe and Schwenk, 2019b; Yang et al., 2020; Reimers and Gurevych, 2020; Feng et al., 2022; Heffernan et al., 2022; Mao and Nakagawa, 2023) align different languages in a shared representation space, facilitating similarity-based bitext mining that extracts parallel sentences for learning multilingual neural machine translation (NMT) systems (Schwenk et al., 2021a,b). Specifically, LASER3 (Heffernan et al., 2022) scales the original LASER (Artetxe and Schwenk, 2019b) beyond the 93 widely used languages and achieves the state-of-the-art (SOTA) performance on the multilingual sentence alignment tasks over 200 languages.

[1]In its original context, LASER3 refers solely to the language-specific models presented in Heffernan et al. (2022). For simplicity, we use LASER3 as an umbrella term encompassing the multilingual model LASER2 and the language-specific models discussed in this paper.

[2]Previous presentations of this work are available at https://arxiv.org/abs/2306.06919.

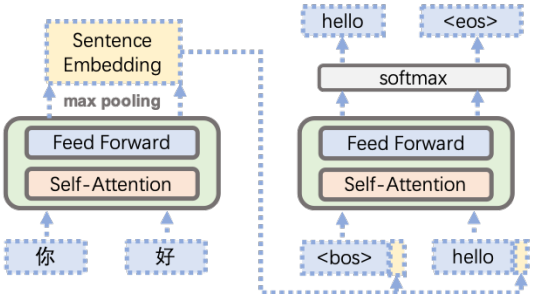

Figure 1: The model architecture of our approach for learning multilingual sentence representations.

Although LASER3 exhibits remarkable performance, it is not a one-for-all multilingual sentence representation model. Instead, it comprises of one multilingual model called LASER2 and 147 language-specific models, which are learned through a teacher-student training mechanism. Such model strategy, although effective, results in substantial storage overhead of 78GB and degraded transfer performance from high-resource to low-resource languages, which hinders its practical value in natural language processing (NLP).

In this paper, our primary goal is to learn a unified multilingual sentence encoder, MuSR, to handle a wide range of languages such that semantic-equivalent sentences in different languages are close to each other in the representation space. Inspired by the cross-lingual consistency for multilingual NMT (Gao et al., 2023), we learn multilingual sentence embeddings by utilizing a many-to-one multilingual NMT training paradigm with cross-lingual consistency regularization (Figures 1 and 2). In order to support a wide range of languages, we collect about 5.5 billion English-centric parallel sentences covering 223 languages from both open-source and in-house datasets. To the best of our knowledge, MuSR is the first one-for-all multilingual sentence representation model that supports more than 220 languages. The contributions of this paper can be summarized as follows:

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 243–262
December 6-10, 2023

Method	#Models	#Parameters	#Languages	Task	Architecture	Monolingual	Pretrain
LASER2	1	45M	93	Seq2Seq	Bi-LSTM		
LASER3	1 + 147	N/A	205	Dual Encoder	Transformer	✓	
LaBSE	1	471M	109	Dual Encoder	Transformer	✓	✓
MuSR	1	434M	223	Seq2Seq	Transformer		

Table 1: Comparison between the related works and our approach. Note that language-specific models in LASER3 have different vocabulary size, and the number of parameters for each model can be approximately calculated as 202M + vocabulary size $\times$ 1024. "Monolingual" denotes whether the monolingual data is used for training. "Pretrain" denotes whether the model relies on the language model pretraining.

- We learn a one-for-all multilingual sentence representation model, MuSR, by leveraging many-to-one multilingual NMT training with CrossConST regularization over 5.5 billion English-centric parallel corpora.
- Our experimental results show that MuSR achieves impressive performance on the multilingual benchmarks and outperforms the SOTA models LaBSE (Feng et al., 2022) and LASER3 (Heffernan et al., 2022).
- We publicly release MuSR, the multilingual sentence representation model that supports 223 languages.[3]

2 Background

2.1 Multilingual Sentence Representation

As an important component of cross-lingual and multilingual NLP, multilingual sentence representation has attracted increasing attention in the NLP community. One direction is to leverage dual-encoder architecture to learn language-agnostic representations. Guo et al. (2018) demonstrate the effectiveness of the dual-encoder model for learning bilingual sentence embeddings, and Yang et al. (2019) extend the dual-encoder model with additive margin softmax loss. Based on these works, LaBSE (Feng et al., 2022) utilizes dual Transformer encoders to learn language-agnostic embeddings over 109 languages with additive margin softmax loss, which is also pretrained with masked language modeling (MLM) and translation language modeling (TLM) (Conneau and Lample, 2019). LEALLA (Mao and Nakagawa, 2023) further constructs low-dimensional sentence embeddings by leveraging knowledge distillation based on LaBSE.

Another direction is to utilize encoders from multilingual NMT to produce universal representations across different languages. LASER (Artetxe and Schwenk, 2019b) learns the multilingual sentence embeddings over 93 languages based on the NMT model with a Bi-LSTM encoder and a LSTM decoder. Heffernan et al. (2022) replace the original LASER model with LASER2 by introducing SentencePiece (Kudo and Richardson, 2018) vocabulary, up-sampling the low-resource languages, and adopting a new fairseq[4] implementation. LASER2 is used as the teacher, and 147 language-specific sentence representation models are learned by utilizing teacher-student and MLM training mechanisms. LASER3 refers to a group of LASER2 and 147 language-specific models across 205 languages. The comparison between the existing works and our approach are summarized in Table 1.

2.2 Cross-lingual Consistency Regularization for Multilingual NMT

The multilingual NMT model refers to a neural network with an encoder-decoder architecture, which receives a sentence in one language as input and returns a translated sentence in another language as output. Assume $\mathbf{x}$ and $\mathbf{y}$ correspond to the source and target sentences respectively, and let $\mathcal{S}$ denotes the multilingual training corpus. The standard training objective is to minimize the empirical risk:

$$\mathcal{L}_{ce}(\theta) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\in\mathcal{S}}[\ell(f(\mathbf{x},\mathbf{y};\theta),\ddot{\mathbf{y}})], \quad (1)$$

where ℓ denotes the cross-entropy loss, θ is a set of model parameters, $f(\mathbf{x},\mathbf{y};\theta)$ is a sequence of probability predictions, i.e.,

$$f_j(\mathbf{x},\mathbf{y};\theta) = P(y|\mathbf{x},\mathbf{y}_{<j};\theta), \quad (2)$$

and $\ddot{\mathbf{y}}$ is a sequence of one-hot label vectors for $\mathbf{y}$.

Gao et al. (2023) introduce a cross-lingual consistency regularization, CrossConST, to bridge the representation gap among different languages in the training of multilingual NMT model. For each

[3] Our implementations are available at https://github.com/gpengzhi/CrossConST-SR.

[4] https://github.com/facebookresearch/fairseq

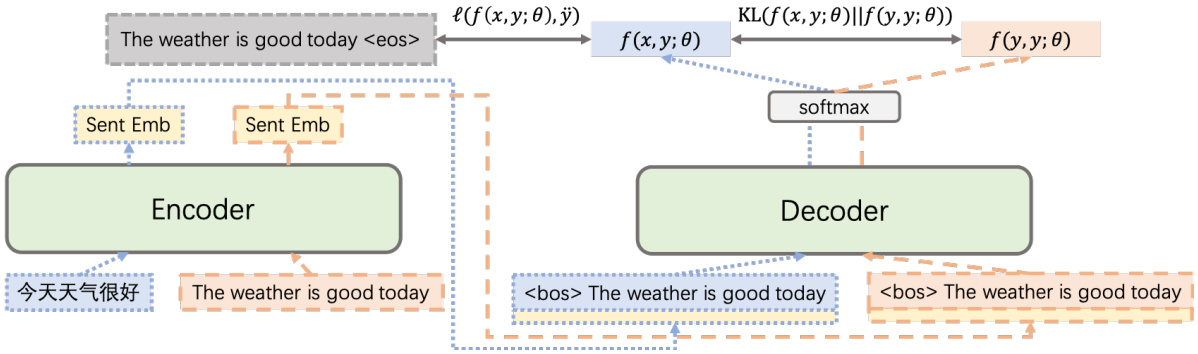

Figure 2: Illustration of CrossConST regularization for learning multilingual sentence representations, where the original Chinese-English sentence pair ("今天天气很好", "The weather is good today") and the copied English-English sentence pair ("The weather is good today", "The weather is good today") are fed into the multilingual NMT model to generate two output distributions $f(\mathbf{x}, \mathbf{y}; \theta)$ and $f(\mathbf{y}, \mathbf{y}; \theta)$.

sentence pair $(\mathbf{x}, \mathbf{y})$, the training objective of CrossConST is defined as:

$$\mathcal{L}_{CrossConST}(\theta) = \mathcal{L}_{ce}(\theta) + \alpha \mathcal{L}_{kl}(\theta), \quad (3)$$

where

$$\mathcal{L}_{kl}(\theta) = \mathrm{KL}(f(\mathbf{x}, \mathbf{y}; \theta) \| f(\mathbf{y}, \mathbf{y}; \theta)), \quad (4)$$

$\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler (KL) divergence between two distributions, and α is a scalar hyper-parameter that balances $\mathcal{L}_{ce}(\theta)$ and $\mathcal{L}_{kl}(\theta)$.

3 Methodology

Following the similar problem formulation of Artetxe and Schwenk (2019b), our approach is based on a Transformer encoder-decoder architecture trained with English-centric parallel corpora. We discuss the details of our model architecture and training strategy as follows.

3.1 Model Architecture

The overall model architecture is illustrated in Figure 1. Multilingual sentence embeddings are calculated by applying a max-pooling operation over the Transformer encoder's output, which is subsequently concatenated to the word embeddings at the Transformer decoder's input. Note that we discard the cross-attention module in the Transformer decoder. The sentence embeddings are the only connection between the encoder and the decoder such that all relevant information of the input sentences are captured by the corresponding sentence representations. Note that our model does not need language tags, as many-to-one multilingual NMT does not rely on them, unlike LASER in Artetxe and Schwenk (2019b).

3.2 Training Strategy

Following Gao et al. (2023), we adopt a two-stage training strategy to stabilize the multilingual NMT training procedure and accelerate the convergence of the multilingual NMT model. Instead of utilizing two target languages (English and Spanish) as in Artetxe and Schwenk (2019b), we consider only one target language (English) and formulate our problem as a many-to-one multilingual NMT task. We first train a multilingual NMT model as the pretrained model and then finetune the model with CrossConST objective function (3). Figure 2 illustrates CrossConST regularization for learning multilingual sentence representations. Through the application of CrossConST, sentence embeddings of the target language are aligned to the representation space of the source languages. The alignment process is facilitated by our many-to-one multilingual NMT model, which effectively encodes all languages into a shared representation space.

4 Datasets and Training Configurations

4.1 Datasets

We use a combination of open-source datasets and in-house datasets in our experiments.[5]

Open-source Dataset We collect all English-centric parallel datasets from the OPUS collection[6] (Tiedemann, 2012) up to October 2022, which is comprised of multiple corpora, ranging from movie subtitles (Tiedemann, 2016) to Bible (Christodouloupoulos and Steedman, 2015) to web crawled datasets (El-Kishky et al., 2020; Schwenk

[5] See the list of the supported languages in Table 5.
[6] http://www.opus.nlpl.eu

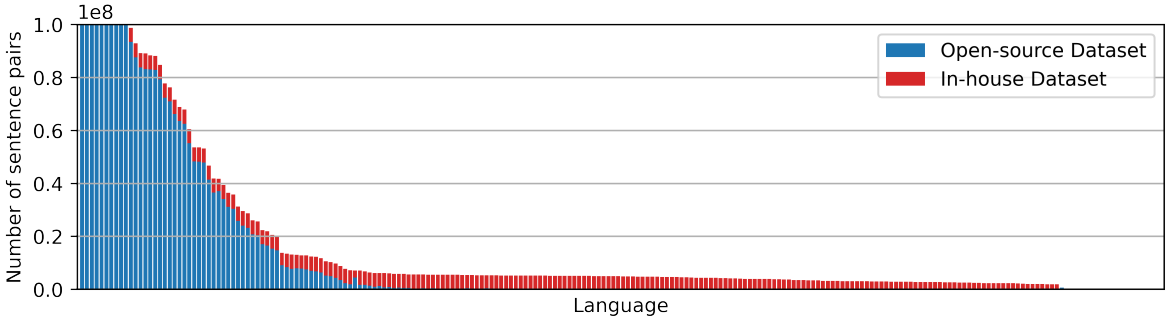

Figure 3: The distribution of the open-source and in-house cleaned datasets for each language in our training dataset. Note that the sentences for each language are capped at 100 million for better illustration. Please check Figure 6 for the complete distribution with the corresponding language name.

et al., 2021b). We download all available English-centric corpora and concatenate them without curating the datasets or trying to balance the representation of different domains.

In-house Dataset We also leverage all English-centric in-house datasets which consists of the following resources: 1) The parallel sentences are constructed from web pages by utilizing a bitext mining system. The extracted sentence pairs are filtered by a predefined scoring threshold. 2) We adopt the 3.3B multilingual NMT model released by the No Language Left Behind (NLLB) project[7] and translate the English sentences from the ParaCrawl project[8] (Bañón et al., 2020) into different languages. 3) We leverage our in-house multilingual NMT model to translate the in-house English corpus into different languages.

After we collect all parallel datasets, we adopt the data cleaning process as follows: 1) We remove duplicate sentence pairs and also discard sentence pairs wherein the English sentences exceed 5000 characters. 2) Language identification filtering is applied by utilizing fastText toolkit (Joulin et al., 2016, 2017). If the language is not supported by the identification model[9], we simply check whether the language is non-English. 3) Dual conditional cross-entropy filtering (Junczys-Dowmunt, 2018) is performed based on our in-house multilingual NMT models. Specifically, for a sentence pair $(\mathbf{x}, \mathbf{y})$, we identify they are translations of each other by leveraging the score defined as follows:

$$|H(y|x) - H(x|y)| + \frac{1}{2}(H(y|x) + H(x|y)),$$

where $H(\cdot|\cdot)$ denotes the word-normalized conditional cross-entropy loss based on the multilingual NMT model. After the cleaning process, we discard the languages which have less than 1000 sentence pairs. In summary, we collect about 5.5 billion cleaned English-centric sentence pairs covering 223 languages including English. The distribution of our training datasets for each language is illustrated in Figure 3.

We can see that there is a discrepancy of 5 orders of magnitude between the highest (Spanish) and the lowest (Algerian Arabic) resource languages. To strike a balance between high and low resource language pairs, we adopt a temperature-based sampling strategy (Arivazhagan et al., 2019; Bapna and Firat, 2019). Sentence pairs are sampled according to a multinomial distribution with probability $\{q_i\}_{i=1,\ldots,N}$, where

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^N p_j^\alpha} \quad \text{with} \quad p_i = \frac{n_i}{\sum_{k=1}^N n_k}, \qquad (5)$$

N denotes the number of languages, and n_i denotes the number of sentence pairs for each language. We consider $\alpha = 0.5$ in our experiments. Sampling with this distribution increases the number of sentence pairs associated to low resource languages and alleviates the bias towards high resource languages. We collect 500 million sentences with such sampling strategy and learn a shared dictionary with 256K byte-pair-encoding (BPE) (Sennrich et al., 2016) types using SentencePiece[10]. We keep tokens occurring no less than 20, which results in a subword vocabulary of 344,276 tokens.

[7] https://github.com/facebookresearch/fairseq/tree/nllb

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

[9] https://fasttext.cc/docs/en/language-identification.html

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

Model	Tatoeba	Flores-101			Flores-200		
	xx ↔ en	xx ↔ en	xx ↔ zh	xx ↔ yy	xx ↔ en	xx ↔ zh	xx ↔ yy
LASER2	69.95	67.78	64.47	44.90	56.98	52.76	31.96
LaBSE	83.23	96.43	95.46	91.00	88.48	86.06	74.92
LASER3	78.08	98.30	96.18	93.62	93.71	90.64	82.26
MuSR	**83.96**	**99.23**	**98.48**	**97.83**	**97.37**	**95.95**	**93.21**

Table 2: Our approach achieves the superior performance over the existing SOTA models on the Tatoeba and Flores benchmarks. The detailed experimental results in English (xx ↔ en) and Chinese (xx ↔ zh) directions are summarized in Tables 6, 7, 8, 9, and 10. The experimental results on the Flores-200 benchmark in all language (xx ↔ yy) directions are illustrated in Figure 5.

4.2 Training Configurations

We implement our approach on top of the Transformer (Vaswani et al., 2017). We apply a Transformer with 12 encoder layers and 3 decoder layers, 8 attention heads, embedding size 768, and FFN layer dimension 768×4 and $768 \times 2 \times 4$ for encoder and decoder respectively. We apply cross-entropy loss with label smoothing rate 0.1 and set max tokens per batch to be 1024. We use the Adam optimizer with Beta $(0.9, 0.98)$, 10000 warmup updates, and inverse square root learning rate scheduler with initial learning rates $7e^{-4}$. We set max source positions and max target positions to be 256 and use dropout rate 0.1. We apply the same training configurations in both pretraining and finetuning stages. We fix α to be 1.0 in (3) for CrossConST. We train all models until convergence on 8×4 NVIDIA Tesla V100 GPUs.

5 Experimental Evaluation

Following the evaluation setup of Heffernan et al. (2022), we here investigate the performance of multilingual sentence embeddings on two tasks: multilingual similarity search and bitext mining.

5.1 Multilingual Similarity Search

Given the parallel sentence pairs, we find the nearest neighbor for each sentence in the other language according to the sentence embedding cosine similarity and compute the corresponding accuracy. We conduct our experiments on the following datasets:

Tatoeba Tatoeba is a multilingual dataset covering 112 languages (Artetxe and Schwenk, 2019b), which contains up to 1000 sentences per language along with their English translations.[11]

Flores-200 Flores-200 is a multilingual dataset made publicly available by the NLLB project (Costa-jussà et al., 2022), which covers 204 languages.[12] We perform the evaluation on the devtest which includes 1012 sentences for each language. We also evaluate on Flores-101 which is a subset of Flores-200 and covers 102 languages.

We report the averaged bidirectional similarity search accuracy on the Tatoeba, Flores-101, and Flores-200 benchmarks in Table 2. The English direction represents the supervised performance of MuSR, while the Chinese direction exemplifies the effectiveness in the zero-shot scenario. Note that there are 5151 and 20706 bidirectional language directions (xx ↔ yy) in Flores-101 and Flores-200 benchmarks respectively. We can see that our approach significantly outperforms the current SOTA models LaBSE and LASER3. It is worth mentioning that MuSR achieves an improvement of over 4.7% accuracy on average over LASER3 that consists of 148 independent sentence embedding models. The performance gap between English and Chinese in LaBSE, the model with the smallest discrepancy, stands at 0.97% and 2.42% on Flores-101 and Flores-200 respectively. In contrast, MuSR exhibits a substantially smaller divergence of 0.75% and 1.42% on these two directions, indicating our superior capability to model various languages within the shared representation space.

As discussed in Heffernan et al. (2022), Tatoeba is less reliable for evaluating multilingual sentence embeddings since it mainly contains very short sentences which can introduce a strong bias towards a particular model or training corpus. We here illustrate the distribution of the averaged bidirectional accuracy of the strong baselines and MuSR on the Flores-200 benchmark in Figure 4. Note that the language order in the x-axis is selected by the descending similarity search accuracy of MuSR on the Flores-200 benchmark. We can see that our approach performs strongly across a wide range

[11] https://github.com/facebookresearch/LASER/tree/main/data/tatoeba/v1

[12] https://github.com/facebookresearch/flores/tree/main/flores200

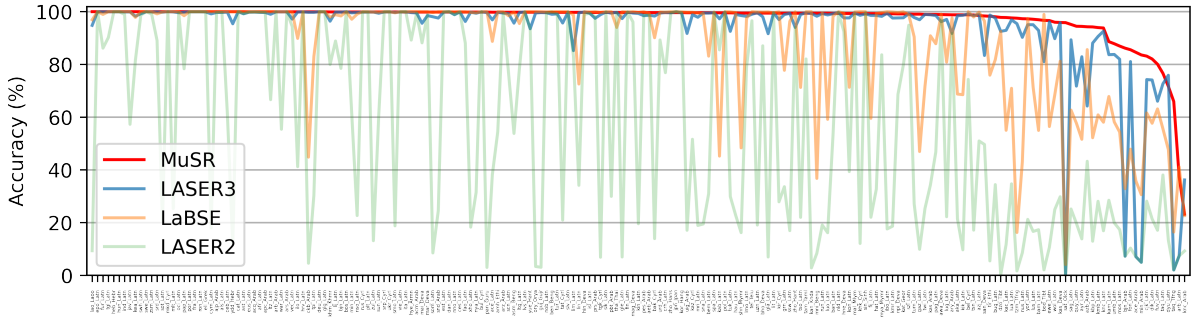

Figure 4: The distribution of the averaged bidirectional accuracy with English of the multilingual similarity search on the Flores-200 benchmark.

of languages, with over 150 languages achieving a similarity search accuracy exceeding 99%. LASER2 shows high variance across languages, and it could be resolved to some extent by incorporating language-specific models in LASER3.

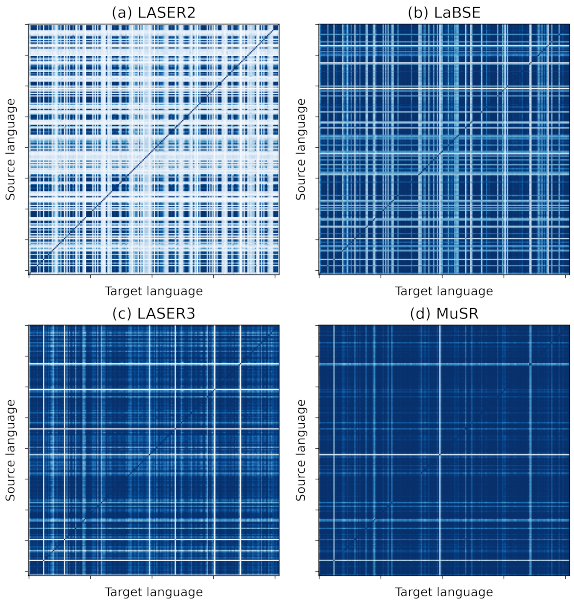

Figure 5: The accuracy distribution of the similarity search task from the source language to the target language on the Flores-200 benchmark. The darker the entry shows, the higher the accuracy is. Please check Figures 7, 8, 9, and 10 for better illustration with the corresponding similarity search accuracy.

The multilingual similarity search performance across all languages (xx ← yy and xx → yy) of the strong baselines and MuSR on the Flores-200 benchmark are visualized in Figure 5, where each entry of the 204 × 204 matrix stands for the corresponding accuracy of the similarity search task from the source language to the target language. We can see that MuSR consistently outperforms the strong baselines across a wide range of languages, with over 80% of language directions achieving a similarity search accuracy exceeding 90%. Note that LASER2, LaBSE, and LASER3 only have around 12%, 49%, and 56% of language directions achieving similarity search accuracy exceeding 90% on the Flores-200 benchmark.

5.2 Bitext Mining

Given two comparable corpora in different languages, we identify the sentence pairs that are translations of each other by leveraging the score (Artetxe and Schwenk, 2019a) defined as follows:

$$\frac{\cos(\mathbf{x},\mathbf{y})}{\sum_{\mathbf{z}\in \mathrm{NN}_k(\mathbf{x})} \frac{\cos(\mathbf{x},\mathbf{z})}{2k} + \sum_{\mathbf{z}\in \mathrm{NN}_k(\mathbf{y})} \frac{\cos(\mathbf{y},\mathbf{z})}{2k}}, \quad (6)$$

where $\mathbf{x}$ and $\mathbf{y}$ are the source and target sentence embeddings respectively, and $\mathrm{NN}_k(\mathbf{x})$ denotes the k nearest neighbors of $\mathbf{x}$ in the other languages. We score each sentence pair by calculating (6), and the parallel sentences are extracted and filtered by setting a fixed threshold over this score.

We conduct experiments on the BUCC dataset (Zweigenbaum et al., 2018) containing comparable corpora between English and four other languages: German (de), French (fr), Russian (ru), and Chinese (zh), using exact same hyperparameters as Artetxe and Schwenk (2019a)[13]. We set k to be 4 in our experiments. Given the monolingual corpora and the gold translation pairs, we extract the translation pairs from the monolingual data and evaluate against the ground truth. Following Feng et al. (2022), we evaluate the performance by F1 score on the training dataset since the ground truth for the test dataset is not released.

We report the F1 scores of the strong baselines and our approach in Table 3. We can see that MuSR

[13] https://github.com/facebookresearch/LASER/tree/main/tasks/bucc

Model	de	fr	ru	zh	avg.
LASER2	**95.36**	92.15	91.95	91.07	92.63
LaBSE	**95.86**	**92.52**	**92.46**	**92.99**	**93.46**
LASER3	**95.36**	92.15	91.95	91.07	92.63
MuSR	94.91	**92.66**	**92.25**	**92.94**	**93.19**

Table 3: Our approach achieves the superior or comparable performance over the existing models on the BUCC benchmark. Note that LASER2 and LASER3 share the same model for the tested languages. We mark the best two scores in bold.

achieves strong performance on the bitext mining task. It is worth noting that all models perform similarly on the BUCC benchmark since the tested languages are all high resource languages. Our model however covers much more languages within a single model than LASER2 and LaBSE.

5.3 Analysis

Method	D	H	Tatoeba	Flores-200	
			↔ en	↔ en	↔ zh
Phase 1	512	8	78.89	95.30	94.38
Phase 2	512	8	82.69	96.25	94.76
Phase 1	768	12	80.76	96.36	95.33
Phase 2	768	12	83.96	97.37	95.95
Phase 1	1024	16	81.16	96.21	95.06
Phase 2	1024	16	84.25	97.29	96.02

Table 4: The averaged bidirectional similarity search accuracy according to different training stages and model architectures. D and H denote the sentence embedding dimension and the number of attention heads. Phase 1 denotes the multilingual NMT pretraining, and Phase 2 denotes the CrossConST finetuning.

We here investigate the impact of the cross-lingual consistency regularization and the model architectures on learning MuSR. We keep the training configurations the same except for the sentence embedding dimension and the number of attention heads. The experimental results on multilingual similarity search are summarized in Table 4. By checking model performance under different combinations of training stage and architecture, we have the following observations: 1) The sentence representation model with multilingual NMT pretraining could achieve decent performance for non-English alignment, and CrossConST finetuning further boosts the model performance especially for English alignment. 2) The model performance consistently improves with the increasing of the sentence embedding dimension and the number of attention heads, while the models with 768 and 1024 embedding dimensions perform similarly, which is in line with Feng et al. (2022). Considering the computationally-heavy inference introduced by 655M parameters of the 1024-dim model, we choose 768 as the sentence embedding dimension.

6 Conclusion

In this paper, we propose MuSR: a one-for-all multilingual sentence representation model supporting 223 languages. Experimental results show that MuSR could yield strong performance on various bitext retrieval and mining tasks compare with the SOTA models LaBSE and LASER3, while also providing increased language coverage in a single model. Extensive analysis shows that CrossConST and the sentence embedding dimension play the key roles in learning multilingual sentence representations. As for future work, we could explore the development of lightweight models by distilling knowledge from MuSR for multilingual sentence alignment, which would potentially lower the computational requirements and make the model more accessible for a variety of applications.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. arXiv preprint arXiv:1907.05019.

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.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. Language resources and evaluation, 49:375–395.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5960–5969, Online. Association for Computational Linguistics.

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

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. Improving zero-shot multilingual neural machine translation by leveraging cross-lingual consistency regularization. In Findings of the Association for Computational Linguistics: ACL 2023, pages 12103–12119, Toronto, Canada. Association for Computational Linguistics.

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.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Zhuoyuan Mao and Tetsuji Nakagawa. 2023. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1886–1894, Dubrovnik, Croatia. 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.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. 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, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In Proceedings of the 59th Annual Meeting of the

Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA).

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, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 87–94, Online. Association for Computational Linguistics.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yunhsuan 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-19, pages 5370–5378. International Joint Conferences on Artificial Intelligence Organization.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In Proceedings of 11th workshop on building and using comparable corpora, pages 39–42.

Appendix

Language	Language	Language	Language
Acehnese (Arabic script)	Georgian	Mossi	Tsonga
Acehnese (Latin script)	German	Najdi Arabic	Tswana
Afrikaans	Greek	Nepali	Tumbuka
Akan	Guarani	Nigerian Fulfulde	Tunisian Arabic
Algerian Arabic	Gujarati	North Azerbaijani	Turkish
Amharic	Haitian Creole	North Levantine Arabic	Turkmen
Armenian	Halh Mongolian	Northern Kurdish	Twi
Assamese	Hausa	Northern Sotho	Ukrainian
Asturian	Hebrew	Northern Uzbek	Umbundu
Awadhi	Hindi	Norwegian Bokmål	Upper Sorbian
Ayacucho Quechua	Hungarian	Norwegian Nynorsk	Urdu
Balinese	Icelandic	Nuer	Uyghur
Bambara	Ido	Nyanja	Venetian
Banjar (Arabic script)	Igbo	Occitan	Vietnamese
Banjar (Latin script)	Ilocano / Iloko	Odia	Walloon
Bashkir	Indonesian	Pangasinan	Waray
Basque	Interlingua	Papiamento	Welsh
Belarusian	Interlingue	Plateau Malagasy	West Central Oromo
Bemba	Irish	Polish	Western Frisian
Bengali	Italian	Portuguese	Western Persian
Berber languages	Japanese	Romanian	Wolof
Bhojpuri	Javanese	Rundi	Xhosa
Bosnian	Jingpho	Russian	Yoruba
Breton	Kabiyè	Samoan	Yue Chinese
Buginese	Kabuverdianu	Sango	Zulu
Bulgarian	Kabyle	Sanskrit	
Burmese	Kamba	Santali	
Catalan	Kannada	Sardinian	
Cebuano	Kashmiri (Arabic script)	Scottish Gaelic	
Central Atlas Tamazight	Kashmiri (Devanagari script)	Serbian	
Central Aymara	Kashubian	Serbo-Croatian	
Central Kanuri (Arabic script)	Kazakh	Shan	
Central Kanuri (Latin script)	Khmer	Shanghainese	
Central Kurdish	Kikongo	Shona	
Chamorro	Kikuyu	Sicilian	
Chhattisgarhi	Kimbundu	Silesian	
Chinese (Simplified)	Kinyarwanda	Sindhi	
Chinese (Traditional)	Korean	Sinhala	
Chokwe	Kyrgyz	Slovak	
Chuvash	Lao	Slovenian	
Cornish	Latgalian	Somali	
Crimean Tatar	Latin	South Azerbaijani	
Croatian	Ligurian	South Levantine Arabic	
Czech	Limburgish	Southern Pashto	
Danish	Lingala	Southern Sotho	
Dari	Lingua Franca Nova	Southwestern Dinka	
Divehi	Lithuanian	Spanish	
Dutch	Lojban	Standard Latvian	
Dyula	Lombard	Standard Malay	
Dzongkha	Low German	Standard Tibetan	
Eastern Panjabi	Luba-Kasai	Sundanese	
Eastern Yiddish	Luo	Swahili	
Egyptian Arabic	Luxembourgish	Swati	
English	Macedonian	Swedish	
Esperanto	Magahi	Tagalog	
Estonian	Maithili	Tajik	
Ewe	Malayalam	Tamasheq (Latin script)	
Faroese	Maltese	Tamasheq (Tifinagh script)	
Fijian	Maori	Tamil	
Filipino	Marathi	Tatar	
Finnish	Meitei (Bengali script)	Ta'izzi-Adeni Arabic	
Fon	Mesopotamian Arabic	Telugu	
French	Minangkabau (Latin script)	Thai	
Friulian	Mizo	Tigrinya	
Galician	Modern Standard Arabic	Tok Pisin	
Ganda	Moroccan Arabic	Tosk Albanian	

Table 5: The supported languages of MuSR.

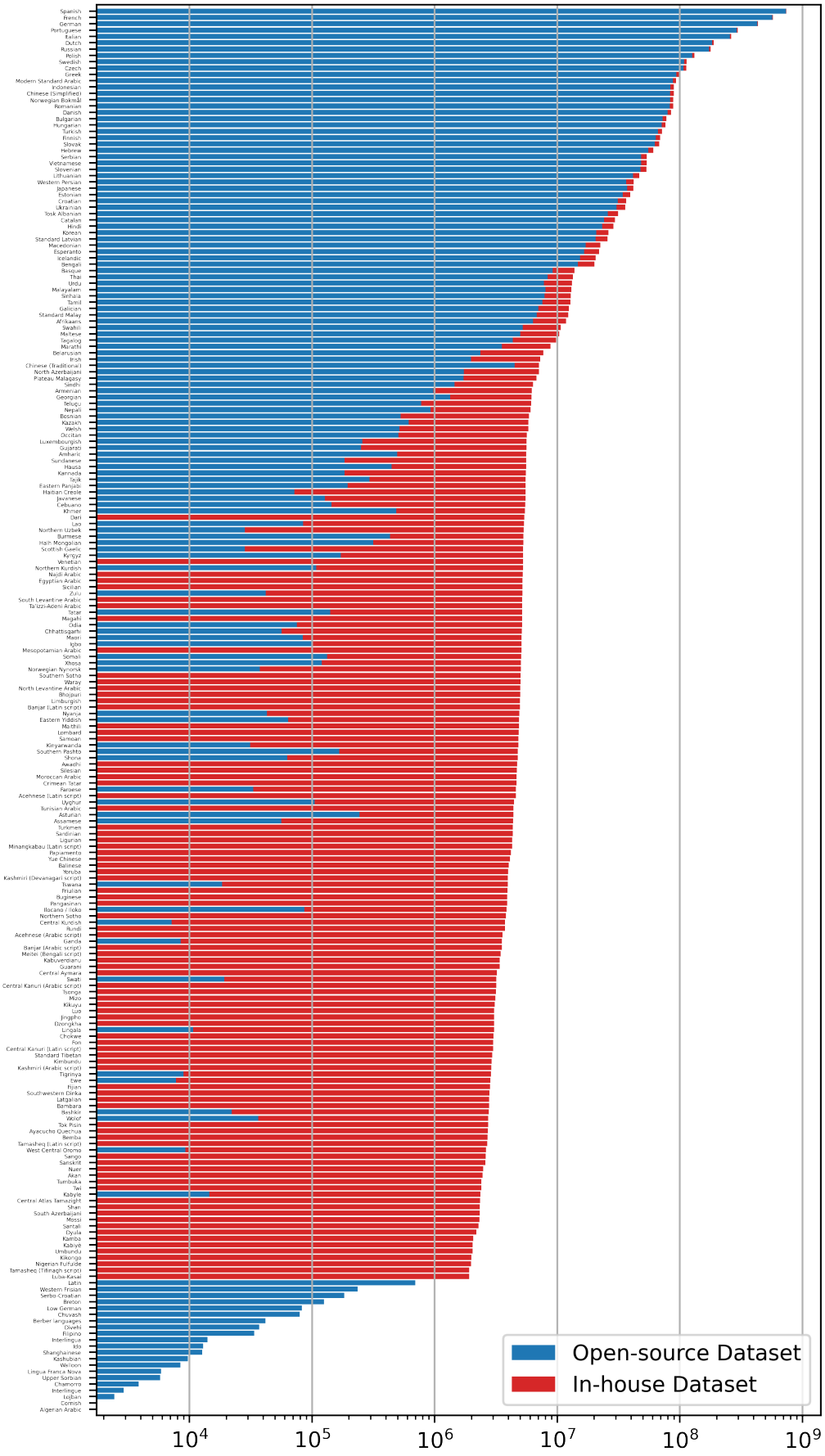

Figure 6: The distribution of the open-source and in-house cleaned datasets for each language in our training dataset.

Language	LASER2	LASER3	LaBSE	MuSR	Language	LASER2	LASER3	LaBSE	MuSR
afr	93.2	-	**97.4**	95.85	kaz	55.83	80.61	**90.52**	87.48
amh	80.06	86.31	**94.05**	88.39	khm	77.49	53.32	**83.17**	77.35
ang	37.31	-	**64.55**	57.84	kor	91.35	-	**93.5**	89.9
ara	**92.25**	-	90.85	90.45	kur	23.41	-	**87.2**	78.54
arq	33.04	-	46.16	**65.59**	kzj	8.65	-	**14.25**	13.95
arz	70.02	-	78.41	**82.39**	lat	68.9	-	**81.9**	70.5
ast	80.71	-	**90.55**	90.16	lfn	67.85	-	71.25	**84.9**
awa	39.39	80.74	73.16	**85.93**	lit	96.95	-	**97.3**	95.8
aze	81.65	91.5	**96.1**	92.95	lvs	96.6	-	**96.8**	94.7
bel	83.4	94.05	**96.15**	95.05	mal	98.4	97.82	**98.91**	97.67
ben	91.3	90.1	**91.35**	89.4	mar	**94.75**	91.1	94.7	94.5
ber	**81.75**	-	10.5	74.7	max	45.42	-	**71.13**	66.02
bos	**96.89**	-	96.33	96.75	mhr	10	-	**19.5**	12.3
bre	**36.6**	-	17.35	21.65	mkd	**95.1**	-	94.85	94.65
bul	95.15	-	**95.7**	95.05	mon	7.27	87.73	**96.48**	88.52
cat	96.55	-	**96.6**	96.25	nds	80.2	-	81.35	**88.75**
cbk	79.75	-	**82.4**	77.2	nld	96.35	-	**97.25**	96.45
ceb	15.92	**80**	71	62.17	nno	77.25	-	95.85	**96**
ces	96.85	-	**97.5**	96.25	nob	95.6	-	**98.9**	98.5
cha	26.64	-	39.05	**44.53**	nov	67.51	-	78.21	**85.02**
cmn	84.3	-	**96.2**	94.85	oci	63.35	-	69.75	**76.85**
cor	7.2	-	12.75	**24.95**	orv	30.24	-	**47.07**	44.01
csb	38.34	-	56.13	**66.21**	pam	5.5	-	**13.55**	13.2
cym	9.74	89.04	**93.65**	87.22	pes	92.9	93.4	**96.05**	94.45
dan	95.9	-	**96.45**	96.25	pms	45.14	-	66.95	**86.67**
deu	99.3	-	**99.35**	98.95	pol	**98**	-	97.85	97.85
dsb	51.25	-	**69.31**	69	por	**95.75**	-	95.55	95.4
dtp	11.5	-	13.35	**21.8**	ron	97.25	-	**97.85**	97.45
ell	**96.85**	-	96.6	96.55	rus	94.35	-	**95.3**	95
epo	97.45	-	**98.35**	97.65	slk	96.6	-	**97.3**	96.55
est	97	-	**97.7**	96.45	slv	**96.78**	-	96.72	95.63
eus	93.85	-	**95.75**	94	spa	97.9	-	**98.45**	97.75
fao	64.12	73.66	90.46	**93.32**	sqi	**97.85**	**97.85**	97.65	97.05
fin	**97.3**	-	97.05	95.85	srp	95.05	-	**96.2**	95.9
fra	95.5	-	**96.05**	95.6	swe	95.85	-	**96.55**	96.45
fry	51.45	-	**90.17**	71.97	swg	45.09	-	**65.18**	**65.18**
gla	3.32	70.27	**88.9**	82.51	swh	57.69	81.41	**88.46**	80.13
gle	9.15	78.55	**95**	88.75	tam	85.99	58.79	**90.72**	85.18
glg	96.75	-	**97.25**	95.5	tat	30.7	64.7	**87.9**	86.5
gsw	36.32	-	52.56	**66.67**	tel	97.01	80.56	**98.29**	92.31
heb	91.75	-	**92.95**	91.85	tgl	68.85	95	**97.45**	91.6
hin	96.1	95.55	**97.75**	97.05	tha	96.99	96.53	**97.08**	95.71
hrv	97.45	-	**97.8**	97.5	tuk	22.17	58.37	80.05	**86.45**
hsb	54.04	-	71.12	**80.43**	tur	98.15	97.2	**98.35**	97.85
hun	96.1	-	**97.2**	96.15	tzl	41.35	-	**62.98**	57.69
hye	90.03	90.63	**95.01**	92.18	uig	51.45	76.3	**93.7**	89.3
ido	84.1	-	90.8	**94.5**	ukr	95.05	-	**95.25**	95.1
ile	88.85	-	87.05	**95.85**	urd	82.6	89.85	**95.35**	92.55
ina	95.5	-	95.85	**96.75**	uzb	26.4	78.39	**86.8**	74.65
ind	94.8	94.75	**95.3**	94.75	vie	97.15	-	**97.85**	96.55
isl	95.8	-	96.15	**96.25**	war	13.35	**75.35**	65.4	70.1
ita	**95.55**	-	94.65	95.25	wuu	79.4	-	**90.3**	89.45
jav	18.78	**86.34**	84.39	81.22	xho	5.63	**93.66**	91.9	91.2
jpn	96	-	**96.45**	94.35	yid	5.19	**94.16**	90.98	89.86
kab	71.45	**89.65**	6	72.55	yue	87.65	-	**92.1**	86.35
kat	81.97	75	**95.91**	93.43	zsm	96.25	96.1	**96.9**	95.85

Table 6: The averaged bidirectional similarity search accuracy (xx $\leftrightarrow$ en) on the Tatoeba benchmark.

Language	LASER2	LASER3	LaBSE	MuSR	Language	LASER2	LASER3	LaBSE	MuSR
ace_Arab	7.11	-	35.82	**83.84**	gaz_Latn	9.93	96.94	46.99	**99.01**
ace_Latn	38.24	96.89	88.74	**99.6**	gla_Latn	7.02	91.65	**99.9**	99.65
acm_Arab	99.51	-	**100**	99.9	gle_Latn	7.02	97.38	**100**	99.7
acq_Arab	99.85	-	**100**	**100**	glg_Latn	99.95	-	**100**	99.95
aeb_Arab	98.67	-	99.41	**99.65**	grn_Latn	33.65	98.91	77.77	**99.31**
afr_Latn	99.75	-	**100**	99.95	guj_Gujr	3.11	99.65	**100**	99.95
ajp_Arab	99.7	-	**99.95**	**99.95**	hat_Latn	32.71	98.57	**99.31**	99.21
aka_Latn	21.49	98.47	68.77	**99.06**	hau_Latn	22.78	98.96	**99.7**	99.56
als_Latn	99.7	-	**100**	**100**	heb_Hebr	99.95	-	**100**	**100**
amh_Ethi	54.5	99.75	**100**	99.9	hin_Deva	98.96	99.9	**100**	99.85
apc_Arab	99.7	-	**100**	99.95	hne_Deva	92.49	97.63	**99.51**	**99.51**
arb_Arab	99.95	-	**100**	**100**	hrv_Latn	99.9	-	**100**	99.95
arb_Latn	7.46	-	**41.16**	35.52	hun_Latn	99.95	-	**100**	**100**
ars_Arab	99.95	-	**100**	**100**	hye_Armn	89.23	99.65	**100**	99.85
ary_Arab	91.75	-	97.63	**98.81**	ibo_Latn	17.64	99.41	**100**	99.65
arz_Arab	99.46	-	**99.95**	99.85	ilo_Latn	41.25	99.85	89.87	**100**
asm_Beng	53.85	95.65	**99.9**	99.75	ind_Latn	98.96	99.9	**100**	**100**
ast_Latn	99.21	-	99.95	**100**	isl_Latn	99.41	-	**99.9**	99.75
awa_Deva	96.89	96.2	**99.06**	99.01	ita_Latn	99.95	-	**100**	99.9
ayr_Latn	13.88	82.91	51.63	**94.47**	jav_Latn	57.31	99.9	**100**	99.95
azb_Arab	43.28	64.23	85.62	**93.82**	jpn_Jpan	**100**	-	**100**	99.7
azj_Latn	50.99	99.06	**99.85**	98.67	kab_Latn	85.52	97.28	45.26	**99.26**
bak_Cyrl	13.98	98.32	90.12	**99.7**	kac_Latn	11.76	92.93	55.04	**98.22**
bam_Latn	17.34	92.89	54.99	**96.49**	kam_Latn	28.51	83.7	67.84	**86.91**
ban_Latn	53.46	99.21	98.27	**99.41**	kan_Knda	2.87	99.31	**100**	99.7
bel_Cyrl	74.31	99.16	**100**	99.11	kas_Arab	34.29	98.81	90.86	**99.01**
bem_Latn	31.03	99.46	83.15	**99.6**	kas_Deva	29.84	**95.8**	81.23	95.06
ben_Beng	99.9	99.01	**100**	99.85	kat_Geor	79.79	97.68	**99.95**	99.36
bho_Deva	87.06	98.07	**99.85**	99.7	kaz_Cyrl	51.63	98.86	**99.8**	99.56
bjn_Arab	7.31	-	32.91	**83.55**	kbp_Latn	12.99	88.09	52.22	**93.82**
bjn_Latn	78.51	**99.8**	98.37	**99.8**	kea_Latn	81.67	98.27	97.83	**100**
bod_Tibt	2.12	81.03	**98.96**	97.48	khk_Cyrl	12.15	98.62	**100**	99.51
bos_Latn	**100**	-	**100**	99.9	khm_Khmr	79.99	96.39	97.92	**99.95**
bug_Latn	34.44	97.58	81.82	**97.97**	kik_Latn	9.73	**98.62**	68.53	**98.62**
bul_Cyrl	99.95	-	**100**	99.75	kin_Latn	19.61	99.31	**99.75**	**99.75**
cat_Latn	**100**	-	**100**	**100**	kir_Cyrl	27.92	96.99	**99.95**	99.11
ceb_Latn	61.41	99.8	**100**	**100**	kmb_Latn	28.11	90.61	60.87	**93.58**
ces_Latn	99.9	-	**100**	99.9	kmr_Latn	18.68	97.58	**99.9**	99.51
cjk_Latn	28.16	74.26	61.61	**82.31**	knc_Arab	9.29	**36.22**	22.68	21.99
ckb_Arab	4.64	99.75	44.86	**99.95**	knc_Latn	16.95	92.59	58.1	**93.13**
crh_Latn	76.88	99.7	**99.85**	99.7	kon_Latn	39.38	97.63	71.34	**99.26**
cym_Latn	18.03	99.16	**100**	**100**	kor_Hang	99.56	-	**99.95**	99.8
dan_Latn	**100**	-	**100**	99.85	lao_Laoo	9.39	94.81	96.94	**100**
deu_Latn	**100**	-	**100**	99.95	lij_Latn	88.88	**99.85**	98.86	**99.85**
dik_Latn	21.44	74.11	57.71	**82.21**	lim_Latn	83.1	85.23	98.72	**99.75**
dyu_Latn	13.39	**75.89**	47.73	70.06	lin_Latn	34.19	99.56	72.58	**99.7**
dzo_Tibt	0.25	92.54	92.54	**98.37**	lit_Latn	99.56	-	**99.6**	99.46
ell_Grek	99.9	-	**100**	**100**	lmo_Latn	78.9	98.22	97.48	**99.7**
eng_Latn	-	-	-	-	ltg_Latn	78.26	99.65	95.5	**99.85**
epo_Latn	**100**	-	**100**	**100**	ltz_Latn	66.65	99.01	**100**	99.95
est_Latn	99.85	-	**100**	99.85	lua_Latn	34.73	96.89	70.95	**97.63**
eus_Latn	99.8	-	99.95	**100**	lug_Latn	22.28	97.08	80.88	**98.67**
ewe_Latn	10.67	96.15	56.47	**96.54**	luo_Latn	16.21	98.76	59.19	**99.6**
fao_Latn	88.09	96.29	**99.95**	**99.95**	lus_Latn	16.7	95.06	71.29	**97.97**
fij_Latn	22.08	98.57	59.58	**99.41**	lvs_Latn	99.9	-	**100**	99.75
fin_Latn	99.85	-	**99.9**	99.6	mag_Deva	96.1	99.46	**100**	99.75
fon_Latn	10.38	81.08	47.88	**84.63**	mai_Deva	88.19	95.6	**100**	**100**
fra_Latn	99.95	-	**100**	**100**	mal_Mlym	99.06	99.51	**99.9**	99.46
fur_Latn	86.17	99.9	98.96	**100**	mar_Deva	98.91	98.52	**100**	99.9
fuv_Latn	17.14	66.06	63.14	**79.35**	min_Arab	4.99	-	30.63	**82.46**

Table 7: The averaged bidirectional similarity search accuracy (xx $\leftrightarrow$ en) on the Flores-200 benchmark (Part I).

Language	LASER2	LASER3	LaBSE	MuSR	Language	LASER2	LASER3	LaBSE	MuSR
min_Latn	61.46	99.56	97.13	**99.9**	spa_Latn	99.6	-	**99.9**	99.51
mkd_Cyrl	**100**	-	**100**	99.95	srd_Latn	89.08	99.9	99.16	**100**
mlt_Latn	25.4	99.9	**100**	**100**	srp_Cyrl	99.9	-	**100**	99.9
mni_Beng	8.4	98.27	36.81	**99.26**	ssw_Latn	17	99.36	96.34	**99.6**
mos_Latn	17.39	81.97	54.35	**86.31**	sun_Latn	61.02	99.41	99.8	**99.9**
mri_Latn	18.97	97.88	**99.51**	99.36	swe_Latn	**100**	-	**100**	**100**
mya_Mymr	83.65	98.22	**99.7**	99.36	swh_Latn	98.72	99.21	**100**	**100**
nld_Latn	99.7	-	**100**	99.51	szl_Latn	94.86	**99.21**	98.86	**99.21**
nno_Latn	98.86	-	**99.9**	**99.9**	tam_Taml	82.07	99.56	**100**	99.41
nob_Latn	99.6	-	**99.9**	99.75	taq_Latn	38.09	72.68	55.58	**76.19**
npi_Deva	68.63	97.63	**99.7**	99.41	taq_Tfng	2.08	-	16.45	**61.17**
nso_Latn	22.73	99.7	99.06	**99.9**	tat_Cyrl	21	95.7	**100**	99.8
nus_Latn	8.6	90.27	43.03	**96.79**	tel_Telu	96.54	99.01	**100**	99.7
nya_Latn	31.52	99.41	99.6	**99.8**	tgk_Cyrl	6.92	98.86	**99.75**	99.7
oci_Latn	99.6	-	99.95	**100**	tgl_Latn	90.22	99.95	**100**	**100**
ory_Orya	3.41	99.51	**100**	99.46	tha_Thai	99.56	**99.75**	94.02	**99.75**
pag_Latn	46.84	98.52	87.85	**99.16**	tir_Ethi	5.53	**98.72**	75.94	98.52
pan_Guru	3.06	99.65	**100**	99.9	tpi_Latn	30.39	99.75	83.05	**100**
pap_Latn	78.36	99.8	98.47	**100**	tsn_Latn	17.19	98.47	97.97	**98.76**
pbt_Arab	29.99	99.41	**100**	99.7	tso_Latn	22.04	98.91	71.29	**99.36**
pes_Arab	98.81	98.47	**100**	99.75	tuk_Latn	29.94	92.54	**99.95**	99.75
plt_Latn	99.9	99.85	**99.95**	**99.95**	tum_Latn	27.12	97.78	90.46	**99.06**
pol_Latn	99.85	-	**100**	99.6	tur_Latn	99.06	99.16	**100**	99.9
por_Latn	99.95	-	**100**	**100**	twi_Latn	25.44	98.96	71.79	**99.06**
prs_Arab	98.12	97.48	**100**	99.75	tzm_Tfng	1.73	95.45	16.3	**97.38**
quy_Latn	19.76	71.79	57.71	**93.63**	uig_Arab	17.14	91.75	**99.8**	99.51
ron_Latn	99.95	-	**100**	**100**	ukr_Cyrl	99.95	-	**100**	99.95
run_Latn	19.12	99.26	**99.51**	99.46	umb_Latn	19.96	83.79	58.2	**87.15**
rus_Cyrl	99.85	-	**100**	99.95	urd_Arab	89.28	99.46	**99.9**	99.56
sag_Latn	25.2	89.33	62.7	**94.86**	uzn_Latn	19.12	99.6	**99.9**	99.51
san_Deva	49.65	83.4	96.44	**98.57**	vec_Latn	94.32	97.18	99.8	**99.95**
sat_Olck	0.3	-	4.15	**95.41**	vie_Latn	99.9	-	**100**	99.9
scn_Latn	76.63	99.26	98.42	**99.85**	war_Latn	55.43	99.9	99.95	**100**
shn_Mymr	16.25	98.52	48.37	**99.51**	wol_Latn	25	89.77	68.48	**95.7**
sin_Sinh	99.65	99.16	**100**	99.26	xho_Latn	18.33	**99.8**	99.7	**99.8**
slk_Latn	99.85	-	**100**	99.75	ydd_Hebr	11.91	95.41	99.95	**100**
slv_Latn	99.85	-	**100**	99.8	yor_Latn	21.25	95.06	**97.43**	97.18
smo_Latn	18.82	99.7	99.56	**99.85**	yue_Hant	93.53	-	**100**	99.85
sna_Latn	19.52	99.46	99.26	**99.65**	zho_Hans	99.56	-	**100**	99.6
snd_Arab	24.51	97.58	**100**	99.7	zho_Hant	94.02	-	**99.95**	99.46
som_Latn	8.55	98.07	99.65	**99.7**	zsm_Latn	99.11	99.9	**100**	**100**
sot_Latn	20.85	99.8	99.9	**100**	zul_Latn	13.19	99.85	99.85	**99.9**

Table 8: The averaged bidirectional similarity search accuracy (xx $\leftrightarrow$ en) on the Flores-200 benchmark (Part II).

Language	LASER2	LASER3	LaBSE	MuSR	Language	LASER2	LASER3	LaBSE	MuSR
ace_Arab	6.27	-	29.2	**73.57**	gaz_Latn	7.51	92.59	40.96	**97.88**
ace_Latn	29.69	91.4	81.92	**97.68**	gla_Latn	4.84	81.27	**99.85**	98.76
acm_Arab	98.52	-	**99.9**	99.56	gle_Latn	5.09	92.64	**99.95**	98.86
acq_Arab	98.76	-	**99.95**	99.7	glg_Latn	99.56	-	**100**	99.65
aeb_Arab	96.99	-	**98.86**	**98.86**	grn_Latn	26.53	96.25	71.49	**97.38**
afr_Latn	97.83	-	**100**	99.46	guj_Gujr	2.57	98.81	**100**	99.65
ajp_Arab	98.52	-	**99.75**	99.56	hat_Latn	24.31	96.25	**99.21**	98.42
aka_Latn	16.55	94.52	58.89	**96.59**	hau_Latn	16.11	97.13	**99.11**	99.01
als_Latn	98.91	-	**100**	99.21	heb_Hebr	99.21	-	**100**	99.51
amh_Ethi	47.48	99.01	**99.9**	99.65	hin_Deva	97.83	99.51	**99.95**	99.6
apc_Arab	98.47	-	**99.7**	**99.7**	hne_Deva	87.15	96.64	98.91	**99.11**
arb_Arab	99.56	-	**100**	99.7	hrv_Latn	99.31	-	**99.95**	99.56
arb_Latn	5.78	-	**36.51**	31.72	hun_Latn	99.51	-	**100**	99.8
ars_Arab	99.51	-	**100**	99.6	hye_Armn	77.72	98.52	**100**	99.6
ary_Arab	87.5	-	96.1	**97.48**	ibo_Latn	13.29	96.94	**99.01**	98.42
arz_Arab	98.17	-	**99.7**	99.31	ilo_Latn	30.93	99.16	81.82	**99.41**
asm_Beng	49.31	91.5	**99.51**	99.11	ind_Latn	98.22	99.31	**100**	99.65
ast_Latn	95.06	-	**99.75**	98.91	isl_Latn	97.48	-	**99.85**	99.11
awa_Deva	93.97	91.9	**99.06**	98.86	ita_Latn	99.65	-	**100**	99.8
ayr_Latn	11.26	75.59	46.25	**92.54**	jav_Latn	45.36	98.02	**100**	99.46
azb_Arab	41.01	55.34	81.57	**92.59**	jpn_Jpan	99.21	-	**100**	99.41
azj_Latn	49.06	97.78	**99.6**	98.57	kab_Latn	70.75	89.97	37.2	**95.8**
bak_Cyrl	12.35	96.15	84.73	**99.56**	kac_Latn	10.03	86.51	48.62	**95.9**
bam_Latn	13.24	87.25	48.27	**92**	kam_Latn	21.74	72.92	58.79	**79.79**
ban_Latn	46.25	97.48	95.9	**98.42**	kan_Knda	1.88	97.53	**100**	99.46
bel_Cyrl	67.98	97.53	**100**	98.62	kas_Arab	31.42	97.08	86.46	**98.17**
bem_Latn	24.85	96.99	72.92	**97.78**	kas_Deva	25.84	89.67	72.38	**92.93**
ben_Beng	99.21	97.38	**99.95**	99.6	kat_Geor	70.01	94.91	**100**	99.06
bho_Deva	82.02	96.25	98.72	**99.36**	kaz_Cyrl	47.08	97.33	**99.8**	99.31
bjn_Arab	6.08	-	24.26	**74.7**	kbp_Latn	9.88	83.35	45.31	**90.91**
bjn_Latn	69.12	98.22	96.64	**98.81**	kea_Latn	64.62	92.69	93.53	**99.21**
bod_Tibt	2.42	76.33	**98.07**	96.84	khk_Cyrl	11.46	95.95	**100**	99.46
bos_Latn	99.7	-	**100**	99.51	khm_Khmr	69.07	88.24	97.83	**99.31**
bug_Latn	26.38	92.34	76.53	**94.96**	kik_Latn	8.05	95.36	57.56	**96.54**
bul_Cyrl	99.36	-	**100**	99.6	kin_Latn	15.02	98.32	**99.56**	99.21
cat_Latn	99.51	-	**100**	99.51	kir_Cyrl	26.73	93.82	**99.8**	98.96
ceb_Latn	46.74	98.52	**99.95**	99.56	kmb_Latn	20.8	80.29	51.43	**84.78**
ces_Latn	99.6	-	**100**	99.8	kmr_Latn	14.87	92.98	**99.65**	98.96
cjk_Latn	21.15	62.06	53.26	**73.22**	knc_Arab	7.41	**29.74**	20.11	17.59
ckb_Arab	3.51	98.86	37.35	**99.16**	knc_Latn	12.75	83.3	50.49	**88.29**
crh_Latn	71.25	98.57	99.21	**99.56**	kon_Latn	31.82	94.86	61.71	**98.07**
cym_Latn	12.99	96.15	**100**	99.65	kor_Hang	98.67	-	**99.9**	99.65
dan_Latn	99.56	-	**100**	99.46	lao_Laoo	7.81	88.59	96.59	**99.6**
deu_Latn	99.6	-	**100**	99.7	lij_Latn	73.96	98.62	95.45	**99.41**
dik_Latn	15.22	61.91	50.15	**73.07**	lim_Latn	70.06	71.1	96.59	**98.52**
dyu_Latn	9.83	**65.51**	41.21	62.3	lin_Latn	28.61	97.68	61.81	**98.47**
dzo_Tibt	0.3	88.54	89.03	**97.08**	lit_Latn	99.21	-	**99.51**	99.26
ell_Grek	99.36	-	**100**	99.7	lmo_Latn	60.67	93.28	92.59	**97.92**
eng_Latn	99.56	-	**100**	99.6	ltg_Latn	66.35	98.67	91.35	**99.11**
epo_Latn	99.26	-	**100**	99.56	ltz_Latn	51.14	94.37	**99.85**	99.7
est_Latn	99.41	-	**99.95**	99.8	lua_Latn	26.88	90.46	62.06	**93.58**
eus_Latn	98.12	-	**99.95**	99.7	lug_Latn	15.51	92.05	69.52	**96.1**
ewe_Latn	8.2	93.28	50.59	**94.71**	luo_Latn	11.76	94.47	51.04	**97.68**
fao_Latn	76.53	87.9	**99.75**	99.41	lus_Latn	12.8	88.44	63.64	**95.85**
fij_Latn	15.56	96.15	51.09	**97.78**	lvs_Latn	99.51	-	**99.95**	99.6
fin_Latn	99.36	-	**99.85**	99.51	mag_Deva	91.9	98.62	99.65	**99.75**
fon_Latn	8	73.12	43.38	**79.2**	mai_Deva	81.92	90.46	99.65	**99.8**
fra_Latn	99.6	-	**100**	99.65	mal_Mlym	97.04	98.76	**99.85**	99.21
fur_Latn	72.83	98.37	96.39	**99.51**	mar_Deva	96.29	96.39	**99.9**	99.6
fuv_Latn	11.91	55.58	55.88	**71.15**	min_Arab	3.75	-	23.67	**72.78**

Table 9: The averaged bidirectional similarity search accuracy (xx $\leftrightarrow$ zh) on the Flores-200 benchmark (Part I).

Language	LASER2	LASER3	LaBSE	MuSR	Language	LASER2	LASER3	LaBSE	MuSR
min_Latn	50.89	97.88	93.97	**99.31**	spa_Latn	99.36	-	**99.9**	99.11
mkd_Cyrl	99.6	-	**100**	99.85	srd_Latn	72.48	96.34	96.54	**99.21**
mlt_Latn	18.92	98.72	**100**	99.56	srp_Cyrl	98.62	-	**100**	99.7
mni_Beng	7.36	93.82	30.63	**98.52**	ssw_Latn	11.86	98.22	90.56	**98.57**
mos_Latn	13.39	72.73	47.68	**79.79**	sun_Latn	51.28	97.48	**99.7**	99.11
mri_Latn	14.72	94.27	**98.42**	97.58	swe_Latn	99.65	-	**100**	99.6
mya_Mymr	79	96.64	**99.65**	99.26	swh_Latn	95.95	96.05	**99.95**	99.21
nld_Latn	98.96	-	**100**	99.46	szl_Latn	85.08	98.27	97.83	**98.62**
nno_Latn	95.06	-	**99.85**	99.41	tam_Taml	76.28	98.02	**99.95**	98.76
nob_Latn	98.27	-	**99.8**	99.41	taq_Latn	27.72	59.88	49.16	**69.17**
npi_Deva	61.56	94.27	**99.7**	99.06	taq_Tfng	1.68	-	13.69	**53.85**
nso_Latn	17.34	98.52	96.1	**99.01**	tat_Cyrl	16.85	91.6	**100**	99.6
nus_Latn	7.36	79.5	36.26	**92.34**	tel_Telu	90.81	97.58	**100**	99.31
nya_Latn	24.56	97.78	**98.81**	98.72	tgk_Cyrl	4.79	96.89	**99.75**	99.16
oci_Latn	95.6	-	**99.7**	99.56	tgl_Latn	77.37	99.31	**99.9**	99.51
ory_Orya	2.77	99.01	**100**	99.31	tha_Thai	**99.36**	99.21	93.73	99.31
pag_Latn	35.67	96.1	82.91	**97.88**	tir_Ethi	5.93	95.75	68.73	**97.63**
pan_Guru	2.57	98.57	**100**	99.51	tpi_Latn	22.83	94.52	73.57	**99.06**
pap_Latn	63.34	98.96	95.36	**99.65**	tsn_Latn	12.9	96.94	94.81	**97.53**
pbt_Arab	26.73	97.33	**99.36**	99.31	tso_Latn	16.7	97.68	59.14	**98.57**
pes_Arab	97.92	95.85	**100**	99.6	tuk_Latn	26.58	85.72	**99.75**	99.41
plt_Latn	99.41	98.86	**99.56**	99.06	tum_Latn	21.49	95.5	85.47	**97.53**
pol_Latn	99.26	-	**99.95**	99.6	tur_Latn	98.12	97.73	**100**	99.75
por_Latn	99.56	-	**100**	99.51	twi_Latn	17.64	95.55	62.01	**97.08**
prs_Arab	97.28	93.92	**100**	99.7	tzm_Tfng	1.63	87.65	14.33	**92.49**
quy_Latn	14.48	61.76	51.78	**88.64**	uig_Arab	14.08	86.71	**99.85**	99.21
ron_Latn	99.06	-	**100**	99.56	ukr_Cyrl	99.26	-	**100**	99.65
run_Latn	14.97	97.68	98.12	**98.86**	umb_Latn	15.66	75.59	51.73	**79.35**
rus_Cyrl	98.96	-	**100**	99.75	urd_Arab	86.12	98.37	**99.8**	99.46
sag_Latn	19.61	80.78	54.5	**89.58**	uzn_Latn	15.61	98.27	**99.85**	99.21
san_Deva	43.63	78.26	93.08	**97.48**	vec_Latn	85.42	89.87	98.47	**99.51**
sat_Olck	0.25	-	2.62	**91.25**	vie_Latn	99.41	-	**100**	99.51
scn_Latn	61.61	97.04	95.16	**98.86**	war_Latn	39.72	99.16	**99.56**	99.41
shn_Mymr	12.5	95.11	42.29	**98.67**	wol_Latn	18.48	76.93	60.77	**90.46**
sin_Sinh	98.47	97.92	**99.9**	99.06	xho_Latn	12.3	98.76	98.91	**99.11**
slk_Latn	99.41	-	**100**	99.56	ydd_Hebr	9.63	77.77	**99.36**	99.01
slv_Latn	99.26	-	**100**	99.41	yor_Latn	15.07	90.76	93.73	**94.32**
smo_Latn	13.44	98.52	**99.06**	98.62	yue_Hant	93.68	-	**100**	99.85
sna_Latn	13.93	97.58	97.68	**98.72**	zho_Hans	-	-	-	-
snd_Arab	21.34	94.07	**99.7**	99.06	zho_Hant	94.32	-	**99.9**	99.56
som_Latn	6.97	93.68	98.67	**98.76**	zsm_Latn	98.42	99.46	**100**	99.51
sot_Latn	14.48	99.11	98.76	**99.16**	zul_Latn	9.29	99.26	**99.51**	99.31

Table 10: The averaged bidirectional similarity search accuracy (xx $\leftrightarrow$ zh) on the Flores-200 benchmark (Part II).

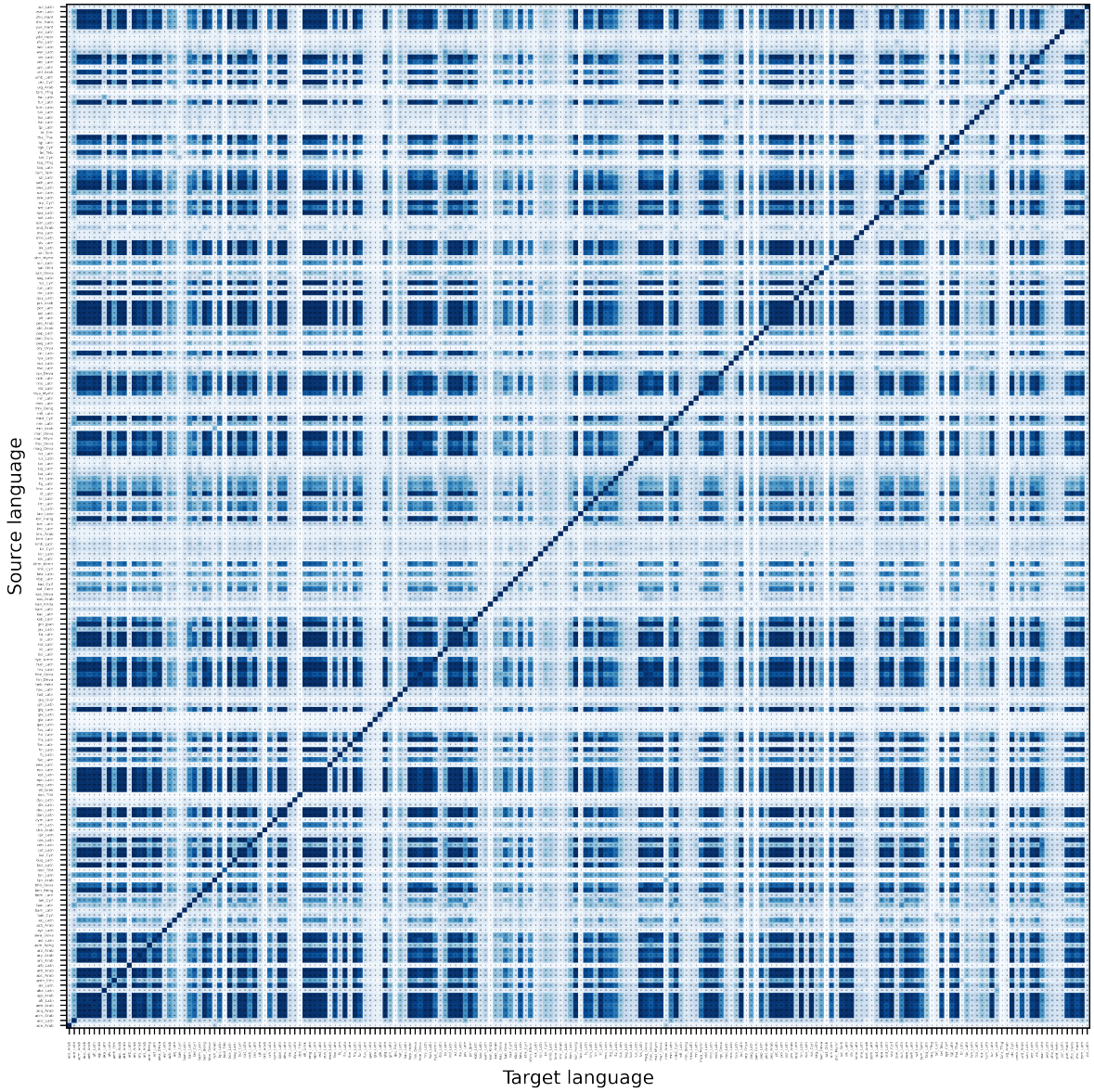

Figure 7: The multilingual similarity search performance of LASER2 on the Flores-200 benchmark.

Figure 8: The multilingual similarity search performance of LaBSE on the Flores-200 benchmark.

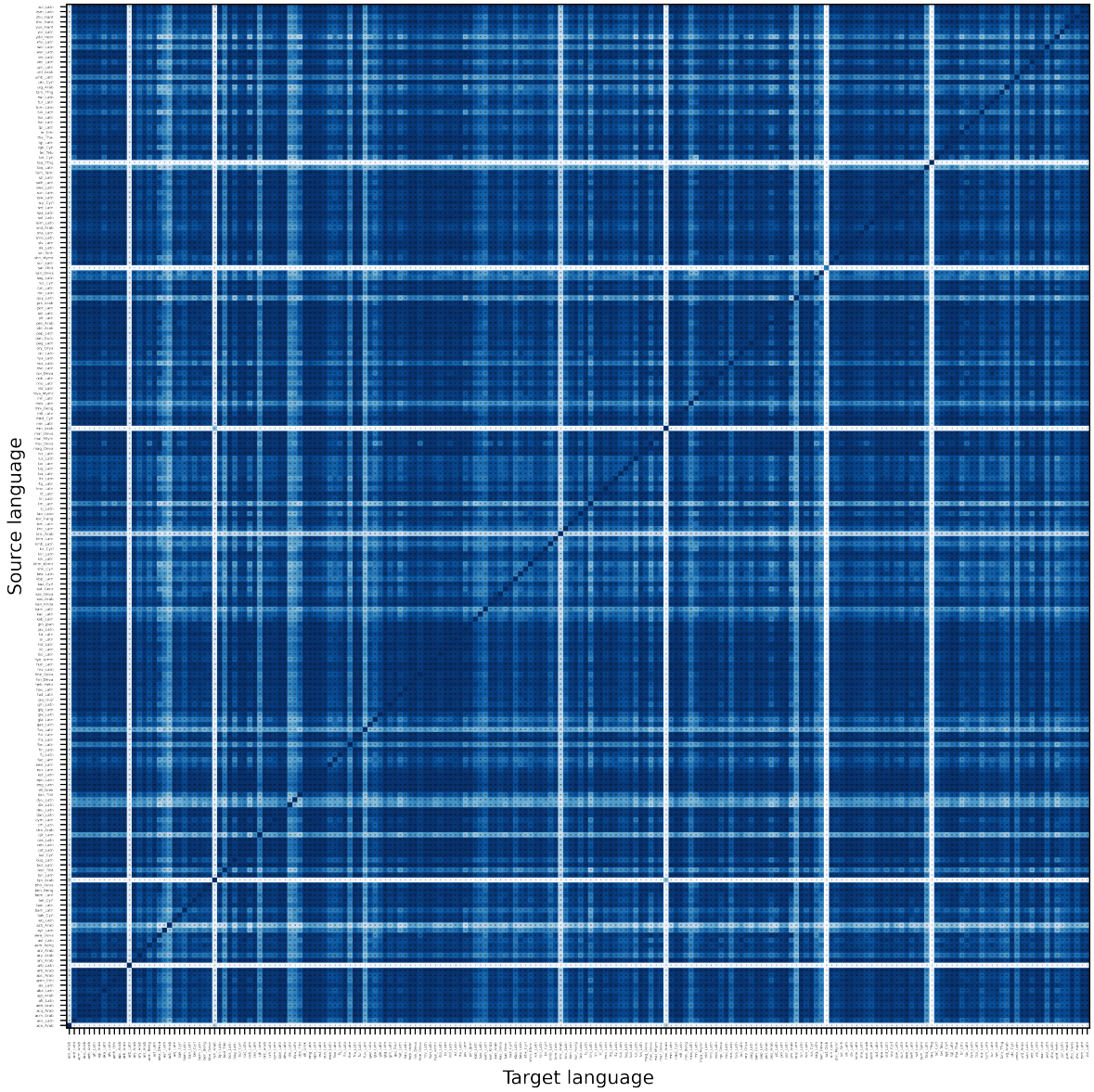

Figure 9: The multilingual similarity search performance of LASER3 on the Flores-200 benchmark.

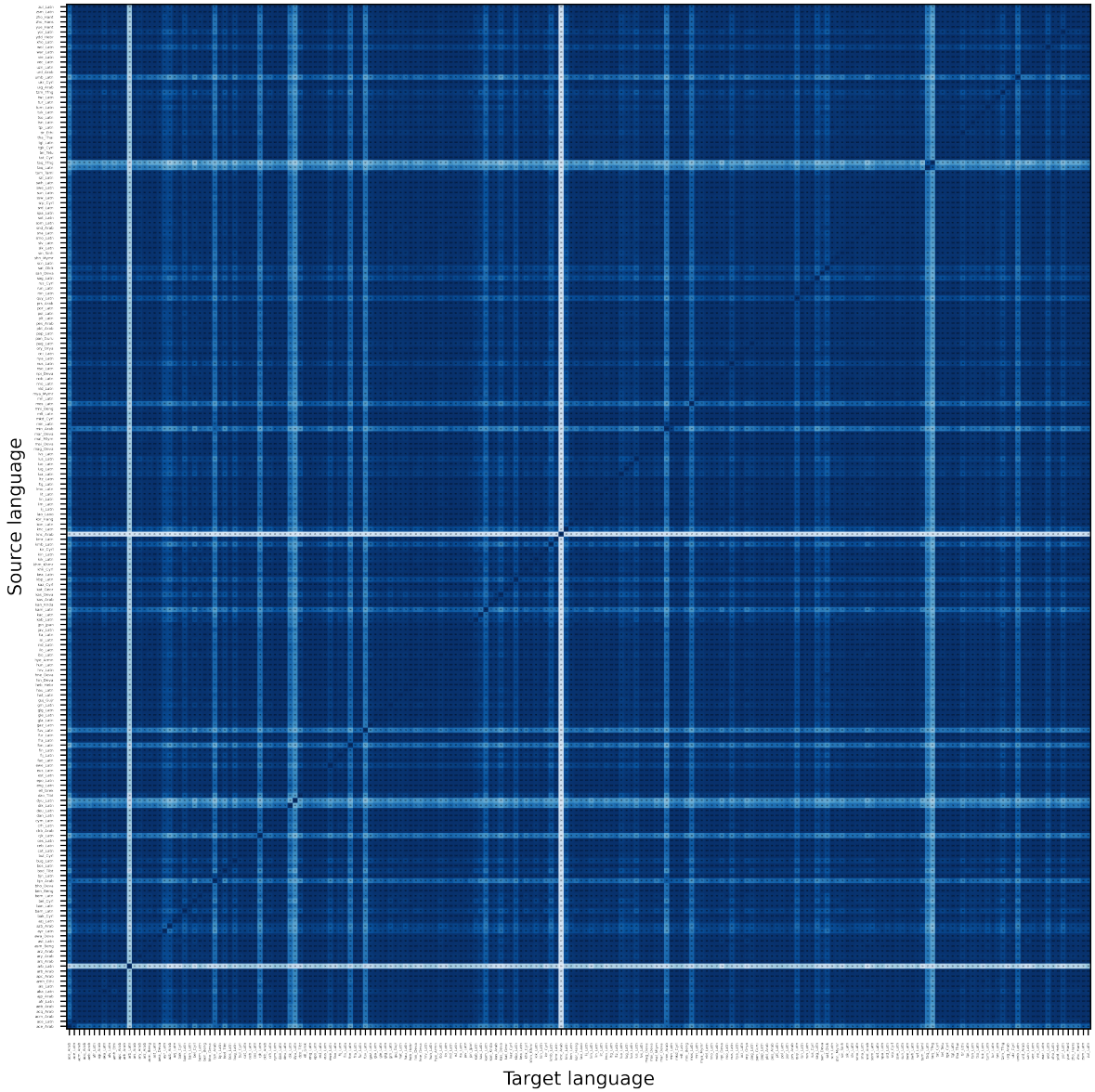

Figure 10: The multilingual similarity search performance of MuSR on the Flores-200 benchmark.