Cross-lingual Transfer or Machine Translation? On Data Augmentation for Monolingual Semantic Textual Similarity

Sho Hoshino, Akihiko Kato, Soichiro Murakami, Peinan Zhang

CyberAgent

{hoshino_sho, kato_akihiko, murakami_soichiro, zhang_peinan}@cyberagent.co.jp

Abstract

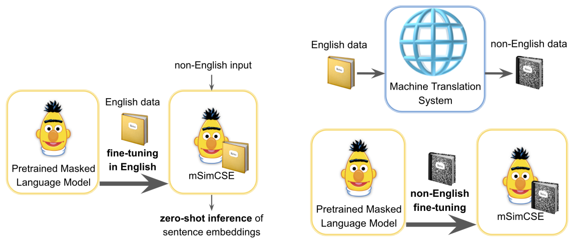
Learning better sentence embeddings leads to improved performance for natural language understanding tasks including semantic textual similarity (STS) and natural language inference (NLI). As prior studies leverage large-scale labeled NLI datasets for fine-tuning masked language models to yield sentence embeddings, task performance for languages other than English is often left behind. In this study, we directly compared two data augmentation techniques as potential solutions for monolingual STS: (a) *cross-lingual transfer* that exploits English resources alone as training data to yield non-English sentence embeddings as zero-shot inference, and (b) *machine translation* that coverts English data into pseudo non-English training data in advance. In our experiments on monolingual STS in Japanese and Korean, we find that the two data techniques yield performance on par. Rather, we find a superiority of the Wikipedia domain over the NLI domain for these languages, in contrast to prior studies that focused on NLI as training data. Combining our findings, we demonstrate that the cross-lingual transfer of Wikipedia data exhibits improved performance, and that native Wikipedia data can further improve performance for monolingual STS.

Keywords: sentence embeddings, cross-lingual transfer, machine translation

1. Introduction

Monolingual semantic textual similarity (STS; Agirre et al., 2016) has been used as a progress milestone for the learning of sentence embeddings (Reimers and Gurevych, 2019; Gao et al., 2021). Given two sentences in a target language, our task was to predict the similarity between the two sentences. Monolingual STS is a core part of natural language understanding tasks (Wang et al., 2018) and is related to natural language inference (NLI; Dagan et al., 2005) where the task was to predict whether one sentence entails another. NLI data have been used for unsupervised STS without using any STS-specific data and are therefore considered to be suitable training data for monolingual STS.

Because such a quantity of training data is not always available in languages other than English, data augmentation techniques including (a) *cross-lingual transfer* (Conneau et al., 2018; Gogoulou et al., 2022; Wang et al., 2022) and (b) *machine translation* (Conneau et al., 2018; Ham et al., 2020; Yanaka and Mineshima, 2022) have been extensively applied. However, aside from the monumental prior study on NLI (Conneau et al., 2018), there have been few to no comprehensive studies that directly compared the two different data augmentation techniques particularly on monolingual STS, which differs from multilingual STS (Cer et al., 2017) in that the given two sentences are in the same language. Therefore, we investigated the following research questions to search for the most suitable data augmentation technique for monolingual STS, which would have the greatest need for monolingual data: **RQ1** *Between cross-lingual transfer or machine translation, which is better?* **RQ2** *Do these approaches yield performance on par with that of a state-of-the-art multilingual model?*



(a) cross-lingual transfer (b) machine translation

Figure 1: Illustration of the two different data augmentation techniques applied from English to non-English.

In this study, we empirically evaluated the two different data augmentation techniques using Japanese and Korean, which can be seen as relatively low-resourced and linguistically dissimilar languages when compared with English, and thus are challenging for the two techniques. We used unsupervised multilingual SimCSE (mSimCSE; Wang et al., 2022) as our test bed, which is a multilingual extension of unsupervised SimCSE (Gao et al., 2021) that uses sentence-pair contrastive learning for self-supervised learning of unlabeled data.

Specifically, we trained mSimCSE models in two ways, each corresponding to the two data augmentation techniques: (a) by using English training data alone as cross-lingual transfer, and (b) by using

LREC-COLING 2024, pages 4164–4173

machine-translated data from English into Korean and Japanese. We also compared trained mSimCSE models with LaBSE (Feng et al., 2022), a state-of-the-art multilingual model.

In our experiments on monolingual STS, we demonstrate that cross-lingual transfer achieves performance on par with that of machine translation (**RQ1**). We further demonstrate that the combination of cross-lingual transfer and Wikipedia domain data exhibits the best performance outperforming or comparable to that of LaBSE (**RQ2**). In contrast to prior studies, we also observed that Wikipedia domain data can be used as an alternative drop-in replacement for NLI domain data when used as unlabeled training data for monolingual STS.

Since Wikipedia domain data is easier to obtain than NLI domain data, our results suggest a recipe for training better sentence embeddings using a large-scale multilingual Wikipedia dataset. As such a pilot study, we actually report improved results using a Japanese portion of the Wikipedia data.

2. Data Augmentation Techniques

Data augmentation (Feng et al., 2021) is a generic strategy to deal with relatively low-resourced situations in a target language to increase the number of training examples (Figure 1). We explain two different data augmentation techniques as follows.

Cross-lingual transfer leverages training data in other languages but not in the target language. We conduct fine-tuning of pretrained masked language models using the available data, without leveraging target language resources.[1] After that, we use the trained model to perform zero-shot inference in the target language. This cost-effective approach has virtually no cost for obtaining data. Conneau et al. (2018) studied the performance of such cross-lingual transfer from English to other languages using a multilingual NLI dataset (XNLI).

Machine translation leverages the same data, but this time by using a machine translation system to translate it from other languages into the target language. We then perform fine-tuning of pretrained masked language models using the machine-translated data. After that, we use the trained model to perform normal inference in the target language, unlike cross-lingual transfer. This pay-as-you-go approach has two variations. Conneau et al. (2018) created the training portion of the XNLI dataset using a neural machine translation system (TRANSLATE TRAIN). They also attained comparable results by translating test data only at runtime (TRANSLATE TEST).

[1] In this context, the pretrained models must be capable of processing a target language and its vocabulary, unlike cross-lingual transfer to unseen languages (Artetxe et al., 2020).

Hyperparameter	Values
# Epochs	1
Learning rate	5e-5
Warmup rate	10%
Batch size	{32, 64, 128}

Table 1: Hyperparameters.

3. Experiments

We compared the performance of cross-lingual transfer and machine translation on unsupervised learning for monolingual STS (Agirre et al., 2016). Given two sentences in a target language, our task was to predict the similarity between the two sentences using unlabeled training data. This task differs from supervised learning that exploits labeled data, because our focus is on the relatively low-resourced situations in Japanese and Korean. The task also differs from multilingual STS in that the given two sentences are in the same language.

3.1. Setup

More specifically, we investigated the performance of sentence embeddings obtained as a result of unsupervised learning for monolingual STS. Following mSimCSE (Wang et al., 2022), we fine-tuned multilingual pretrained masked language models by feeding various training data as different data augmentation techniques. Since XLM-R (Conneau et al., 2020) is used as a common baseline in the relevant literature, we used XLM-R as the multilingual base model and performed fine-tuning using unsupervised SimCSE (Gao et al., 2021). We compared our models with LaBSE (Feng et al., 2022) without fine-tuning as a strong fully supervised baseline.

We used random hardware available at Google Colaboratory[2] and Sentence-Transformers 2.2.0 (Reimers and Gurevych, 2019) for our implementation.[3] We followed the hyperparameters used in Sentence-Transformers as best practices, unless listed in Table 1.

3.2. Datasets

Training Table 2 summarizes various datasets in English, Japanese, and Korean that are primarily used as our training data. Following prior studies on monolingual STS (Reimers and Gurevych, 2019; Gao et al., 2021), we used NLI datasets that contain premises, hypotheses, and their relationship labels such as entailment, contradiction, and neutral (Dagan et al., 2005). Specifically, we used premises

[2] https://colab.research.google.com/

[3] The details of our implementation are described in the Appendix A, followed by the replication study of unsupervised SimCSE reported in the Appendix D.

Name	Size	Domain	Lang.	Sources
SNLI	0.5M	NLI	en	
MNLI	0.5M	NLI	en	
JSNLI	0.5M	NLI	MT_{ja}	SNLI
KorNLI	1.0M	NLI	MT_{ko}	SNLI, MNLI
Wiki	1.0M	Wikipedia	en	
Wiki-40B	1.0M	Wikipedia	ja	

Table 2: The statistics of the various training datasets we used. The size column reports the approximate number of training examples. MT denotes pseudo data machine-translated from the corresponding resources shown in the last column.

Models	Japanese JSICK-STS	Japanese JGLUE-JSTS	Korean KorSTS	Korean KLUE-STS
Cross-lingual transfer				
$\text{mSimCSE}_{\text{en}}$				
+ NLI	79.55	74.12	74.68	72.87
+ SNLI	**80.31**	74.73	75.21	65.83
Machine translation				
$\text{mSimCSE}_{\text{MT}_{\text{ja}}}$				
+ JSNLI	78.41	**75.55**		
$\text{mSimCSE}_{\text{MT}_{\text{kr}}}$				
+ KorNLI			**75.43**	**74.45**

Table 3: Comparisons of cross-lingual transfer and machine translation as data augmentation. We report Spearman's correlation [%]. Higher is better.

Models	English STS-B
Unsupervised	
$\text{RoBERTa}_{\text{large}}$	56.29
SimCSE	
+ Wiki	**85.95**

(a) Monolingual models.

Models	English STS-B
Unsupervised	
$\text{XLM-R}_{\text{large}}$	43.54
$\text{mSimCSE}_{\text{en}}$	
+ Wiki	**83.54**
+ NLI	76.98

(b) Multilingual models.

Table 4: Comparisons of different data domains.

and hypotheses as training examples of unsupervised learning, while simply discarding NLI labels. We also used English and Japanese Wikipedia.[4]

For fine-tuning in English, we used the Stanford Natural Language Inference corpus (SNLI; Bowman et al., 2015) and the Multi-Genre Natural Language Inference corpus (MNLI; Williams et al., 2018) as well as English Wikipedia (Wiki; Gao et al., 2021). To match the data size of NLI datasets with its Wikipedia counterpart (1 million sentences), we used them in two ways[5]: (i) only the premises portions of SNLI and MNLI (**NLI**; 942,854 sentences), and (ii) both the premise and hypothesis portions of SNLI (**SNLI**; 1,100,304 sentences).

For fine-tuning in Japanese and Korean, machine-translated English datasets are used instead. For instance, the JSNLI dataset (Yoshikoshi et al., 2020) contains machine-translated SNLI and the KorNLI dataset (Ham et al., 2020) contains machine-translated SNLI and MNLI, respectively. We also used Japanese Wikipedia data created from multilingual Wikipedia (Wiki-40B; Guo et al., 2020).

Evaluation We used the following labeled STS datasets as our evaluation data: STS-B (Cer et al., 2017) for English; JSICK-STS (Yanaka and Mineshima, 2022), a human-translated SICK (Marelli et al., 2014), and JGLUE-JSTS (Kurihara et al., 2022) for Japanese; and KorSTS (Ham et al., 2020), a human-translated STS-B (Cer et al., 2017), and KLUE-STS (Park et al., 2021) for Korean.

3.3. Evaluation Protocol

Following Gao et al. (2021), we evaluate sentence embeddings by measuring the correlation between the human labels $[0, 5]$ and the cosine distance between two given sentences, using Spearman's rank correlation coefficient (Spearman, 1904). However, instead of using test portions of the labeled STS datasets, we used development portions as our evaluation data to avoid the unnecessary overfitting to English development data (Keung et al., 2020), which would caused by applying the standard evaluation protocol to cross-lingual transfer experiments.

3.4. Results

Table 3 summarizes the comparisons of the two data augmentation techniques under a controlled setting within the same NLI domain. In this fair setting, machine translation ($\text{mSimCSE}_{\text{MT}}$) slightly outperformed cross-lingual transfer ($\text{mSimCSE}_{\text{en}}$). Specifically, in the case of Korean, $\text{mSimCSE}_{\text{MT}_{\text{kr}}}$ almost outperformed $\text{mSimCSE}_{\text{en}}$. Similarly, in the case of Japanese, $\text{mSimCSE}_{\text{MT}_{\text{ja}}}$ outperformed $\text{mSimCSE}_{\text{en}}$ with the exception on JSICK-STS.

However, we are yet to conclude that machine translation is better than cross-lingual transfer. Table 4 summarizes the comparisons between NLI and Wikipedia domains in English. Unlike prior studies focusing on NLI as training data, surprisingly, we obtained different results in this domain-aware setting. Specifically, the $\text{mSimCSE}_{\text{en}}$ model trained using Wikipedia data outperformed its counterpart using NLI data, suggesting a superiority of

[4]The details of data preprocessing are described in the Appendix B.

[5]An ablation study with different data sizes is reported in the Appendix E.

Models	Japanese JSICK-STS	JGLUE-JSTS	Korean KorSTS	KLUE-STS
Unsupervised				
XLM-R_{large}	61.87	54.19	49.26	20.13
mSimCSE$_{en}$				
+ Wiki	**81.02**	**77.62**	**80.38**	81.42
+ SNLI	80.31	74.73	75.21	65.83
Fully supervised				
LaBSE	76.77	76.12	73.01	**82.81**

Table 5: Comparisons of the best-performing cross-lingual transfer of Wikipedia data and LaBSE.

Models	English STS-B	Japanese JSICK-STS	JGLUE-JSTS
Cross-lingual transfer			
mSimCSE$_{en}$			
+ Wiki	**83.54**	81.02	77.62
Native data			
mSimCSE$_{ja}$			
+ Wiki-40B	81.89	**81.05**	**78.71**

Table 6: Comparisons of English Wikipedia data cross-lingual transfer and native Japanese data.

Wikipedia as unlabeled training data.

These new results led us to the combination of our findings, namely the cross-lingual transfer of Wikipedia domain data. Table 5 summarizes the comparisons involving both the Wikipedia and NLI domains. In this best-performing setting following Gao et al. (2021), the cross-lingual transfer of Wikpedia outperformed the machine translation of NLI, resulting in performance almost outperforming that of LaBSE. Specifically, the mSimCSE$_{en}$ model trained using Wikipedia data outperformed LaBSE with the only exception on KLUE-STS.[6]

4. Discussion

Between cross-lingual transfer or machine translation, which is better? It depends. On one hand, machine translation can outperform cross-lingual transfer if we used the same NLI domain as training data. On the other hand, the cross-lingual transfer of Wikipedia domain data outperformed the machine-translated NLI domain data. These results rather suggest the effectiveness of Wikipedia as unlabeled training data.

Do these approaches yield performance on par with that of a state-of-the-art multilingual model? Yes, we found that the cross-lingual transfer of Wikipedia data can outperform LaBSE. This posed us a few more questions: Should we pursue the direction of English as training data proxy, similar to mSimCSE? Are there some benefit from using native multilingual data, if we could create it without using machine translation, similar to LaBSE?

Table 6 summarizes additional results using native Japanese Wikipedia data. This pilot study suggests that using native data can indeed improve performance over cross-lingual transfer of English data. However, we are yet to answer the ultimate question of English as training data proxy.

[6]We also performed an analysis of the obtained sentence embeddings using KLUE-STS, which is reported in the Appendix C

5. Related Work

Learning Sentence Embeddings Sentence embeddings are learned representations of sentences within a single dense matrix, unlike word embeddings, which are represented in multiple dense matrices. Several methods for fine-tuning pretrained masked language models have been proposed, including (a) SBERT (Reimers and Gurevych, 2019) using sentence-pair regression for supervised learning of labeled STS data, and (b) unsupervised SimCSE (Gao et al., 2021) using sentence-pair contrastive learning for self-supervised learning of unlabeled data. Both methods yield good sentence embeddings in terms of cosine distance in STS.

Cross-lingual Transferability of Multilingual Models Artetxe and Schwenk (2019) proposed LASER, language-agnostic representations of sentence embeddings learned from dedicated parallel data. They studied cross-lingual transferability of LASER on XNLI (Conneau et al., 2018) and found improved performance in various languages. Feng et al. (2022) proposed LaBSE, which outperformed LASER in downstream tasks by using additional monolingual data with parallel data. Wang et al. (2022) proposed mSimCSE, a multilingual extension of SimCSE, which is perhaps the most related work. They investigated cross-lingual transferability of mSimCSE by using various tasks including multilingual STS.

Cross-lingual Transferability of Monolingual Models Artetxe et al. (2020) studied cross-lingual transferability from one language to unseen languages. They showed that the transfer learning of monolingual BERT (Devlin et al., 2019) at the lexical level outperformed multilingual BERT on XQuAD. Reimers and Gurevych (2020) studied transfer learning of SBERT at the sentence level, using multilingual knowledge distillation in English and Korean as monolingual and multilingual STS. They observed performance better than that of monolingual fine-tuning in Korean. Gogoulou et al. (2022) studied cross-lingual transferability from non-English languages into English by using various GLUE tasks (Wang et al., 2018) including monolingual STS.

6. Conclusion

In this study, we empirically compared cross-lingual transfer and machine translation in terms of monolingual STS performance. We chose Japanese and Korean as our test bed, because these languages are relatively low-resourced and linguistically dissimilar compared with English, and thus are challenging for the two data augmentation techniques. We found that the cross-lingual transfer exhibits performance comparable to that of the machine translation depending on data domain. We also found that, in contrast to prior studies, cross-lingual transfer of Wikipedia data achieved the best performance, outperforming or comparable to that of the state-of-the-art LaBSE. Our future work will include fine-grained analysis of which types of data are better suitable for monolingual STS.

Limitations

Our study focuses primarily on Japanese and Korean, which are often considered high-resource or mid-resource languages, but, at the same time, are relatively low-resourced with respect to STS training data and therefore suitable for data augmentation. For this reason, our results are not directly applicable to other low-resource and regional languages, as even human-labeled STS evaluation data are lacking. In addition, we did not conduct experiments using Wikipedia data machine-translated from English into Japanese and Korean, which will be part of our future work.

7. Bibliographical References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 497–511.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

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.

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*, pages 632–642.

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

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*, pages 8440–8451.

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*, pages 2475–2485.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the 1st PASCAL Machine Learning Challenges Workshop*, pages 177–190.

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*, 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*, pages 878–891.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

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*, pages 6894–6910.

Evangelia Gogoulou, Ariel Ekgren, Tim Isbister, and Magnus Sahlgren. 2022. Cross-lingual transfer of monolingual models. In *Proceedings of the 13th International Conference on Language Resources and Evaluation*, pages 948–955.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452.

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 2020*, pages 422–430.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 549–554.

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.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 2957–2966.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. Preprint, arXiv:1907.11692.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 216–223.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook 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, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

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*, 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*, pages 4512–4525.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*, pages 353–355.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9929–9939.

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.

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*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

Takumi Yoshikoshi, Daisuke Kawahara, and Sadao Kurohashi. 2020. Multilingualization of a natural language inference dataset using machine translation. *IPSJ SIG Technical Report*, NL-244(6):1–8. In Japanese.

A. Implementation Details

In this study, we reimplemented the unsupervised SimCSE (Gao et al., 2021) instead of utilizing the original implementation provided by the authors[7]. We used Sentence-Transformers 2.2.0 (Reimers and Gurevych, 2019), Hugging Face Transformers 4.18.0 (Wolf et al., 2020), and SentencePiece 0.1.96 (Kudo and Richardson, 2018). We also used the publicly available models of XLM-R[8], LaBSE[9], RoBERTa[10], KLUE-RoBERTa[11], and SBERT[12], as well as the unsupervised SimCSE models distributed by the original authors[13]. We report results obtained in a single run with the random seed fixed, while freezing hyperparameters for the sake of reproducibility (Keung et al., 2020).

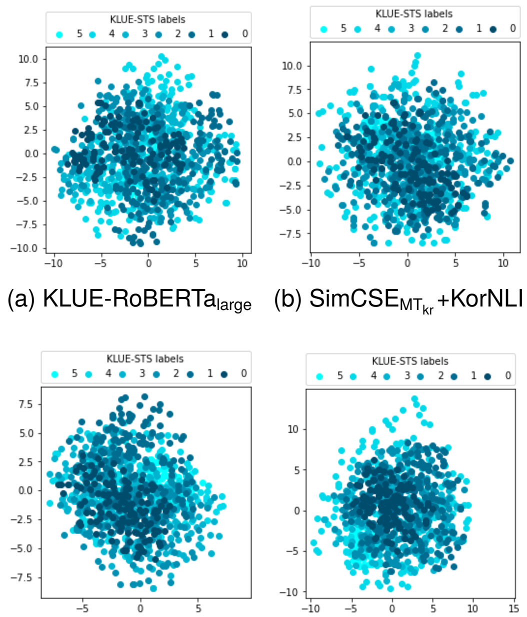
B. Data Preprocessing

Here, we describe the details of our data preprocessing to make the data used in this study reproducible. We have removed empty lines from our NLI and Wikipedia datasets and applied additional preprocessing as follows.

NLI datasets For JSNLI, we roughly detokenized the already-tokenized dataset by applying NFKC normalization and eliminating white spaces. For JSICK-STS, we used its test portion to make it comparable with the other datasets.

Wikipedia datasets For English Wikipedia, we used the data[14] distributed with the original implementation of SimCSE (Gao et al., 2021), which contains exactly 1 million sentences but the authors did not report its details.

Therefore, we created our own Japanese Wikipedia data from the multilingual Wiki-40B dataset[15] (Guo et al., 2020) as a replication study. Specifically, we extracted Japanese paragraphs from Wiki-40B and applied sentence splitting using new line symbols and Japanese period markers. To match the data size of the Japanese Wikipedia data with its English counterpart, we used only the first 1 million sentences.



(a) KLUE-RoBERTa$_{\text{large}}$ (b) SimCSE$_{\text{MT}_{\text{kr}}}$+KorNLI

(c) XLM-R$_{\text{large}}$ (d) mSimCSE$_{\text{en}}$ + Wiki

Figure 2: 2D visualization of sentence embeddings on KLUE-STS. Different scales are used to better illustrate the isotropic nature of SimCSE fine-tuning.

C. Analysis of Sentence Embeddings

We performed an analysis of the obtained sentence embeddings using visualization and metrics.

Visualization Figure 2 provides a visualization of the sentence embeddings on KLUE-STS, which were obtained by applying t-distributed stochastic neighbor embedding (t-SNE; van der Maaten and Hinton, 2008) for two-dimensional reduction. We compared cross-lingual transfer (2d) with a Korean model (2b), which was fine-tuned using KLUE-RoBERTa (Liu et al., 2019; Park et al., 2021), unsupervised SimCSE (Gao et al., 2021), and KorNLI. We observe that cross-lingual transfer can maintain the isotropic nature of SimCSE fine-tuning.

Alignment and Uniformity We further confirmed the aforementioned trend by using alignment and uniformity metrics (Wang and Isola, 2020) to quantify the isotropic nature of SimCSE fine-tuning. The monolingual model (2b) scored $\ell_{\text{align}} = 0.3540$ and $\ell_{\text{uniform}} = -3.3486$, while cross-lingual transfer (2d) scored similar values of $\ell_{\text{align}} = 0.2684$ and $\ell_{\text{uniform}} = -3.2484$.

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

[8] https://huggingface.co/xlm-roberta-large

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

[10] https://huggingface.co/roberta-large

[11] https://huggingface.co/klue/roberta-large

[12] https://huggingface.co/sentence-transformers/roberta-large-nli-mean-tokens

[13] https://huggingface.co/princeton-nlp/unsup-simcse-roberta-large

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

[15] https://huggingface.co/datasets/wiki40b

Models	English STS-B
Unsupervised	
$\text{RoBERTa}_{\text{base}}$	64.70
$\text{SimCSE}_{\text{orig}}$ + Wiki	84.26
$\text{SimCSE}_{\text{repro}}$ + Wiki	83.90
$\text{RoBERTa}_{\text{large}}$	56.29
$\text{SimCSE}_{\text{orig}}$ + Wiki	85.52
$\text{SimCSE}_{\text{repro}}$ + Wiki	**85.95**
English NLI Supervised	
$\text{SBERT}_{\text{base}}$ + NLI	80.73
$\text{SBERT}_{\text{large}}$ + NLI	82.51

(a) Monolingual models.

Models	English STS-B
Unsupervised	
$\text{XLM-R}_{\text{base}}$	53.62
$\text{mSimCSE}_{\text{en}}$ + Wiki	76.44
+ NLI	75.79
$\text{XLM-R}_{\text{large}}$	43.54
$\text{mSimCSE}_{\text{en}}$ + Wiki	**83.54**
+ NLI	76.98
Fully supervised	
LaBSE	74.13

(b) Multilingual models.

Table 7: Comparison of unsupervised SimCSE in English on different training data and base models. $\text{SimCSE}_{\text{orig}}$ denotes the models distributed by the original authors, whereas $\text{SimCSE}_{\text{repro}}$ and mSimCSE denote our implementation used in this study.

D. Replication Study

We conducted a replication study of unsupervised SimCSE in English (Gao et al., 2021) by using various training data that have been extensively used in the relevant literature but not directly compared in the unsupervised SimCSE setup. We used RoBERTa (Liu et al., 2019) as the monolingual base model and performed monolingual fine-tuning using unsupervised SimCSE.

As summarized in Table 7, our implementation ($\text{SimCSE}_{\text{repro}}$ and mSimCSE) exhibited performance comparable to that of the models distributed by the original authors ($\text{SimCSE}_{\text{orig}}$).

E. Ablation Study

We conducted an ablation study on different sizes of training data in English, Japanese, and Korean. In this comparison, we used NLI datasets, combining them in two ways: (a) only the premise portion, and (b) both the premise and hypothesis portions. As a result, our data sizes varied from roughly 0.5M examples to up to 2.0M examples.

Table 8 summarizes the results of the ablation study. Our findings are twofold. (i) There was a minimum practical size of 1.0M examples. All models trained using lower sizes suffered serious performance degradation. (ii) We also observed that when the data size was already close to 1.0M examples, the case using the premise portion alone (8a) outperformed that using both the premise and hypothesis (8b). This result is convincing, as the hypothesis portion is artificially created from the premise portion (Dagan et al., 2005) and there is not much variation in the relation between them.

Models	Size	English STS-B	Japanese JGLUE-JSTS	Korean KorSTS	KLUE-STS
			Unsupervised		
XLM-R_{base}		53.62	59.28	57.98	30.69
mSimCSE$_{en}$					
+ SNLI	0.5M	72.61	71.56	69.52	52.13
+ NLI	1.0M	75.79	71.83	73.61	60.66
mSimCSE$_{MT_{ja}}$					
+ JSNLI	0.5M	72.87	71.16	73.48	63.69
mSimCSE$_{MT_{kr}}$					
+ KorNLI	1.0M	75.39	73.02	73.34	67.13
XLM-R_{large}		43.54	54.19	49.26	20.13
mSimCSE$_{en}$					
+ NLI	1.0M	78.66	74.12	74.68	72.87
mSimCSE$_{MT_{kr}}$					
+ KorNLI	1.0M	**80.29**	74.93	**75.43**	**74.45**

(a) The premise alone.

Models	Size	English STS-B	Japanese JGLUE-JSTS	Korean KorSTS	KLUE-STS
			Unsupervised		
XLM-R_{base}		53.62	59.28	57.98	30.69
mSimCSE$_{en}$					
+ SNLI	1.0M	75.03	72.69	73.71	58.75
+ NLI	2.0M	75.83	73.12	74.18	61.75
mSimCSE$_{MT_{ja}}$					
+ JSNLI	1.0M	75.26	71.69	74.20	65.39
XLM-R_{large}		43.54	54.19	49.26	20.13
mSimCSE$_{en}$					
+ SNLI	1.0M	75.36	74.73	75.21	65.83
+ NLI	2.0M	76.87	75.19	74.19	70.60
mSimCSE$_{MT_{ja}}$					
+ JSNLI	1.0M	79.40	**75.55**	75.21	73.97

(b) Both the premise and hypothesis.

Table 8: Ablation study on different data sizes, data combinations, and base models. **Boldface** only highlights the best performance values over the two combinations of the premise alone and both the premise and hypothesis.