LUXEMBEDDER: A Cross-Lingual Approach to Enhanced Luxembourgish Sentence Embeddings

Fred Philippy[1,2], Siwen Guo[1], Jacques Klein[2], Tegawendé F. Bissyandé[2]

[1]Zortify S.A., Luxembourg [2]University of Luxembourg, Luxembourg

{fred, siwen}@zortify.com
{jacques.klein, tegawende.bissyande}@uni.lu

Abstract

Sentence embedding models play a key role in various Natural Language Processing tasks, such as in Topic Modeling, Document Clustering and Recommendation Systems. However, these models rely heavily on parallel data, which can be scarce for many low-resource languages, including Luxembourgish. This scarcity results in suboptimal performance of monolingual and cross-lingual sentence embedding models for these languages. To address this issue, we compile a relatively small but high-quality human-generated cross-lingual parallel dataset to train LUXEMBEDDER, an enhanced sentence embedding model for Luxembourgish with strong cross-lingual capabilities. Additionally, we present evidence suggesting that including low-resource languages in parallel training datasets can be more advantageous for other low-resource languages than relying solely on high-resource language pairs. Furthermore, recognizing the lack of sentence embedding benchmarks for low-resource languages, we create a paraphrase detection benchmark specifically for Luxembourgish, aiming to partially fill this gap and promote further research.[1]

1 Introduction

The development of sentence embedding models has been instrumental in applications such as Bitext Mining (Artetxe and Schwenk, 2019), Information Retrieval (Thakur et al., 2021), and most recently Retrieval Augmented Generation (Lewis et al., 2020). Generative Large Language Models are not capable of handling these tasks as effectively, making sentence embedding models crucial in these areas. However, these models depend on large-scale parallel data to function effectively, a resource readily available for high-resource languages but sorely lacking for low-resource languages (Zhou et al., 2018).

One way to address this issue is to apply cross-lingual sentence embedding models (Chidambaram et al., 2019; Artetxe and Schwenk, 2019; Reimers and Gurevych, 2020; Yang et al., 2020; Feng et al., 2022; Wang et al., 2022), which aim to embed various languages into a common shared representation space. This approach is intended to boost the performance of low-resource languages by leveraging cross-lingual transfer, where knowledge gained from high-resource languages contributes to the understanding and processing of low-resource languages. However, due to the significant differences in data availability, these models still exhibit a large performance gap between high-resource and low-resource languages.

Luxembourgish, a West-Germanic language spoken by about 400 000 people, is one of the many languages that face this challenge. While translation models for Luxembourgish exist (NLLB Team et al., 2022; Song et al., 2023), their performance remains significantly inferior to that of high-resource languages, hindering the creation of parallel data using methods like back-translation. This limitation also applies to general-purpose generative LLMs, making the direct creation of synthetic parallel data impractical as well. Our research aims to address this issue by collecting a comprehensive set of high-quality human-generated cross-lingual parallel data specifically for Luxembourgish. With this data, we train a sentence embedding model, LUXEMBEDDER, tailored specifically for Luxembourgish by leveraging cross-lingual transfer.

Although cross-lingual sentence embedding models harness the strength of cross-lingual transfer to improve low-resource language performance, we argue that this does not eliminate the necessity for parallel data in these languages. Our findings demonstrate that incorporating these languages in parallel training datasets is essential, as it significantly improves alignment within cross-lingual models, particularly among other low-resource

[1]https://github.com/fredxlpy/LuxEmbedder

languages, in contrast to relying solely on high-resource language parallel data.

Another major challenge is the evaluation of sentence embedding models in low-resource languages, given that the primary benchmarks, such as MTEB (Muennighoff et al., 2023) and BEIR (Thakur et al., 2021), predominantly support English and a few other high-resource languages. To address this, we establish a new paraphrase detection benchmark for Luxembourgish, facilitating future research and improving the language's representation in NLP. To thoroughly evaluate our enhanced model, LUXEMBEDDER, we use our own benchmark along with three additional evaluation tasks. The results indicate that LUXEMBEDDER outperforms not only other open-source models but also proprietary models in the majority of cases.

2 Dataset & Benchmark Construction

We create cross-lingual parallel data and a Luxembourgish paraphrase detection benchmark. See Appendix A for details and Figure 1 for an overview.

2.1 Cross-Lingual Parallel Data (LUXALIGN)

We collect news articles from RTL.lu, a Luxembourgish news platform that publishes in Luxembourgish (LB), English (EN), and French (FR). Due to the lack of explicit mapping between language versions, we use the OpenAI text embedding model `text-embedding-3-small`[2] to align articles across language pairs. LaBSE (Feng et al., 2022) is then employed to extract parallel sentences from these aligned pairs for LB-FR and LB-EN.

2.2 Luxembourgish Paraphrase Detection (PARALUX) Benchmark

Then, we repeat the same process but focusing exclusively on Luxembourgish articles. Within each article, using the same setup, we extract parallel sentences, which can be considered near-paraphrases, from which we hand-pick high-quality samples for our benchmark. From these paraphrased pairs, we prompt `GPT-4o`[3] to generate adversarial negative samples for each pair. Given its limited language capabilities in Luxembourgish, the generated adversarial negative samples are then checked and, if needed, corrected by a human annotator to ensure high quality and accuracy.

[2] https://platform.openai.com/docs/guides/embeddings/embedding-models

[3] https://openai.com/index/hello-gpt-4o/

Through this methodology, we gather 25 996 LB-EN, 86 293 LB-FR samples for LUXALIGN, and 312 samples for PARALUX.

3 LUXEMBEDDER

3.1 Training

Given its cross-lingual capabilities and its already existing support of Luxembourgish, we use LaBSE (Feng et al., 2022) as our base model, which we further train on both LB-EN & LB-FR parallel subsets from LUXALIGN.

We train the model using a batch size of 16 for 3 epochs with a constant learning rate of 1×10^{-6} using a contrastive loss function. We reserve 1% of the data for evaluation, on which we evaluated every 500 steps, and retained the model with the best loss on the development set. The negative pairs for the loss function are created by randomly pairing each Luxembourgish sentence with the translation of another sentence from the dataset.

3.2 Evaluation

We comprehensively compare LUXEMBEDDER's performance across multiple tasks against a variety of open-source and proprietary baseline models.

3.3 Baselines

We provide more details on the used models in Appendix B.2.1.

Proprietary Models Developed by Cohere, `embed-multilingual-light-v3.0` and `embed-multilingual-v3.0` are multilingual embedding models, designed to handle over 100 languages, including Luxembourgish, producing embeddings of size 384 and 1 024, respectively.

OpenAI's `text-embedding-3-small` and `text-embedding-3-large` models generate embeddings with dimensions of 1 536 and 3 072, respectively. Despite the native API feature for embedding shortening, we use the full dimensions in our experiments. While these models have been assessed on the multilingual MIRACL benchmark (Zhang et al., 2023), there is no official information on the number of supported languages.

Open-Source Models We also compare LUXEMBEDDER against two open-source multilingual sentence embedding models that support Luxembourgish. These models are LaBSE (Feng et al., 2022), which generates cross-lingual sentence embeddings for 109 languages, and LASER (Artetxe

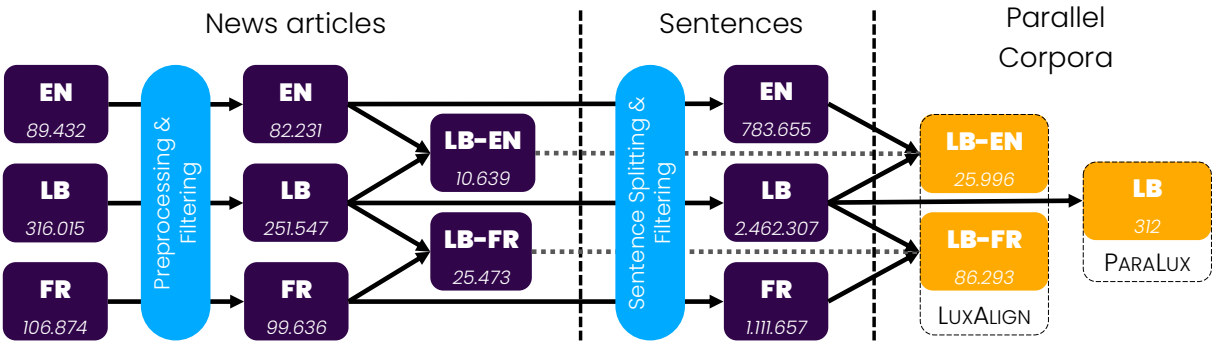

Figure 1: Our data construction workflow involves preprocessing and filtering news articles in English (EN), Luxembourgish (LB), and French (FR), aligning them through sentence embeddings, extracting parallel sentences from aligned article pairs to create LuxAlign, and generating the Luxembourgish paraphrase detection benchmark ParaLux. The numbers in *italics* represent the number of documents used at each stage.

and Schwenk, 2019; Heffernan et al., 2022), which incorporates a multilingual teacher sentence embedding model and language-specific student models for 200 languages.

We further extend our evaluation to include mBERT, a multilingual BERT (Devlin et al., 2019), variant pre-trained on 104 languages, and LuxemBERT (Lothritz et al., 2022), a monolingual Luxembourgish BERT model. In our experiments, we leverage both CLS embeddings and MEAN-pooled embeddings from these models.

3.4 Evaluation Tasks

Additional details on the specific evaluation setup can be found in Appendix B.2.2.

Zero-Shot Classification Using SIB-200 (Adelani et al., 2024), a 7-class classification dataset, we perform similarity-based zero-shot classification. First, we fill each label into a pre-defined template sentence, and separately encode both the input document and all potential template-embedded labels. Then, the class with the most similar embedding to the input document is chosen, assessing the model's ability to generalize to new, unseen tasks without any task-specific training. To account for variability, we repeat this process for 5 different label templates and report the average performance.

Cross-Lingual Transfer For cross-lingual transfer performance, we use the embeddings generated by the respective model to fine-tune a classifier on the SIB-200 dataset in six different high-resource source languages and evaluate directly on the Luxembourgish test set.

Bitext Mining We evaluate the model's proficiency in accurately retrieving or matching parallel sentence pairs from a bilingual corpus using the Tatoeba dataset. Since the original Tatoeba test set (Artetxe and Schwenk, 2019) does not include Luxembourgish, we use the LB-EN, LB-NL, and LB-DE test sets developed by the *Tatoeba Translation Challenge* (Tiedemann, 2020).

ParaLux Lastly, we evaluate the model on our newly created benchmark for paraphrase detection. This task involves determining which of two sentences is a paraphrase of a given anchor sentence. It tests the model's ability to discern nuanced semantic equivalence, which is critical for applications like plagiarism detection, question answering, and information retrieval.

3.5 Results

LuxEmbedder demonstrates superior performance among open-source models in all four tasks and even outperforms all tested proprietary models in 3 out of 4 tasks (Table 1). Only `text-embedding-3-large` model shows superior cross-lingual transfer performance.

In particular, we observe considerable improvements in LuxEmbedder 's performance on both monolingual tasks, Zero-Shot Classification and Paraphrase Detection, relative to its base model, LaBSE. This confirms the efficacy of our cross-lingual approach for Luxembourgish.

4 Cross-Lingual Alignment

In this section, we investigate the impact of fine-tuning models on parallel data for cross-lingual alignment between and within high-resource (HR) and low-resource (LR) languages.

	Model	CL Transfer	Bitext Mining	Zero-Shot Classific.	PARALUX
Proprietary	**Cohere/embed-multilingual-light-v3.0**	70.89	50.10	40.39	37.50
	Cohere/embed-multilingual-v3.0	79.49	59.38	53.33	49.04
	OpenAI/text-embedding-3-small	72.59	39.30	40.20	15.71
	OpenAI/text-embedding-3-large	**86.25**	56.04	58.82	26.28
Open-Source	**mBERT (MEAN)**	70.53	28.44	15.49	5.13
	mBERT(CLS)	70.20	22.27	13.73	4.81
	LuxemBERT (MEAN)	48.47	30.33	14.02	7.69
	LuxemBERT(CLS)	56.86	21.94	33.73	14.42
	LASER	62.70	62.96	11.08	16.03
	LaBSE	80.88	70.11	43.24	38.14
	LUXEMBEDDER	83.39	**70.24**	**65.59**	**52.24**

Table 1: Comparison of LuxEmbedder with various open-source and proprietary models across two cross-lingual and two monolingual tasks. We report accuracy for all 4 tasks. The best overall performance for each task is highlighted in **bold**, while the best performance among open-source models is underlined.

Experimental Setup To measure the cross-lingual alignment, we use Flores-200 (NLLB Team et al., 2022), which includes parallel sentences across 200 languages, making it an ideal resource for assessing cross-lingual alignment. We use the Centered Kernel Alignment (CKA) method (Kornblith et al., 2019) to calculate the level of alignment by comparing the embeddings of parallel sentences from different languages.

We fine-tune LaBSE on three different language pairs: LB-EN, LB-FR, and EN-FR[4], each time using 20 000 parallel sentences from our newly compiled datasets. After fine-tuning, we assess cross-lingual alignment by comparing alignment within HR languages and LR languages, as well as between LR and HR languages[5].

Results Our observations (Figure 2) reveal that when fine-tuning on parallel data, the alignment within the model generally increases. HR languages benefit equally from fine-tuning on any of the three language pairs. However, we observe that the alignment of LR languages benefits more when Luxembourgish is part of the training data compared to fine-tuning on HR language pairs alone.

These results indicate the critical importance of including LR languages, such as Luxembourgish, when collecting parallel data. Incorporating LR in the training process enhances cross-lingual alignment, not only for the respective language pair but also for other LR languages, more effectively than focusing solely on HR languages.

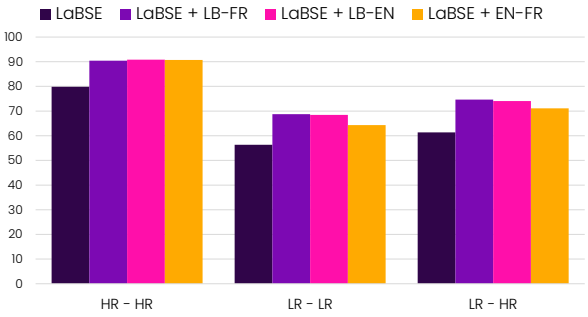

Figure 2: The alignment of language-specific embedding spaces within and between high-resource (HR) and low-resource (LR) languages, measured using the CKA method, is shown for LaBSE before and after fine-tuning on LB-EN, LB-FR, and EN-FR parallel data. The exact values for all language pairs are provided in Figure 3.

5 Conclusion

Sentence embedding models struggle with low-resource languages due to a shortage of parallel data. To address this problem, we collected high-quality, human-generated cross-lingual parallel data for Luxembourgish and developed an enhanced version of a cross-lingual sentence embedding model specifically adapted to Luxembourgish. This model outperforms open-source as well as proprietary models in almost all evaluations conducted

[4]Created using the same process as described in §2.1.

[5]As HR and LR languages we select the 10 languages with the most and least training data in LaBSE which are also covered by Flores-200.

in our study. Our findings also stress the importance of incorporating low-resource languages in parallel data collection, as evidence suggests that this enhances embedding alignment for both the target language and other low-resource languages within the same model more effectively than using high-resource language pairs alone. Therefore, we believe this research encourages further creation of parallel corpora for low-resource languages.

Limitations

It is important to note that we do not compare our embedding model against general-purpose generative LLMs. We acknowledge that some of these models, which are significantly larger in terms of parameter count, may outperform LUXEMBEDDER in certain tasks. Nonetheless, the primary objective of our paper is not to compete with generative models . Instead, our focus is on providing a robust sentence embedding model capable of solving specific tasks such as information retrieval, document clustering, and similar applications where generative language models may not be as effective.

Additionally, we acknowledge that our data is limited to the news domain, due to its availability. However, our goal is to use this data to boost the model's retrieval performance, facilitating future expansion into various other domains by mining a more diverse range of parallel data.

Ethical Statement

In the newly created PARALUX benchmark, the adversarial counterparts of the paraphrases have been edited in a way that some of the edited sentences may contain non-factual information. Therefore, we strongly recommend using this data solely, as designed, for evaluation purposes and not for training, to ensure the integrity of model development.

Furthermore, our datasets, based on news articles, naturally include the names of individuals. As the text is publicly available and anonymization would greatly diminish data quality, we chose not to anonymize it. We believe that preserving the original context of publicly accessible information is essential for maintaining data integrity and the effectiveness of our research.

Acknowledgments

We are grateful to RTL Luxembourg for providing the raw data necessary for our research. Their support significantly facilitated our efforts.

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An Open Dataset and Model for Language Identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

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.

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.

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.

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.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089, Marseille, France. European Language Resources Association.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint*. ArXiv:2207.04672 [cs].

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.

Yewei Song, Saad Ezzini, Jacques Klein, Tegawende Bissyande, Clément Lefebvre, and Anne Goujon. 2023. Letz Translate: Low-Resource Machine Translation for Luxembourgish. In *2023 5th International Conference on Natural Language Processing (ICNLP)*, pages 165–170.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English Contrastive Learning Can Learn Universal Cross-lingual Sentence Embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Zhong Zhou, Matthias Sperber, and Alexander Waibel. 2018. Massively Parallel Cross-Lingual Learning in Low-Resource Target Language Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 232–243, Brussels, Belgium. Association for Computational Linguistics.

A Data Collection & Processing

Here, we outline the method used to create cross-lingual training data and the paraphrase detection benchmark, providing examples in Tables 2 and 3.

Processing Articles We gather news articles from the Luxembourgish news platform RTL[6] written in Luxembourgish, French, and English, covering different time periods: from January 1, 1999 for Luxembourgish, from September 1, 2011 for French, and from January 1, 2018 for English, up until May 10, 2024. We first remove all URL tags and extraneous metadata, and filter out articles with fewer than 100 characters, as these are often just traffic or sports updates, which were not relevant for our study. To ensure linguistic accuracy, we use the OpenLID (Burchell et al., 2023) to identify and exclude articles that are not in the intended language.

[6] https://www.rtl.lu

Article Matching Subsequently, we embed each article using the OpenAI `text-embedding-3-small` model to facilitate cross-language article matching. To identify potential parallel articles in different languages, we first narrow down the candidates by considering only those articles published within a one-day window of the target article. Among these candidates, we select the one with the highest cosine similarity to the target article's embedding, provided the similarity score exceeds 0.65.

Sentence Extraction In parallel, we extract sentences from each article using the NLTK[7] library. For Luxembourgish, in the absence of a dedicated sentence tokenizer, we use the German tokenizer. After splitting the articles into sentences, we employ OpenLID once again to remove any sentences identified as being in the wrong language. Additionally, we filter out sentences with fewer than 10 characters or fewer than three words.

Sentence Matching Next, we embed each sentence using LaBSE, focusing on sentences from articles already matched with articles in another language. For each sentence, we restrict the candidates to sentences from the corresponding matched article, minimizing the risk of false positives. We then select the candidate sentence with the highest cosine similarity, provided it exceeds a similarity threshold of 0.7. After identifying all sentence pairs, we filter out pairs where the length difference is greater than 50%. To create a seed dataset for PARALUX, we replicate this process within Luxembourgish articles alone.

B Training and Evaluation Details for LUXEMBEDDER

All our training processes and experiments were run on 4 A100 GPUs within a few hours.

B.1 Training

Given a sentence embedding model $\mathcal{M}_\theta$ with parameters θ, for a sentence pair (x_1, x_2) and its label y (1 if positive pair, 0 if negative pair), the contrastive loss function is defined as:

$$\mathcal{L}(\theta, (x_1, x_2, y)) = \frac{1}{2}\left[y \cdot \mathrm{D}^2 + (1-y) \cdot \max(0, \mathrm{m} - \mathrm{D})^2\right] \quad (1)$$

where

- $D = d\left(\mathcal{M}_\theta(x_1), \mathcal{M}_\theta(x_2)\right)$
- m is the margin value, defining the minimum distance that samples withing a negative pair should have

with $m = 0.5$ and d being the cosine distance in our experiments.

B.2 Evaluation

B.2.1 Baseline Models

Due to the proprietary nature of Cohere's models, `embed-multilingual-light-v3.0` and `embed-multilingual-v3.0`, as well as OpenAI's `text-embedding-3-small` and `text-embedding-3-large`, detailed information about their training data and model architecture is not publicly available. We refer readers to their online documentation[8] [9] for any details.

Our experiments with open-source models involve base multilingual BERT (cased) (Devlin et al., 2019) and LuxemBERT (Lothritz et al., 2022). These models feature identical architectures, including 12 attention heads and 12 transformer blocks, each with a hidden size of 768. mBERT's vocabulary size is 30 000, whereas LuxemBERT's is 119 547. Both models have about 110 million parameters.

Additionally, we incorporate LaBSE (Feng et al., 2022), which also serves as the foundational model for LUXEMBEDDER. LaBSE is derived from the base multilingual BERT (cased) but features an expanded vocabulary of 501 153 tokens. It has been trained using a combination of monolingual data and bilingual translation pairs.

B.2.2 Evaluation Tasks

Cross-Lingual Transfer

To assess cross-lingual transfer performance, we use embeddings from the respective model to fine-tune a classifier on the SIB-200 (Adelani et al., 2024) dataset in several high-resource source languages, then evaluate it directly on the Luxembourgish test set.

The SIB-200 dataset includes over 200 languages, with 701 training, 99 development and 204 test samples per language.

In our experiments, however, we only train separately on French, English, German, Japanese, Chinese, and Russian. Additionally, we fine-tune on

[7] https://www.nltk.org

[8] https://cohere.com/blog/introducing-embed-v3

[9] https://openai.com/index/new-embedding-models-and-api-updates/

Luxembourgish, but this is not included in the average performance reported in Table 1. The classifier is a simple linear layer with 7 output nodes, trained with the Adam optimizer and the cross-entropy loss function. Training is performed for 500 epochs with a constant learning rate of $1e^{-2}$. We evaluate the classifier once per epoch and select the model with the best development loss. Each training process is repeated 4 times using different seeds to ensure robustness, and we report the average performance per source language in Table 4.

Zero-Shot Classification

To assess the zero-shot classification capabilities of different model, we again use the SIB-200 dataset (Adelani et al., 2024). We independently encode the input and all potential labels, integrating each label within a prompt template. The class whose embedding has the highest cosine similarity to the input document is selected.

We use five different prompt templates to evaluate the classification performance and report the average performance per template in Table 5. These templates are:

1. `[LABEL]`
2. `An dësem Beispill geet et em [LABEL].`
   This example is about [LABEL].
3. `D'Thema vun dësem Text ass [LABEL].`
   The topic of this text is [LABEL].
4. `Hei gëtt iwwer [LABEL] geschwat.`
   Here we are talking about [LABEL].
5. `Dëst Dokument beschäftegt sech mat [LABEL].`
   This document deals with [LABEL].

The labels in Luxembourgish we use in this classification task are `Technologie` (*technology*), `Reesen` (*travel*), `Politik` (*politics*), `Gesondheet` (*health*), `Ennerhalung` (*entertainment*), `Geographie` (*geography*) and `Sport` (*sports*).

Bitext Mining

We initially considered the Tatoeba dataset, but it lacks Luxembourgish in the original set. Instead, we used Luxembourgish-English, Luxembourgish-Dutch, and Luxembourgish-German test sets from the *Tatoeba Translation Challenge* (Tiedemann, 2020), which include 346 LB-EN, 291 LB-EN, and 292 LB-DE sample pairs.[10] We conducted experiments in both retrieval directions and reported the full results in Table 6.

PARALUX

To assess performance on PARALUX, the model encoded the anchor sentence and both paraphrase candidates. The candidate with the greatest cosine similarity to the anchor was chosen as the predicted paraphrase.

C Full Results

Here, we report the full experimental results from the evaluations on Cross-Lingual Transfer (Table 4), Zero-Shot Classification (Table 5) and Bitext Mining (Table 6) conducted in Section 3.5.

D Details on the Cross-Lingual Alignment Experiments

In Section 4, we measure the alignment of language-specific subspaces using the Centered Kernel Alignment (CKA) method (Kornblith et al., 2019). The CKA score of two representation matrices $X \in \mathbb{R}^{N \times m}$ and $Y \in \mathbb{R}^{N \times m}$, where N is the number of samples and m is the embedding dimension of the model, when using a linear kernel, is given by

$$CKA(X, Y) = 1 - \frac{\|XY^T\|_F^2}{\|XX^T\|_F \|YY^T\|_F}$$

where $\|\cdot\|_F$ is the Frobenius norm.

Since parallel cross-lingual data is essential for computing the CKA across various languages, we use the Flores-200 dataset (NLLB Team et al., 2022), which includes human-curated translations between English and 204 other languages. Specifically, we use the devtest split, containing 1 012 aligned sentences per language.

We choose the 10 languages with the highest and lowest amounts of training data in LaBSE, which are also included in Flores-200, to represent the HR and LR languages. As LR languages, we use `bod`, `snd`, `tuk`, `ydd`, `wol`, `asm`, `smo`, `xho`, `nya`, and `sot`. As HR languages, we use `eng`, `rus`, `jpn`, `zho`, `fra`, `deu`, `por`, `nld`, `spa`, and `pol`.

The exact CKA values across all language pairs are provided in Figure 3.

[10] https://huggingface.co/datasets/Helsinki-NLP/tatoeba_mt

Luxembourgish Sentence	English/French Sentence
D'Police sicht no engem Mann, deen an der Stad mat enger geklauter Kreditkaart Suen opgehuewen huet.	The police is looking for a man who withdrew money with a stolen credit card in Luxembourg City.
D'Temperaturen am Grand-Duché sinn an der Moyenne em 1.3 Grad an d'Luucht gaangen.	Temperatures in the Grand Duchy have risen by 1.3 degrees on average.
Déi Petitioun ass vun 336.000 Persounen aus 112 Länner ënnerschriwwe ginn.	Cette pétition a été signée par 336.000 personnes originaires de 112 pays.
Am September 2013 hat fir d'éischte Kéier e Lëtzebuerger den Jackpot gewonnen.	En septembre 2013, un Luxembourgeois avait pour la 1e fois remporté le jackpot.

Table 2: Examples from the compiled parallel LB-EN & LB-FR dataset LUXALIGN.

Anchor Sentence	Paraphrase	Not Paraphrase
Mexiko gewënnt 3-1 géint Kroatien. *Mexico wins 3-1 against Croatia.*	Kroatien verléiert 1-3 géint Mexiko. *Croatia loses 3-1 against Mexico.*	Kroatien gewënnt 3-1 géint Mexiko. *Croatia wins 3-1 against Mexico.*
De Sträit tëscht Süd- a Nordkorea spëtzt sech weider zou. *The dispute between South and North Korea continues to escalate.*	D'Verhältnis tëscht Nord- a Südkorea gëtt ëmmer méi schlecht. *The relationship between South and North Korea is getting worse and worse.*	De Sträit tëscht Süd- a Nordkorea entspaant sech weider. *The dispute between South and North Korea continues to ease.*

Table 3: Examples from the newly created Luxembourgish paraphrase detection benchmark PARALUX.

	Model	Source Language						
		de	en	fr	jp	ru	zh	lb
Proprietary	**Cohere/embed-multilingual-light-v3.0**	69.24	70.10	69.00	73.28	74.26	69.49	76.96
	Cohere/embed-multilingual-v3.0	79.90	82.60	76.47	79.29	78.19	80.51	83.09
	OpenAI/text-embedding-3-small	76.96	73.53	68.14	72.18	74.02	70.71	77.08
	OpenAI/text-embedding-3-large	**87.75**	**84.44**	**85.54**	**85.54**	**87.75**	**86.52**	**88.36**
Open-Source	**mBERT (MEAN)**	72.67	70.96	73.28	65.44	72.55	68.26	75.12
	mBERT (CLS)	70.10	68.75	71.08	67.89	73.28	70.10	75.25
	LuxemBERT (MEAN)	79.41	74.02	81.74	23.41	9.56	22.67	82.48
	LuxemBERT (CLS)	80.02	81.86	79.29	35.05	28.92	36.03	<u>84.44</u>
	LASER	62.01	61.76	63.36	61.76	62.38	64.95	57.84
	LaBSE	80.51	79.90	80.88	82.35	80.39	81.25	80.15
	LUXEMBEDDER	<u>83.82</u>	<u>82.48</u>	<u>82.84</u>	<u>83.58</u>	<u>83.33</u>	<u>84.31</u>	84.31

Table 4: **Cross-lingual Transfer Performance**: Comparative results of models on the SIB-200 dataset. Linear classifiers were trained using model embeddings in various source languages and evaluated on the Luxembourgish test set. The table shows average performance from 4 experiment iterations. The best overall performance for each source language is highlighted in **bold** while the best performance among open-source models is <u>underlined</u>.

	Model	Label Template				
		1	**2**	**3**	**4**	**5**
Proprietary	**Cohere/embed-multilingual-light-v3.0**	42.65	42.65	35.29	42.16	39.22
	Cohere/embed-multilingual-v3.0	45.59	56.86	53.92	58.82	51.47
	OpenAI/text-embedding-3-small	24.02	46.08	35.29	48.04	47.06
	OpenAI/text-embedding-3-large	42.65	65.20	67.16	57.84	61.27
Open-Source	**mBERT (MEAN)**	10.78	15.69	15.69	17.65	17.65
	mBERT (CLS)	11.27	16.67	11.76	16.18	12.75
	LuxemBERT (MEAN)	9.31	9.31	14.71	15.69	21.08
	Luxembert (CLS)	9.31	41.67	23.53	50.49	43.63
	LASER	13.73	10.78	10.78	10.29	9.80
	LaBSE	38.24	45.10	44.12	47.55	41.18
	LUXEMBEDDER	**55.88**	**68.14**	**68.14**	**69.61**	**66.18**

Table 5: **Zero-Shot Classification Performance** on the SIB-200 datasets for five different label templates.

	Model	Language Pair and Direction					
		lb←de	**lb→de**	**lb←en**	**lb→en**	**lb←nl**	**lb→nl**
Proprietary	**Cohere/embed-multilingual-light-v3.0**	43.64	54.91	46.58	54.11	44.33	57.04
	Cohere/embed-multilingual-v3.0	52.60	58.38	57.88	65.75	52.23	69.42
	OpenAI/text-embedding-3-large	56.36	51.73	53.08	58.22	59.11	57.73
	OpenAI/text-embedding-3-small	46.24	37.57	34.25	38.36	42.61	36.77
Open-Source	**mBERT (CLS)**	25.14	25.43	11.64	18.15	25.43	27.84
	mBERT (MEAN)	36.71	26.88	22.60	22.26	34.02	28.18
	LuxemBERT (CLS)	47.40	54.05	6.85	7.53	9.28	6.53
	LuxemBERT (MEAN)	62.14	65.32	11.99	15.41	13.75	13.40
	LASER	57.80	59.25	64.73	66.44	62.54	67.01
	LaBSE	**67.63**	67.63	**70.89**	**70.89**	**73.54**	70.10
	LUXEMBEDDER	66.47	**68.50**	**70.89**	69.18	73.20	**73.20**

Table 6: **Bitex Mining Performance** on the Tatoeba dataset for three different language pairs. Retrieval accuracy values are provided for each language pair in both retrieval directions.

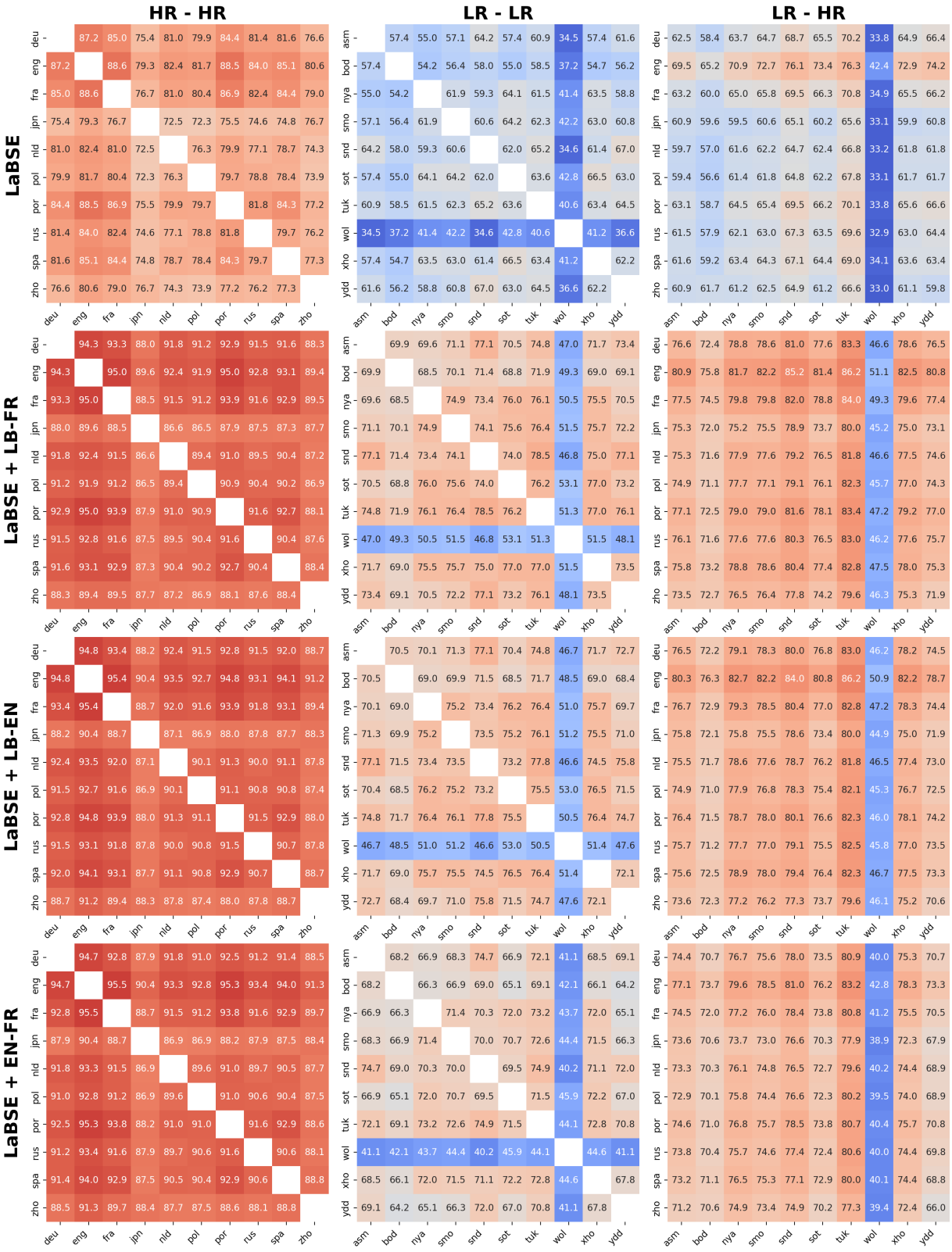

Figure 3: Alignment of language-specific embedding spaces within and between high-resource (HR) and low-resource (LR) languages, measured using the CKA method for LaBSE before and after fine-tuning on LB-EN, LB-FR, and EN-FR parallel data.