# Multilingual Sentence-T5:
# Scalable Sentence Encoders for Multilingual Applications

**Chihiro Yano**[1,2] **Akihiko Fukuchi**[2] **Shoko Fukasawa**[2]
**Hideyuki Tachibana**[2] **Yotaro Watanabe**[2]
[1]Graduate School of Informatics, Nagoya University [2]PKSHA Technology Inc.
yano.chihiro.j3@s.mail.nagoya-u.ac.jp
{akihiko_fukuchi,shoko_fukasawa,h_tachibana,y_watanabe}@pkshatech.com

## Abstract

Prior work on multilingual sentence embedding has demonstrated that the efficient use of natural language inference (NLI) data to build high-performance models can outperform conventional methods. However, the potential benefits from the recent "exponential" growth of language models with billions of parameters have not yet been fully explored. In this paper, we introduce Multilingual Sentence T5 (m-ST5), as a larger model of NLI-based multilingual sentence embedding, by extending Sentence T5, an existing monolingual model. By employing the low-rank adaptation (LoRA) technique, we have achieved a successful scaling of the model's size to 5.7 billion parameters. We conducted experiments to evaluate the performance of sentence embedding and verified that the method outperforms the NLI-based prior approach. Furthermore, we also have confirmed a positive correlation between the size of the model and its performance. It was particularly noteworthy that languages with fewer resources or those with less linguistic similarity to English benefited more from the parameter increase. Our model is available at https://huggingface.co/pkshatech/m-ST5.

**Keywords:** sentence embedding, multilingual, encoder-decoder model

## 1. Introduction

Sentence embedding is a versatile and fundamental technique of NLP and has been studied extensively (Kiros et al., 2015; Logeswaran and Lee, 2018; Reimers and Gurevych, 2019; Yan et al., 2021; Meng et al., 2021; Carlsson et al., 2021; Kim et al., 2021; Muennighoff, 2022). In particular, the recently proposed SimCSE (Gao et al., 2021), a simple and data-efficient method based on contrastive fine-tuning of existing pre-trained text encoders such as BERT, greatly advanced the frontier and attracted much attention. This technique can be naturally used with other kinds of model architectures. For example, in their Sentence T5, Ni et al. (2022) adopted T5, an encoder-decoder model.

Multilingual sentence embedding, which projects sentences from diverse languages into a shared semantic space, is an important extension of this problem, and many techniques have been proposed (Section 2). Of these, we particularly focus on a multilingual extension of SimCSE, namely mSimCSE (Wang et al., 2022), because of its data efficiency. In particular, even though the fine-tuning requires just a natural language inference (NLI) dataset consisting of around 2 million sentences, it showed comparable results with supervised techniques based on larger parallel corpora.

Now, the natural question that arises here is whether such a learning strategy scales to larger models with billions of parameters. To answer this question, this paper examines the performance of a fine-tuned model based on mT5 (Xue et al., 2021), a larger model than XLM-RoBERTa (Conneau et al., 2020), which is the base model of mSimCSE. We extend the existing large-scale monolingual model Sentence T5 to multilingual scenarios.

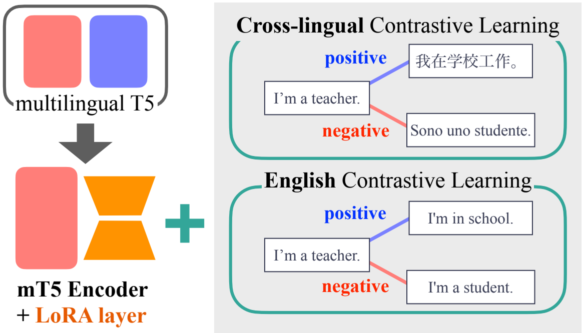

Figure 1: Concept diagram of m-ST5.

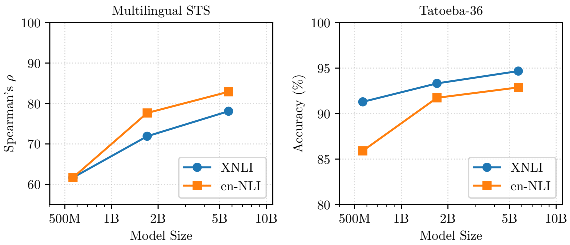

Figure 2: Comparison of model size and performance of the proposed method.

The proposed method performed well on some benchmarks, including cross-lingual STS (XSTS)

*LREC-COLING 2024*, pages 11849–11858
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

(Cer et al., 2017a) and sentence retrieval (Artetxe and Schwenk, 2019; Zweigenbaum et al., 2017). Besides, it outperformed a monolingual model in Japanese, a language far distant from the ones used for training. We have further confirmed a positive correlation between the model size and its performance, as shown in Figure 2, which is often referred to as the scaling law and found in other language models (Ni et al., 2022; Kaplan et al., 2020). The observation that the scaling law holds for multilingual sentence embeddings suggests that the constraint of insufficient amount of training data in low-resource languages may be alleviated by using large-scale pre-trained models.

## 2. Related Work

Recently, multilingual models (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020; Wei et al., 2021; Chi et al., 2022; Xue et al., 2021) have been intensively studied as they have language transferability, the ability to adapt to a new language in a few or zero shots. It should also be noted that the transferability varies by task and language pair and is not always effective, especially between distant languages (Pires et al., 2019; Lauscher et al., 2020).

As an important branch of multilingual NLP, various techniques for multilingual sentence embedding have been studied. A major challenge in this field is how to acquire semantic proximity between sentences in different languages. A natural approach would be the use of parallel corpora (Artetxe and Schwenk, 2019; Yang et al., 2020; Feng et al., 2022), though this approach is data-hungry, and it is costly to maintain such corpora. Some techniques have been proposed that can alleviate such problems, such as a distillation-based approach (Reimers and Gurevych, 2020), introduction of adversarial training strategy to reduce language identifier information from semantic vectors (Chen et al., 2019; Keung et al., 2019), the use of word alignment between parallel sentences (Cao et al., 2020), and so on.

In this paper, we particularly focus on the NLI-based contrastive training strategy, specifically the multilingual extension of SimCSE (Gao et al., 2021), namely mSimCSE (Wang et al., 2022). It was shown that mSimCSE could acquire inter-language alignments without explicit parallel corpora, and even monolingual NLI corpora could yield good fine-tuning results.

## 3. Proposed Method: m-ST5

In this paper, we propose the *Multilingual Sentence T5* (m-ST5) as a new extension of Sentence T5 (Ni et al., 2022). Similarly to ST5, our method is based on fine-tuning of a pre-trained T5 model (Raffel et al., 2020), which is one of the most popular encoder-decoder language models. However, since we are interested in multilingual sentence embedding, we need to use a multilingual model as our baseline. We then used Multilingual T5 (mT5) (Xue et al., 2021), which was pre-trained on mC4, a large-scale multilingual corpus covering 101 languages. To build a sentence embedding model, only the encoder part of the enc-dec model is needed, as in Sentence T5. For example, the encoder module (5.7B params) out of the pre-trained mT5-xxl (13B params) is extracted. The encoder converts a sentence into token-wise embedding, and these token representations are averaged together to produce a sentence embedding.

Naturally, the vector obtained this way is not sufficient for sentence embedding in quality, and fine-tuning of the encoder is required. To this end, following mSimCSE (Wang et al., 2022), we trained m-ST5 in a contrastive manner using the NLI dataset for a task to predict whether a given hypothesis sentence is an *entailment*, a *contradiction*, or *neutral* to another premise sentence. Specifically, m-ST5 is trained to minimize the distances between positive pairs (entailment) and maximize the distances between negative ones (contradiction). Furthermore, unrelated in-batch sentences are also incorporated as negative samples because such a trick promotes the uniformity of the semantic space (Gao et al., 2021).

Additionally, in multilingual learning scenarios using cross-lingual NLI data (XNLI) (Conneau et al., 2018), it should also be taken into account which languages the positive and negative samples are drawn from. In this study, following mSimCSE, we draw triplets, each of which consists of a premise and two hypotheses (entailment and contradiction) from different languages, as shown in Figure 1.

## 4. Experiment

### 4.1. Cross-lingual Experiments

We first evaluated the quality of sentence embedding by sentence retrieval tasks (Tatoeba, Artetxe and Schwenk, 2019; and BUCC, Zweigenbaum et al., 2017) and cross-lingual STS task (XSTS, Cer et al., 2017a). Details of the evaluation task are provided in Section 7. For all tasks, cosine similarity was used as the measure of similarity. We compared the performance with the following methods: mSimCSE (Wang et al., 2022), LASER (Artetxe and Schwenk, 2019), and LaBSE (Feng et al., 2022). Of these, LASER and LaBSE were trained in a fully supervised manner.

In order to feasibly train our models on a single A100 GPU with 80GB of VRAM, we used the tech-

| | Train | Fine Tuning | Tatoeba tasks (Accuracy) | | BUCC ($F_1$) | XSTS tasks (Spearman's $\rho$) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | method | Data | Tatoeba-14 | Tatoeba-36 | avg. | ar-ar | ar-en | es-es | es-en | tr-en |
| **Contrastive Learning** | | | | | | | | | | |
| mSimCSE | Full FT | XNLI | 93.2$^*$ | 91.4$^*$ | 95.2$^*$ | 79.4$^*$ | 72.1$^*$ | 85.3$^*$ | 77.8$^*$ | 74.2$^*$ |
| | | en-NLI | 89.9$^*$ | 87.7$^*$ | 93.6$^*$ | 81.6$^*$ | 71.5$^*$ | 87.5$^*$ | 79.6$^*$ | 71.1$^*$ |
| m-ST5 | LoRA (query+value) | XNLI | 96.3 | 94.7 | 97.6 | 76.2 | 78.6 | 84.4 | 76.2 | 75.1 |
| | | en-NLI | 93.9 | 92.9 | 96.6 | 83.2 | 79.5 | 87.7 | 84.9 | 79.2 |
| | LoRA (all-linear) | XNLI | **96.5** | 94.8 | **97.7** | 77.3 | 77.8 | 85.0 | 77.7 | 75.0 |
| | | en-NLI | 94.0 | 93.1 | 96.7 | **84.5** | **82.9** | **89.2** | **86.3** | **79.7** |
| **Fully Supervised** | | | | | | | | | | |
| LASER | - | - | 95.3$^\dagger$ | 84.4$^\dagger$ | 93.0$^\ddagger$ | 68.9$^\ddagger$ | 66.5$^\ddagger$ | 79.7$^\ddagger$ | 57.9$^\ddagger$ | 72.0$^\ddagger$ |
| LaBSE | - | - | 95.3$^\dagger$ | **95.0**$^\dagger$ | 93.5$^\ddagger$ | 69.1$^\ddagger$ | 74.5$^\ddagger$ | 80.8$^\ddagger$ | 65.5$^\ddagger$ | 72.0$^\ddagger$ |

Table 1: Evaluation results using Tatoeba, BUCC, and XSTS. Each score is the average of the performance over three trials with different random seeds. Scores with $^*$, $^\dagger$ and $^\ddagger$ were excerpted from (Wang et al., 2022), (Feng et al., 2022), and (Reimers and Gurevych, 2020), respectively.

| Model | Train method | FT Data | hi | fr | de | af | te | tl | ga | ka | am | sw |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Contrastive Learning** | | | | | | | | | | | | |
| mSimCSE | Full FT | XNLI | 96.2 | 94.8 | 98.8 | 90.6 | 96.2 | 80.9 | 65.1 | 92.4 | 82.4 | 67.8 |
| | | en-NLI | 94.4 | 93.9 | 98.6 | 85.6 | 92.9 | 70.0 | 54.8 | 89.2 | 79.5 | 42.1 |
| m-ST5 | LoRA (query+value) | XNLI | **98.0** | 96.2 | **99.7** | 95.6 | 98.2 | 94.3 | 83.0 | 95.4 | 93.1 | 91.2 |
| | | en-NLI | 97.5 | 95.7 | 99.4 | 94.5 | 97.3 | 93.1 | 81.6 | 93.3 | 90.7 | 68.5 |
| | LoRA (all-linear) | XNLI | 97.8 | **96.4** | 99.6 | 95.6 | 97.9 | 94.1 | 84.3 | 95.6 | **94.3** | **91.5** |
| | | en-NLI | 97.6 | 95.7 | 99.3 | 94.5 | 97.2 | 93.5 | 82.4 | 93.9 | 91.7 | 68.8 |
| **Fully Supervised** | | | | | | | | | | | | |
| LASER | - | - | 94.7 | 95.7 | 99.0 | 89.4 | 79.7 | - | 5.2 | 35.9 | 42.0 | 42.4 |
| LaBSE | - | - | 97.8 | 96.0 | 99.4 | **97.4** | **98.3** | **97.4** | **95.0** | **95.9** | 94.0 | 88.5 |

Table 2: Accuracy of Tatoeba retrieval task. Target languages are the same as in (Wang et al., 2022).

nique of LoRA (Hu et al., 2021), which enables training of very large models with limited computational resources. In this paper, we examined two LoRA conditions. One is to apply the LoRA technique only to the query and value matrices, following the original LoRA paper (Hu et al., 2021), which reported that fine-tuning only the query and value matrices is effective. The other is to apply LoRA to all linear layers, following the QLoRA paper (Dettmers et al., 2023) which showed that the highest performance is obtained by fine-tuning all linear layers. In both cases, the rank of the matrices was $r = 8$, and the batch size was 128 in our experiment. Other details of the experimental setup are in Section 8. The training data was chosen from the following NLI datasets: en-NLI = SNLI (Bowman et al., 2015) + MNLI (Williams et al., 2018), and XNLI (Conneau et al., 2018). Details of the training data are in Section 9.

Table 1 shows the evaluation results on the above three tasks. We may observe that the proposed method (m-ST5) outperformed the existing mSimCSE. Notably, even monolingual (en-NLI) fine-tuning of m-ST5 outperformed multilingual (XNLI) fine-tuning of mSimCSE. It is also observed that all-linear LoRA gave better results than query+value LoRA. The difference was especially noticeable when en-NLI data was used for training.

We also found that it depends on the task which of XNLI/en-NLI fine-tuning data gave better results. Specifically, we have observed that cross-lingual training data (XNLI) was notably more effective in sentence retrieval tasks, while monolingual data (en-NLI) was more effective for the XSTS task. It may reflect differences in the nature of the evaluation metrics, i.e. the sentence retrieval tasks are only concerned with the sentences with top relevance, while the STS task considers ranking that requires more elaborate knowledge on each concept. Here, the XNLI-based learning prioritizes the alignment of different languages, resulting in sparse occurrences of each word and making it harder to acquire elaborated word knowledge. In this respect, monolingual learning using en-NLI, with its opposite properties where the same word often appears both in positive and negative samples, would have been more effective in STS.

Table 2 further details the accuracy of English-* sentence retrieval in various languages. While mSimCSE did not perform well for low-resource languages (e.g. ga) or phylogenetically distant languages from English (e.g. sw), the proposed method produced high scores of over 90% for all of such languages except Irish (ga). Overall, the performance has improved, and the results are approaching those of fully supervised methods.

## 4.2. Comparison with Monolingual Models

To evaluate the transferability of the proposed method, we conducted experiments on Japanese, Korean and Chinese languages which are phylogenetically very distant from English but are rich in evaluation resources and the population of potential users. The performance of our method (m-ST5) on monolingual STS tasks was compared with these monolingual models, as well as the multilingual models.

**Baseline Models** As monolingual baselines, we used the Japanese BERT-large[1], the Korean RoBERTa-large from KLUE (Park et al., 2021), and the Chinese RoBERTa-large (Cui et al., 2020). We will call these models 'ja-BERT', 'ko-RoBERTa' and 'zh-RoBERTa' for simplicity. As multilingual models, we used the LaBSE and mSimCSE. The LaBSE model was from Hugging Face Hub[2], and the mSimCSE model was reproduced based on the original paper (Wang et al., 2022).

**Fine-tuning and Evaluation Data** The training data used for fine-tuning were XLNI, en-NLI, and monolingual NLI datasets for each language. (Note that XNLI do not contain Japanese and Korean languages.) For evaluation, STS data in each language, as well as English STS data (STS-B, Cer et al., 2017b) were used. Note that the English STS is not the main focus of this section, but is for reference.

The Japanese monolingual NLI dataset was JSNLI (Yoshikoshi et al., 2020), and the evaluation STS dataset was JSTS in JGLUE (Kurihara et al., 2022). The Korean dataset was KorNLI/KorSTS (Ham et al., 2020). The Chinese NLI dataset was CMNLI in CLUE (Xu et al., 2020), and the STS dataset was STS-B test set in C-MTEB (Xiao et al., 2023). We will refer to these fine-tuning and evaluation data as {ja, ko, zh}-{NLI, STS}, for simplicity.

Table 3 shows the evaluation results of STS tasks in three languages mentioned. The proposed model with the LoRA layer added only to query+value matrices showed inferior results to the existing multilingual model, mSimCSE. Nevertheless, this problem was solved by adding the LoRA layers to all linear layers, and by increasing the number of trainable parameters (Dettmers et al., 2023).

The average score of the results for these languages was higher than that of the existing multilingual models. The performance for each language was as follows: in Chinese, the proposed method outperforms the monolingual counterpart (zh-RoBERTa), and in Japanese, the performance of the proposed method is equivalent to that of the monolingual counterpart (ja-BERT). In particular, even when the target language data was not used for training at all (i.e., only en-NLI was used), the performance of m-ST5 was comparable to these monolingual models.

This could be attributed to two factors: The first would be that m-ST5 has high cross-lingual transferability, and the second would be the large size and high quality of the en-NLI dataset.

These results suggest that fine-tuning multilingual models for monolingual tasks is a promising option when pre-trained large monolingual models are not available. Moreover, high performance could be achieved without using target language data during fine-tuning.

In Korean language, however, the opposite trend has been observed. The monolingual model trained on the monolingual corpus was significantly better than m-ST5, transfer-learned from a multilingual model. In our observation, this could be attributed to the quality of tokenizers. In general, the tokenization of Korean language is not a straightforward task (Park et al., 2021), and the tokenizer of our base multilingual model does not seem to be sufficiently tuned in this respect. On the other hand, the tokenizer of ko-RoBERTa seems to have been carefully crafted.

## 4.3. Scaling Law

It has been suggested that the performance of language models scale with the increase of the model size. In this section, we investigate whether this is the case for our approach as well as Sentence T5 (Ni et al., 2022). We compared the performance of three pre-trained mT5 models of different sizes (564M, 1.7B, and 5.7B) when fine-tuned query and value matrices using XNLI or en-NLI.

Figure 2 shows the scaling law of the performance of Spearman's $\rho$ on XSTS and the accuracy of multilingual sentence retrieval on Tatoeba-36. Table 4 details the performance on XSTS for various language pairs. In this table, a trend could be observed that languages far from English (i.e., ar and tr) tended to benefit more from the increase in model size. Particularly interesting in this table is the fact that monolingual fine-tuning becomes more effective as the model size is scaled up. This result

[1] https://huggingface.co/cl-tohoku/bert-large-japanese-v2

[2] https://huggingface.co/sentence-transformers/LaBSE

| Model | Train method | FT Data | en-STS | ja-STS | ko-STS | zh-STS | avg. |
|---|---|---|---|---|---|---|---|
| ja-BERT | Full FT | ja-NLI | – | 83.6 | – | – | – |
| ko-RoBERTa | | ko-NLI | – | – | **83.4** | – | – |
| zh-RoBERTa | | zh-NLI | – | – | – | 71.2 | – |
| m-ST5<br>(ours) | LoRA<br>(query+ value) | XNLI | 80.3 | 81.4 | 73.4 | 73.0 | 77.3 |
| | | en-NLI | 85.6 | 82.7 | 77.2 | 77.3 | 80.7 |
| | LoRA<br>(all-linear) | XNLI | 84.2 | 82.1 | 78.0 | 74.7 | 79.8 |
| | | en-NLI | **88.1** | **84.1** | 81.1 | 79.6 | **83.2** |
| mSimCSE | Full FT | XNLI | 78.3 | 79.0 | 72.9 | 69.6 | 75.0 |
| | | en-NLI | 87.2 | 81.2 | 80.1 | **80.4** | 82.2 |
| LaBSE | – | – | 74.1 | 76.1 | 70.5 | 68.4 | 72.3 |

Table 3: Comparison with monolingual models. ‘ja’, ‘ko’ and ‘zh’ refer to Japanese, Korean and Chinese, respectively. LaBSE was evaluated using a model published on Hugging Face Hub, and mSimCSE was evaluated using a model reproduced based on the original paper.

| Model | FT Data | ar-ar | ar-en | es-es | es-en | tr-en |
|---|---|---|---|---|---|---|
| mT5-large<br>(564M) | XNLI | 62.0 | 58.2 | 77.6 | 56.8 | 53.8 |
| | en-NLI | 68.8 | 49.1 | 82.1 | 59.5 | 48.8 |
| mT5-xl<br>(1.7B) | XNLI | 71.8 | 71.7 | 81.7 | 68.0 | 66.3 |
| | en-NLI | 78.2 | 73.9 | 87.3 | 76.8 | 72.1 |
| mT5-xxl<br>(5.7B) | XNLI | 76.2 | 78.6 | 84.4 | 76.2 | 75.1 |
| | en-NLI | **83.2** | **79.5** | **87.7** | **84.9** | **79.2** |

Table 4: Comparisons of models’ performance on SemEval 2017 STS shared task when scaling up model size.

suggests that cross-lingual transferability emerges when the model becomes larger.

## 5. Conclusion

In this paper, we proposed the Multilingual Sentence T5 (m-ST5), an extension of Sentence T5 to multilingual. m-ST5 demonstrated excellent performance in multilingual tasks such as cross-lingual sentence retrieval and cross-lingual STS. It also performed well in a monolingual task, demonstrating the effectiveness of the proposed model in low-resource languages where no large-scale, high-performance model exists. Furthermore, we investigated the correlation between the size of the model’s parameters and changes in performance, confirming that performance changes follow the scaling laws and that performance improvements are particularly notable in low-resource languages. Note that we are planning to release the trained model of the proposed m-ST5.

## Ethics Statement

Since this method is an embedding model and does not generate language, the risk of generating harmful sentences would not need to be considered. On the other hand, since the biases contained in the training data are incorporated as is, the vectors generated may implicitly contain such biases, and the possibility of serious discriminatory results in some applications cannot be denied. In actual applications, maximum measures are needed to prevent such disadvantages.

## 6. Bibliographical References

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 (TACL)*, pages 597–610.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations (ICLR)*.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations (ICLR)*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017a. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, pages 1–14.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017b. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, pages 1–14.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3098–3112.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: Cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6170–6182.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2475–2485.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 657–668.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

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 (NAACL)*, pages 4171–4186.

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 (ACL)*, pages 878–891.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 422–430.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning (ICML)*, pages 4411–4421.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv:2001.08361*.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. 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 1355–1360.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*, pages 2528–2540.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 2957–2966.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations (ICLR)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Yu Meng, Chenyan Xiong, Payal Bajaj, saurabh tiwary, Paul N. Bennett, Jiawei Han, and Xia Song. 2021. COCO-LM: Correcting and contrasting text sequences for language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Niklas Muennighoff. 2022. SGPT: GPT sentence embeddings for semantic search. *arXiv:2202.08904*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics (ACL)*, pages 1864–1874.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4996–5001.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

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.

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 (EMNLP)*, pages 9122–9133.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *International Conference on Learning Representations (ICLR)*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL)*, pages 1112–1122.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-Pack: Packaged resources to advance general Chinese embedding. *arXiv:2309.07597*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng

Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4762–4772.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 483–498.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*, pages 5065–5075.

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 (ACL)*, pages 87–94.

Takumi Yoshikoshi, Daisuke Kawahara, and Sadao Kurohashi. 2020. Multilingualization of a natural language inference dataset using machine translation. In *IPSJ SIG Technical Report*, pages 1–8.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67.

## 7. Evaluate task detail

**Tatoeba-{14 and 36}:** A community-based corpus of English sentences and their translations into more than 400 languages, from which Artetxe and Schwenk (2019) generated a dataset for multilingual NLP tasks. We conduct evaluations on sentence retrieval tasks with pairs of English sentences and sentences in other languages. There are two settings: one with 14 languages and another with 36 languages.

| Data | $r$ | lr | XSTS<br>avg. | tatoeba<br>14 | tatoeba<br>36 | BUCC<br>avg. |
|---|---|---|---|---|---|---|
| XNLI | 4 | 5e-5 | **78.1** | 96.2 | 94.6 | **97.6** |
|  | 8 | 5e-5 | **78.1** | **96.3** | **94.7** | **97.6** |
|  | 16 | 1e-5 | 76.3 | 96.1 | 94.6 | 97.4 |
| en-NLI | 4 | 5e-4 | **83.0** | **94.0** | **92.9** | 96.5 |
|  | 8 | 5e-4 | 82.9 | 93.9 | **92.9** | 96.6 |
|  | 16 | 5e-4 | 82.7 | 93.8 | 92.6 | **96.8** |

Table 5: Performance when changing the rank of the LoRA adaptation matrix.

**BUCC:** It is a bitext mining task to predict translated sentences from a collection of sentences in two languages. It consists of English and one of the 4 languages (German, French, Russian and Chinese) (Zweigenbaum et al., 2017). Following XTREME (Hu et al., 2020), we regarded sentence pairs whose similarity exceeded the pre-defined threshold as translations of each other, and the results were evaluated using F-measure.

**XSTS:** The cross-lingual semantic textual similarity (XSTS) (Cer et al., 2017a), which is a multilingual extension of the vanilla STS that evaluates the correlation of the ranking of semantic similarity with human judgement. The sentence pairs of the dataset are either in the same language or different languages.

## 8. Training detail

**Batch size and number of epochs:** In our experiment, the batch size was 128. In the preliminary experiments, batch sizes of 64, 128, and 256 were considered, but no significant differences were found. Also, the number of epochs was set to 1.

**Learning rate** We used AdamW as the optimizer (Loshchilov and Hutter, 2019). We investigated the learning rate at $10^{-5}$, $5 \times 10^{-5}$, $10^{-4}$, $5 \times 10^{-4}$, and used the one that showed the best performance on the STS Benchmark (Cer et al., 2017a) dev set.

**LoRA configuration:** The rank of the LoRA adaptation matrix was $r = 8$ and the weight was $\alpha = 32$. We tried ranks $r = 4, 8,$ and $16,$ but did not observe significant differences in performance. The details of the differences are shown in Table 5.

**Model size and training data volume** Table 6 compares the model size and training data volume.

| Model | Size | Training Data |
|---|---|---|
| LASER | 0.2B | 200M pairs |
| LaBSE | 1.8B | 17B sents + 6B pairs |
| mSimCSE$_{\text{XNLI}}$ | 0.3B | 2.5TB data + 2M pairs |
| m-ST5$_{\text{XNLI}}$ | 5.7B | 6.3T tokens + 2M pairs |
| m-ST5$_{\text{en-NLI}}$ | 5.7B | 6.3T tokens + 0.2M pairs |

Table 6: Size and data datails of the models used in the experiment.

## 9. Training data detail

In contrastive learning with NLI dataset, premise-hypothesis pairs that share the same premise are concatenated to create a triplet consisting of a premise and two hypotheses (entailment and contradiction).

**en-NLI:** This training dataset is the concatenation of Stanford NLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018). Both datasets contain only English texts. The ready-to-use dataset consisting of triplets made from these datasets can be downloaded from the official repository of SimCSE[3][4]. This dataset consists of 275,601 triplets.

**XNLI:** The XNLI dataset is a crowd-sourced translation of MultiNLI dataset into 15 languages. The number of triplets included in the training data was 1,963,485.

**JSNLI:** The JSNLI dataset is a machine translation of the SNLI dataset into Japanese (Yoshikoshi et al., 2020). The train set contains around 533k premise-hypothesis pairs with labels. By the preprocesssing described above, we obtained 176,309 triplets.

## 10. Languages used in the experiment

Table 7 shows the list of languages used in the experiments in this paper. All the languages displayed in this table are included in mC4, the pre-training data for mT5 (Xue et al., 2021). Note that in the mC4 specification document, Hebrew and Tagalog (Filipino) are denoted as ‘iw’ and ‘fil’, respectively.

[3] https://github.com/princeton-nlp/SimCSE

[4] https://huggingface.co/datasets/princeton-nlp/datasets-for-simcse

| code | language | family | | Tatoeba | | BUCC | XSTS | XNLI | en-NLI | JSNLI | KorNLI | CMNLI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 14 | 36 | Table 2 | | | | | | | |
| en | English | IE / Germanic | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | |
| af | Afrikaans | IE / Germanic | | ✓ | ✓ | | | | | | | |
| am | Amharic | Semitic | | | ✓ | | | | | | | |
| ar | Arabic | Semitic | ✓ | ✓ | | | ✓ | ✓ | | | | |
| bg | Bulgarian | IE / Balto-Slavic | ✓ | ✓ | | | | ✓ | | | | |
| bn | Bengali | IE / Indo-Iranian | | ✓ | | | | | | | | |
| de | German | IE / Germanic | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | |
| el | Greek | IE / Greek | ✓ | ✓ | | | | ✓ | | | | |
| es | Spanish | IE / Italic | ✓ | ✓ | | | ✓ | ✓ | | | | |
| et | Estonian | Uralic | | ✓ | | | | | | | | |
| eu | Basque | *isolate* | | ✓ | | | | | | | | |
| fa | Persian | IE / Indo-Iranian | | ✓ | | | | | | | | |
| fi | Finnish | Uralic | | ✓ | | | | | | | | |
| fr | French | IE / Italic | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | |
| ga | Irish | IE / Celtic | | | ✓ | | | | | | | |
| he (iw) | Hebrew | Semitic | | ✓ | | | | | | | | |
| hi | Hindi | IE / Indo-Iranian | ✓ | ✓ | ✓ | | | ✓ | | | | |
| hu | Hungarian | Uralic | | ✓ | | | | | | | | |
| id | Indonesian | Austronesian | | ✓ | | | | | | | | |
| it | Italian | IE / Italic | | ✓ | | | ✓ | | | | | |
| ja | Japanese | Japonic | | ✓ | | | | | | ✓ | | |
| jv | Javanese | Austronesian | | ✓ | | | | | | | | |
| ka | Georgian | Kartvelian | | ✓ | ✓ | | | | | | | |
| kk | Kazakh | Turkic | | ✓ | | | | | | | | |
| ko | Korean | Koreanic | | ✓ | | | ✓ | | | | ✓ | |
| ml | Malayalam | Dravidian | | ✓ | | | | | | | | |
| mr | Marathi | IE / Indo-Iranian | | ✓ | | | | | | | | |
| nl | Dutch | IE / Germanic | | ✓ | | | ✓ | | | | | |
| pt | Portuguese | IE / Italic | | ✓ | | | | | | | | |
| ru | Russian | IE / Balto-Slavic | ✓ | ✓ | | ✓ | | ✓ | | | | |
| sw | Swahili | Bantu | ✓ | ✓ | ✓ | | | ✓ | | | | |
| ta | Tamil | Dravidian | | ✓ | | | | | | | | |
| te | Telugu | Dravidian | | ✓ | ✓ | | | | | | | |
| th | Thai | Kra-Dai | ✓ | ✓ | | | | ✓ | | | | |
| tl (fil) | Tagalog | Austronesian | | ✓ | ✓ | | | | | | | |
| tr | Turkish | Turkic | ✓ | ✓ | | | ✓ | ✓ | | | | |
| ur | Urdu | IE / Indo-Iranian | ✓ | ✓ | | | | ✓ | | | | |
| vi | Vietnamese | Austroasiatic | ✓ | ✓ | | | | ✓ | | | | |
| zh | Chinese | Sino-Tibetan | ✓ | ✓ | | ✓ | ✓ | ✓ | | | | ✓ |

Table 7: List of languages used in the experiments.