How Can Multimodal Data Improve Low-Resource Language Sentence Embeddings? A Case Study on Tamil and Minangkabau

Howard Tangkulung, Kaiyan Zhao, Yoshimasa Tsuruoka
The University of Tokyo
{howard,kaiyan1006,tsuruoka}@logos.t.u-tokyo.ac.jp

Abstract

The training of multilingual sentence embedding models typically relies on large-scale parallel text corpora, which are often scarce for low-resource languages. Recent works have shown that multimodal data can supplement existing parallel text data or even replace the need for translated text by solely relying on image captioning. This paper explores the use of multimodal data for sentence embeddings on three practical scenarios: *Parallel* (direct translations), *Semi-Parallel* (different captions of the same image), and *Pseudo-Parallel* (single caption per image) on two low-resource languages: Tamil and Minangkabau. Our results demonstrate that incorporating multimodal alignment significantly improves semantic textual similarity (STS) tasks but has a limited impact on bi-text retrieval tasks. Furthermore, our results show that the *Semi-Parallel* scenario, which does not rely on translated texts, can be a cost-effective alternative when parallel text data is unavailable.

1 Introduction

Sentence embeddings are a fundamental building block in natural language processing (NLP). Multilingual sentence embedding models map sentences to fixed-dimensional vectors that represent their semantic meanings, regardless of the input language [1]. Similar sentences are mapped to nearby points and dissimilar sentences to distant points in the embedding space. These embeddings can then be used in various downstream tasks, such as multilingual document retrieval [2], clustering [3], and unsupervised machine translation [4].

Typically, a multilingual sentence embedding model is trained on large corpora of multilingual text data using masked language modeling (MLM) objectives [5, 6], then further fine-tuned on supervised multilingual datasets using contrastive learning objectives [1, 7]. This approach works well for high-resource languages with abundant training data, but is less practical for low-resource languages with limited data. To address this, this paper explores the use of multimodal data and multimodal alignment on three scenarios (Figure 1):

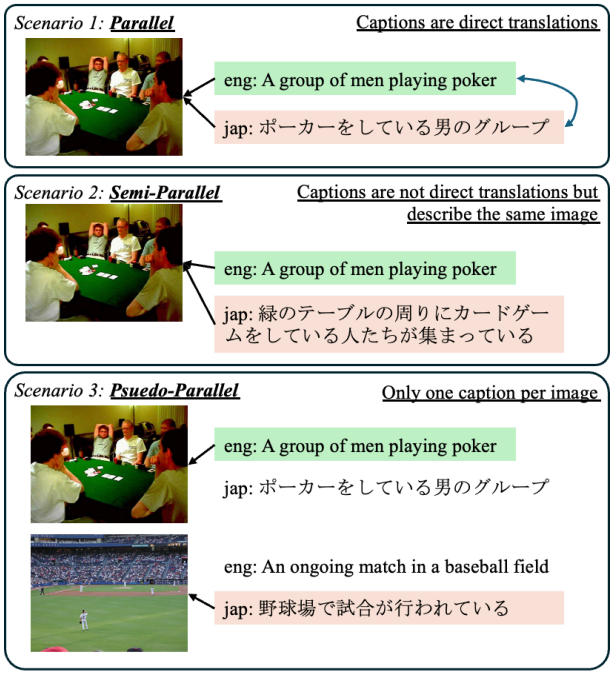

Figure 1 Three scenarios for leveraging multimodal data to improve low-resource language sentence embeddings.

1. **Parallel**: Directly translated caption pairs.
2. **Semi-Parallel**: Caption pairs to the same image, but not direct translations.
3. **Pseudo-Parallel**: Only one caption per image.

In scenario 1 (**Parallel**), we have access to directly translated parallel text data for low-resource languages. In scenario 2 (**Semi-Parallel**), we have access to caption pairs to

the same image, but these caption pairs are *not* direct translations of each other. In scenario 3 (**Pseudo-Parallel**), we only have access to one caption per image, with no parallel text data. Scenario 1 is important to investigate the possibility of squeezing more performance out of existing parallel text data using multimodal alignment. Scenarios 2 and 3 do not require translated captions and thus do not require bilingual experts or translation tools. This paper simulates these two scenarios using low-resource languages accessible through machine translation, but an actual application would involve curating image captions from monolingual speakers where machine translation is not available and bilingual experts are expensive.

This paper investigates these three scenarios, particularly in two low-resource languages: Tamil and Minangkabau. Tamil is a well-studied low-resource language in NLP with an existing STS benchmark [8], and included in the pretraining of XLM-RoBERTa [6]. On the other hand, Minangkabau is a regional language spoken in Indonesia and not included in the pretraining of XLM-RoBERTa [6]. This paper aims to provide insights into the differences in the feasibility of each scenario for low-resource languages that are included and not included in the pretraining of the base multilingual model.

To summarize, this paper (1) explores three practical scenarios for leveraging multimodal data for low-resource language sentence embeddings, (2) compares the effectiveness of each scenario on low-resource languages included and not included in the pretraining, and (3) shows that multimodal alignment improves STS tasks but has less significant impact on bi-text retrieval tasks.

2 Related Work

2.1 Multilingual Sentence Embeddings

Popular multilingual sentence embedding models such as LaBSE [7] and mSimCSE [1] require large amounts of parallel text data for supervised contrastive learning. LaBSE [7] performs translation language modeling (TLM), which extends the MLM objective to translation pairs, on 6B translation pairs from 109 languages. mSimCSE [1] adapts XLM-RoBERTa [6] in a contrastive learning objective on 2.2M triples of cross-lingual natural language inference (XNLI) [9] data. These two approaches perform very well on tasks such as bi-text retrieval and semantic textual similarity (STS) of languages included in their training data. However, training on large amounts of parallel text data as described above is not feasible for some low-resource languages.

2.2 Multimodal Alignment for Sentence Embeddings

MCSE [10] uses SimCSE [11] as a textual baseline and further adds a multimodal alignment objective on English image-caption pairs. MCSE uses 80k image-caption pairs from the MS-COCO dataset [12] and achieves strong performance on English STS tasks. The multimodal alignment here happens in a shared multimodal embedding space where an image encoder's (ViT [13]) and a text encoder's (BERT-base [14]) `[CLS]` embeddings are projected using a linear layer and trained using contrastive learning. This multimodal alignment technique will be the basis for our experiments in scenario 1 (Parallel) and scenario 2 (Semi-Parallel), where there is one caption pair per image.

Krasner et al. [15] investigate the use of multimodal alignment to introduce a new language (Quechua) to an existing multilingual sentence embedding model (XLM-RoBERTa). On the MS-COCO dataset [12], Krasner et al. [15] use contrastive learning to align a caption to its corresponding image in a shared multimodal embedding space. This case study is similar to this paper's scenario 3 (Pseudo-Parallel), with only one caption per image.

This paper differs from these prior works by focusing on the comparison of the three scenarios (**Parallel**, **Semi-Parallel**, and **Pseudo-Parallel**) for using multimodal data and multimodal alignment. This paper also evaluates two low-resource languages (Tamil and Minangkabau) to compare effectiveness on different types of low-resource languages (included and not included in the pretraining) in each scenario.

3 Method

Following the work of Zhang et al. [10], this paper uses a contrastive learning objective to align image and text embeddings in a shared multimodal embedding space, and uses the InfoNCE loss [16] as the training objective. For scenarios 1 (Parallel) and 2 (Semi-Parallel), the objective loss is calculated using a set of image and caption pairs $D^m = \{(x_i^{l_1}, x_i^{l_2}, y_i)\}_{i=1}^{N}$, where $x_i^{l_1}$ and $x_i^{l_2}$ are the captions in two different languages for image y_i. The caption is encoded using the same text encoder f_θ and projection

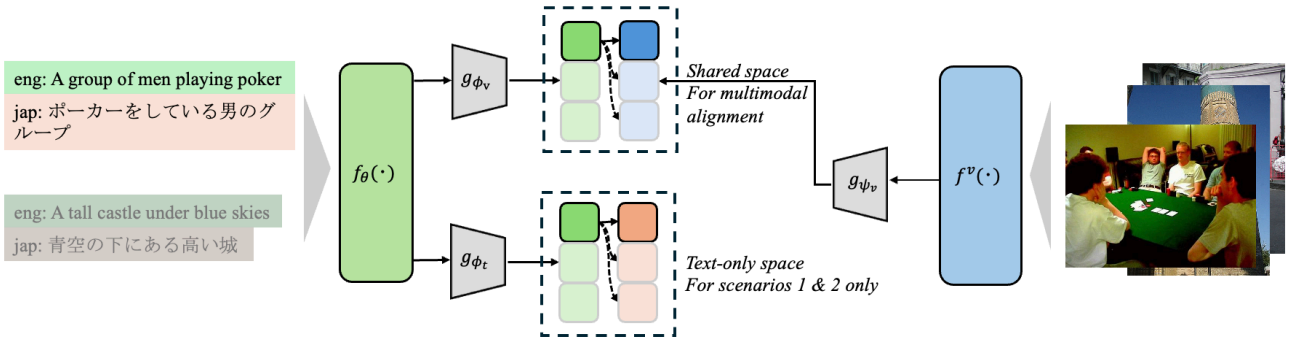

Figure 2 Contrastive learning implementation for text-only and multimodal losses.

head g_{ϕ_t} for the text-only embedding space and a different projection head g_{ϕ_v} for the shared multimodal embedding space. The image is encoded using a pretrained image encoder f^v and a separate projection head g_ψ:

$$\mathbf{s}^l_{i,text} = g_{\phi_t}(f_\theta(x^l_i)), \tag{1}$$

$$\mathbf{s}^l_{i,shared} = g_{\phi_v}(f_\theta(x^l_i)), \tag{2}$$

$$\mathbf{v}_i = g_\psi(f^v(y_i)), \tag{3}$$

where $l \in \{l_1, l_2\}$, $\mathbf{s}^l_{i,text}$ is the text embedding in the text-only embedding space, $\mathbf{s}^l_{i,shared}$ and $\mathbf{v}_i$ is the text and image embeddings in the shared multimodal embedding space, respectively. The multimodal contrastive loss is calculated as:

$$\ell^M_i = -\frac{1}{2}\sum_{l\in\{l_1,l_2\}} \log \frac{\exp(\text{sim}(\mathbf{s}^l_{i,shared}, \mathbf{v}_i)/\tau')}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{s}^l_{i,shared}, \mathbf{v}_j)/\tau')}. \tag{4}$$

The text-only contrastive loss is calculated as:

$$\ell^S_i = -\log \frac{\exp(\text{sim}(\mathbf{s}^{l_1}_{i,text}, \mathbf{s}^{l_2}_{i,text})/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{s}^{l_1}_{i,text}, \mathbf{s}^{l_2}_{j,text})/\tau)}. \tag{5}$$

The final loss is calculated as the weighted sum of the two losses:

$$\ell_i = \lambda\ell^M_i + \ell^S_i, \tag{6}$$

For scenario 3 (Pseudo-Parallel), there is only one caption per image. Using a set of image and caption pairs $D^p = \{(x^l_i, y_i)\}^N_{i=1}$, where x^l_i is the caption in language l for image y_i, the multimodal contrastive loss is calculated as:

$$\ell^M_i = -\log \frac{\exp(\text{sim}(\mathbf{s}^l_{i,shared}, \mathbf{v}_i)/\tau')}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{s}^l_{i,shared}, \mathbf{v}_j)/\tau')}, \tag{7}$$

which serves as the final loss in scenario 3.

4 Experiments

4.1 Dataset

This paper uses the MS-COCO dataset [12] in the Karpathy split [17] as the source of English image-caption pairs. Each image in MS-COCO has 5 English captions, which are then translated using Google Translate to Spanish, Thai, and Japanese (high-resource languages), and to Tamil and Minangkabau (low-resource languages). A total of 113k image-caption pairs are used for training, with 5k pairs for validation and 5k pairs for testing. Caption pairs for scenario 1 (Parallel) are translations of the same original English caption. Caption pairs for scenario 2 (Semi-Parallel) are also translations from English captions, but each comes from different original English captions for the same image.

4.2 Implementation Details

This paper's implementation is illustrated in Figure 2. Following Krasner et al. [15], this paper uses XLM-RoBERTa [6] as the text encoder f_θ and a pretrained ViT-B/32 [13] as the image encoder f^v. For each run, 2 captions per image for scenarios 1 (Parallel) and 2 (Semi-Parallel), and 1 caption per image for scenario 3 (Pseudo-Parallel) are sampled. Each language has an equal probability of being sampled during training. However, each run only includes one low-resource language (Tamil or Minangkabau) at a time.

During the first epoch of training, only high-resource languages are used, after which low-resource languages are introduced in subsequent epochs. For all scenarios, the models are trained on 4 epochs with five random seeds, learning rate set to 2e-5, batch size to 64, and temperature τ, τ' to 0.01. For scenarios 1 (Parallel) and 2 (Semi-

	Model	MS-COCO (Tam)	FLoRes-200 (Tam)	MUSTS (Tam)	MS-COCO (Min)	FLoRes-200 (Min)
	Base (w/o finetuning)	0.53	2.26	2.93	0.71	5.56
Parallel	Text-only ($\lambda = 0$)	97.05	99.24	44.88	96.60	93.79
Parallel	Multimodal ($\lambda = 0.01$)	97.11	99.08	**45.81**	96.47	92.09
Semi-Parallel	Text-only ($\lambda = 0$)	85.07	87.68	42.54	81.69	72.92
Semi-Parallel	Multimodal ($\lambda = 0.01$)	85.04	88.47	**44.14**	83.56	73.44
Pseudo-Parallel	Multimodal	83.48	86.96	40.88	81.37	68.96

Table 1 Main results on Tamil (Tam) and Minangkabau (Min) languages. Bi-text retrieval accuracy is reported for MS-COCO and FLoRes-200, while Spearman's correlation coefficient is reported for MUSTS. The best score for each scenario and language is underlined. Bold scores show statistically significant improvements over the text-only baseline ($p < 0.05$) using the paired t-test.

Parallel), the weight λ is set to 0.01. Cosine similarity is used as the similarity function.

4.3 Evaluation

For the Tamil language, the models are evaluated on the translated MS-COCO captions and FLoRes-200 [18] bi-text retrieval tasks. MS-COCO and FLoRes-200 contain 5k and 1k test translation pairs for `Eng <-> Low-resource` in both dev and test sets, respectively. We report the average accuracy of `Eng <-> Low-resource` language retrieval. MUSTS [8], a recently released STS benchmark for Tamil, is also used for evaluation. Similar to MCSE [10] and mSimCSE [1], the `[CLS]` token embeddings from the text encoder are used for the STS evaluation. For the Minangkabau language, the models are evaluated on the translated MS-COCO captions and FLoRes-200 bi-text retrieval tasks.

4.4 Results and Discussion

The main results are shown in Table 1. For scenario 1 (Parallel), adding multimodal alignment improves the MUSTS Spearman score for Tamil from 44.88 to 45.81. A similar observation is found for scenario 2 (Semi-Parallel), where the MUSTS Spearman score improved from 42.54 to 44.14. These improvements on STS tasks align with the findings of Zhang et al. [10]. For bi-text retrieval tasks, the difference in adding multimodal alignment is not statistically significant. One possible explanation is that bi-text retrieval tasks require precise alignment between translation pairs, which multimodal alignment does not directly optimize. STS tasks, on the other hand, measure the overall semantic relationship between sentences, which can benefit from the additional semantic information provided by multimodal alignment.

A notable observation is that the overall performance on Minangkabau is only slightly lower than Tamil, despite Minangkabau not being included in the pretraining of XLM-RoBERTa [6]. One possible explanation is that Minangkabau is linguistically similar to Indonesian, a mid to high-resource language included in the pretraining.

Results from scenario 2 (Semi-Parallel) and scenario 3 (Pseudo-Parallel) are consistently lower than those from scenario 1 (Parallel). However, it has to be noted that these two scenarios do not require any translation pairs and could solely rely on monolingual speakers for image captioning.

5 Conclusion

This paper explored three practical scenarios for leveraging multimodal data to improve low-resource language sentence embeddings: **Parallel**, **Semi-Parallel**, and **Pseudo-Parallel**. Experimental results show that when incorporating multimodal alignment, there are consistent improvements on STS tasks but limited improvements on bi-text retrieval tasks. Through data curation simulated with machine translation, this paper shows that the **Semi-Parallel** scenario, which relies solely on monolingual speakers' image captioning, can be a cost-effective choice as well. Models trained in the **Semi-Parallel** scenario can then be further bootstrapped for early-stage applications (e.g., retrieval systems) of very low-resource languages, to ultimately aid high-quality translation dataset construction.

References

[1] Yaushian Wang, Ashley Wu, and Graham Neubig. English contrastive learning can learn universal cross-lingual sentence embeddings. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2022.

[2] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 597–610, 09 2019.

[3] Sebastião Miranda, Artūrs Znotiņš, Shay B. Cohen, and Guntis Barzdins. Multilingual clustering of streaming news. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2018.

[4] Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. Cross-lingual retrieval for iterative self-supervised training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 2207–2219. Curran Associates, Inc., 2020.

[5] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding, 2022.

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Association for Computational Linguistics, 2020.

[7] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Association for Computational Linguistics, 2022.

[8] Tharindu Ranasinghe, Hansi Hettiarachchi, Constantin Orasan, and Ruslan Mitkov. MUSTS: MUltilingual semantic textual similarity benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Association for Computational Linguistics, 2025.

[9] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2018.

[10] Miaoran Zhang, Marius Mosbach, David Ifeoluwa Adelani, Michael A. Hedderich, and Dietrich Klakow. MCSE: Multimodal contrastive learning of sentence embeddings. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Association for Computational Linguistics, 2022.

[11] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2021.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, **Computer Vision – ECCV 2014**. Springer International Publishing, 2014.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **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)**. Association for Computational Linguistics, 2019.

[15] Nathaniel Krasner, Nicholas Lanuzo, and Antonios Anastasopoulos. Cross-lingual representation alignment through contrastive image-caption tuning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Association for Computational Linguistics, 2025.

[16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In **arXiv preprint arXiv:1807.03748**, 2018.

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions, 2015.

[18] NLLB Team. No language left behind: Scaling human-centered machine translation.