Disentangling Meaning and Language Components in Diverse Multilingual Sentence Embeddings

Kanade Nonomura[†] **Keita Fukushima**[†] **Risa Kondo**[†] **Tomoyuki Kajiwara**[†‡]
[†] Ehime University, Japan [‡] The University of Osaka, Japan
{nonomura, fukushima, kondo}@ai.cs.ehime-u.ac.jp
kajiwara@cs.ehime-u.ac.jp

Abstract

We disentangle multilingual sentence embeddings into language-dependent and language-agnostic components, leveraging the latter to improve cross-lingual similarity estimation. Previous studies focused on encoder-based approaches that use only the input sentence; in contrast, this study examines the effectiveness of disentanglement methods across a broader range of sentence embeddings, including decoder-based approaches and those that utilize prompts. Experimental results demonstrate that embedding disentanglement is effective for a wide variety of sentence embeddings.

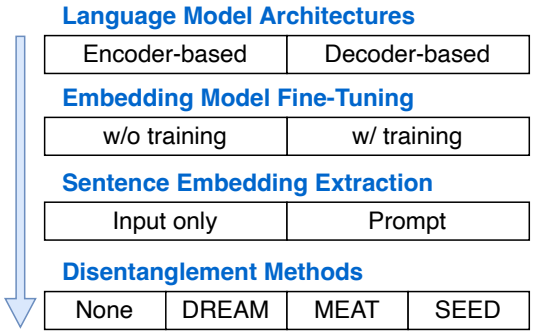

Figure 1: Overview of meaning embedding derivation.

1 Introduction

Multilingual sentence embeddings (Feng et al., 2022; Wang et al., 2024; Zhang et al., 2025; Babakhin et al., 2025) are widely used as a foundational technique for cross-lingual similarity estimation (Cer et al., 2017; Specia et al., 2020). However, previous studies have reported the issue of language specificity, where embeddings are influenced by language-specific information and tend to form language-specific subspaces (Libovický et al., 2020; Tiyajamorn et al., 2021). To address this problem, prior work (Tiyajamorn et al., 2021; Kuroda et al., 2022; Ki et al., 2024; Fukushima et al., 2025) proposed disentangling sentence embeddings into two components: language-specific information (hereafter, *language embeddings*) and language-independent information (hereafter, *meaning embeddings*), showing that using the latter improves cross-lingual task performance.

Existing disentanglement methods have proven effective for Transformer encoder-based approaches (Devlin et al., 2019; Conneau et al., 2020; Feng et al., 2022; Wang et al., 2024). However, their effectiveness for recently introduced, high-performing decoder-based and prompt-based embedding methods (Llama Team, 2024; Qwen Team, 2025; Jiang et al., 2022, 2024; Zhang et al., 2025; Babakhin et al., 2025; Enevoldsen et al., 2025) remains largely underexplored. While a concurrent study has analyzed disentanglement from the perspective of loss functions (Nonomura et al., 2026), a comprehensive evaluation across diverse architectures and extraction methods is still lacking.

As illustrated in Figure 1, this study systematizes the derivation of meaning embeddings across four dimensions: language model architectures, embedding model fine-tuning, sentence embedding extraction, and disentanglement methods. We comprehensively evaluate the effectiveness of disentanglement methods on such systematically organized diverse sentence embeddings.

Experimental results on the machine translation Quality Estimation (QE) task from WMT20 (Specia et al., 2020) show that disentanglement methods consistently improve performance across diverse sentence embeddings. Furthermore, through geometric analysis of embedding spaces, we demonstrate that the fine-tuning of embedding models is a key factor affecting the degree of performance improvement obtained through disentanglement.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1169–1176

2 Taxonomy of Meaning Embeddings

In this section, we systematize meaning embedding derivation across four dimensions: language model architecture, embedding model fine-tuning, extraction methods, and disentanglement methods.

2.1 Language Model Architectures

Current Transformer-based language models are broadly categorized into encoder-based and decoder-based models. While encoder-based models (Devlin et al., 2019; Conneau et al., 2020) are pretrained to produce contextualized word embeddings for language understanding and decoder-based models (Llama Team, 2024; Qwen Team, 2025) generate the next token for language generation, both serve as highly transferable foundational models for various tasks.

2.2 Embedding Model Fine-Tuning

Pretrained language models (Devlin et al., 2019; Conneau et al., 2020; Llama Team, 2024; Qwen Team, 2025) are strongly affected by anisotropy, an uneven distribution of embeddings that degrades downstream performance (Gao et al., 2019; Li et al., 2020; Gao et al., 2021). To mitigate this, contrastive learning has become a standard training approach (Reimers and Gurevych, 2019; Gao et al., 2021), and recently decoder-based models (Zhang et al., 2025; Babakhin et al., 2025) have achieved strong benchmark performance alongside traditional encoder-based models (Feng et al., 2022; Wang et al., 2024; Enevoldsen et al., 2025).

2.3 Sentence Embedding Extraction

Traditional sentence embeddings (Feng et al., 2022; Wang et al., 2024) use only the input sentence with mean pooling or the CLS token. In contrast, recent methods generate task-specific embeddings by switching prompts depending on the target application during the training of embedding models (Zhang et al., 2025; Babakhin et al., 2025). Furthermore, PromptBERT (Jiang et al., 2022) and PromptEOL (Jiang et al., 2024) demonstrated that prompt utilization alone can improve performance, even without fine-tuning the embedding model.

2.4 Disentanglement Methods

Ideally, multilingual sentence embeddings should map sentences with similar meanings across different languages to nearby vectors. However, they often suffer from the language specificity problem, where embeddings cluster by language (Tiyajamorn et al., 2021). To address this, previous studies disentangle an original sentence embedding $\boldsymbol{e} \in \mathbb{R}^d$ into a meaning embedding $\boldsymbol{e}^{(m)}$ and a language embedding $\boldsymbol{e}^{(l)}$, demonstrating that utilizing the former improves cross-lingual task performance (Tiyajamorn et al., 2021; Kuroda et al., 2022; Ki et al., 2024; Fukushima et al., 2025).

To formulate the objectives of these methods, let (x_i, y_i) denote the i-th pair of parallel sentences in languages X and Y within a batch, and let j denote a different index representing non-parallel sentences in the same batch. We describe four representative disentanglement methods below:

DREAM (Tiyajamorn et al., 2021): DREAM extracts meaning and language embeddings via two multi-layer perceptrons: $\boldsymbol{e}^{(m)} = \mathrm{MLP}_M(\boldsymbol{e})$ and $\boldsymbol{e}^{(l)} = \mathrm{MLP}_L(\boldsymbol{e})$. It relies on a meaning loss L_{mean} to align parallel sentences (x_i, y_i) and push apart non-parallel ones:

$$\begin{aligned} L_{\text{mean}} = 1 &- \cos\left(\boldsymbol{e}^{(m)}_{x_i}, \boldsymbol{e}^{(m)}_{y_i}\right) \\ &+ \max\left(0, \cos\left(\boldsymbol{e}^{(m)}_{x_i}, \boldsymbol{e}^{(m)}_{x_j}\right)\right) \\ &+ \max\left(0, \cos\left(\boldsymbol{e}^{(m)}_{y_i}, \boldsymbol{e}^{(m)}_{y_j}\right)\right). \end{aligned}$$

It also utilizes a language loss L_{lang} to maximize similarity between same-language sentences:

$$L_{\text{lang}} = 2 - \cos\left(\boldsymbol{e}^{(l)}_{x_i}, \boldsymbol{e}^{(l)}_{x_j}\right) - \cos\left(\boldsymbol{e}^{(l)}_{y_i}, \boldsymbol{e}^{(l)}_{y_j}\right)$$

In addition, DREAM employs an identification loss L_{id} that trains an auxiliary classifier to predict the language from $\boldsymbol{e}^{(l)}$.

MEAT (Kuroda et al., 2022): MEAT extends DREAM with an adversarial discriminator to eliminate language identifiability from $\boldsymbol{e}^{(m)}$ by driving its predicted language distribution toward a uniform distribution. Furthermore, MEAT replaces MSE in the reconstruction loss with cosine similarity and introduces a cross-reconstruction loss L_{cross}. This ensures meaning and language embeddings are interchangeable across parallel sentences:

$$\begin{aligned} L_{\text{recon}} = 2 &- \cos\left(\boldsymbol{e}_{x_i}, \boldsymbol{e}^{(m)}_{x_i} + \boldsymbol{e}^{(l)}_{x_i}\right) \\ &- \cos\left(\boldsymbol{e}_{y_i}, \boldsymbol{e}^{(m)}_{y_i} + \boldsymbol{e}^{(l)}_{y_i}\right) \\ L_{\text{cross}} = 2 &- \cos\left(\boldsymbol{e}_{x_i}, \boldsymbol{e}^{(m)}_{y_i} + \boldsymbol{e}^{(l)}_{x_i}\right) \\ &- \cos\left(\boldsymbol{e}_{y_i}, \boldsymbol{e}^{(m)}_{x_i} + \boldsymbol{e}^{(l)}_{y_i}\right). \end{aligned}$$

Language Pair	Training Pairs
en–de, en–zh	1,000k
ro–en, et–en	200k
ne–en, si–en	50k

Table 1: Training data size per language pair.

ORACLE (Ki et al., 2024): To mitigate semantic leakage in frameworks like DREAM and MEAT, ORACLE introduces an auxiliary objective that explicitly enforces orthogonality between meaning and language embeddings.

SEED (Fukushima et al., 2025): To prevent information loss caused by using multiple MLPs, SEED uses a single MLP for meaning and defines the language representation strictly as the residual: $\boldsymbol{e}^{(m)} = \mathrm{MLP}(\boldsymbol{e}), \quad \boldsymbol{e}^{(l)} = \boldsymbol{e} - \boldsymbol{e}^{(m)}$. Alongside L_{mean}, L_{lang}, and a separation loss L_{sep}, SEED expands L_{cross} to reconstruct embeddings swapped between both parallel and same-language sentences:

$$
\begin{aligned}
L_{\text{cross}} = 4 &- \cos\left(\boldsymbol{e}_{x_i}, \boldsymbol{e}^{(m)}_{y_i} + \boldsymbol{e}^{(l)}_{x_i}\right) \\
&- \cos\left(\boldsymbol{e}_{y_i}, \boldsymbol{e}^{(m)}_{x_i} + \boldsymbol{e}^{(l)}_{y_i}\right) \\
&- \cos\left(\boldsymbol{e}_{x_i}, \boldsymbol{e}^{(m)}_{x_i} + \boldsymbol{e}^{(l)}_{x_j}\right) \\
&- \cos\left(\boldsymbol{e}_{y_i}, \boldsymbol{e}^{(m)}_{y_i} + \boldsymbol{e}^{(l)}_{y_j}\right).
\end{aligned}
$$

3 Evaluation on the WMT20 QE Task

We evaluate the effectiveness of disentanglement methods (Tiyajamorn et al., 2021; Kuroda et al., 2022; Fukushima et al., 2025) across diverse sentence embeddings using the WMT20 QE task (Specia et al., 2020). We address this task, which estimates translation quality without references, under an unsupervised setting since only parallel corpora are used for training.[1] QE is adopted in this study because it strongly preserves semantic consistency between source and translated sentence pairs, making it a suitable task for evaluating the disentanglement of meaning and language.

3.1 Experimental Setup

Dataset The WMT20 QE task provides 1,000 evaluation sentence pairs annotated with human quality scores across six language pairs.[2] The evaluated machine translation systems are Transformer models (Vaswani et al., 2017; Ott et al., 2019). For training the disentanglement models, we used a subset of the available parallel corpora.[3] We detail the exact dataset sizes per language pair in Table 1.

Model	Dimensions (d)
mBERT	768
LaBSE	768
XLM-R	1024
mE5	1024
Llama3.1	4096
Llama-Emb.	4096
Qwen3	4096
Qwen-Emb.	4096

Table 2: Model embedding dimensions.

Models We pair pretrained language models with their fine-tuned embedding counterparts. For encoder-based models, we use mBERT-base[4] (Devlin et al., 2019) with LaBSE[5] (Feng et al., 2022), and XLM-R-large[6] (Conneau et al., 2020) with mE5-large[7] (Wang et al., 2024). For decoder-based models, we use Llama3.1-8B[8] (Llama Team, 2024) with Llama-Embed-Nemotron-8B[9] (Babakhin et al., 2025), and Qwen3-8B[10] (Qwen Team, 2025) with Qwen3-Embedding-8B[11] (Zhang et al., 2025). For disentanglement, we train a single-layer MLP while keeping the embedding models frozen. Dimensionality details are provided in Table 2.

Methods for Obtaining Sentence Embeddings We compare settings with and without prompts. For the encoder-based embedding models, LaBSE and mE5, when prompts are not used, the `[CLS]` token and mean pooling are used as the sentence embeddings, respectively. For the corresponding language models, mBERT and XLM-R, sentence

[1] In supervised settings, models are trained with source sentences, target sentences, and human scores, but our unsupervised setting does not use human scores during training.

[2] https://github.com/facebookresearch/mlqe English–German, English–Chinese, Romanian–English, Estonian–English, Nepali–English, and Sinhala–English.

[3] Randomly sampled from http://www.statmt.org/wmt20/quality-estimation-task.html in quantities comparable to prior work (Kuroda et al., 2022).

[4] https://huggingface.co/google-bert/bert-base-multilingual-cased

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

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

[7] https://huggingface.co/intfloat/multilingual-e5-large

[8] https://huggingface.co/meta-llama/Llama-3.1-8B

[9] https://huggingface.co/nvidia/llama-embed-nemotron-8b

[10] https://huggingface.co/Qwen/Qwen3-8B

[11] https://huggingface.co/Qwen/Qwen3-Embedding-8B

Embedding model training	Sentence embedding method	Disentanglement	Encoder-based		Decoder-based	
			mBERT	XLM-R	Llama3.1	Qwen3
w/o training	Input only	Baseline	0.107	0.091	0.013	0.010
		DREAM	**0.345**	**0.384**	**0.311**	**0.230**
		MEAT	0.184	0.217	0.252	0.150
		SEED	0.331	0.372	0.204	0.226
	Prompt	Baseline	0.138	0.165	0.271	0.156
		DREAM	**0.306**	**0.389**	0.413	0.331
		MEAT	0.164	0.173	0.393	0.337
		SEED	0.296	0.372	**0.414**	**0.412**
			LaBSE	**mE5**	**Llama-Emb.**	**Qwen-Emb.**
w/ training	Input only	Baseline	0.396	0.407	–	–
		DREAM	0.458	0.494	–	–
		MEAT	**0.491**	0.489	–	–
		SEED	0.482	**0.496**	–	–
	Prompt	Baseline	–	–	0.524	0.456
		DREAM	–	–	0.488	0.369
		MEAT	–	–	**0.531**	0.459
		SEED	–	–	0.527	**0.478**

Table 3: Pearson correlation with human evaluation scores in the WMT20 QE task (average over six language pairs).

embeddings are obtained using the same methods as their corresponding embedding models. When prompts are used, we follow the method of PromptBERT (Jiang et al., 2022).[12] For the decoder-based embedding models, Llama-Embed-Nemotron-8B and Qwen3-Embedding-8B, prompts are incorporated into the model input,[13] and mean pooling and the `[EOS]` token are used as sentence embeddings, respectively (Babakhin et al., 2025). For the corresponding language models, Llama3.1-8B and Qwen3-8B, we adopt the same sentence embedding settings as those used for their embedding model counterparts. When prompts are used, we follow the PromptEOL (Jiang et al., 2024).[14]

Disentanglement Methods We apply three disentanglement methods (DREAM, MEAT, and SEED) and compare their performance against a baseline using original sentence embeddings without disentanglement.

Hyperparameters Training is conducted using HuggingFace Transformers (Wolf et al., 2020), with a batch size of 512 and the Adam optimizer (Kingma and Ba, 2015). The learning rate is set to $1e-4$ for DREAM and SEED, and $1e-5$ for MEAT. For validation, 10% of the training data is randomly sampled. Training is terminated if the validation loss does not improve for three epochs.

3.2 Experimental Results

As shown in Table 3, applying disentanglement generally outperforms the Baseline across diverse sentence embeddings, though the optimal method depends on the embedding type. Notably, models without embedding-specific fine-tuning exhibit substantially larger performance gains. This is likely because fine-tuned models already acquire cross-lingual semantic alignment via contrastive learning, which functionally overlaps with the disentanglement objectives. Furthermore, decoder-based models tend to show smaller improvements than encoder-based models. This suggests that decoder-based models already filter out language-specific information to some extent at the embedding stage, due to their larger scale and use of prompts.

4 Geometric Analysis

To investigate how embedding model fine-tuning affects disentanglement, we analyze embedding spaces using Alignment and Uniformity (Wang and Isola, 2020), metrics widely used to evaluate representation learning (Gao et al., 2021; Chuang et al., 2022; Li et al., 2024). For the analysis, we evaluate one pretrained language model and one fine-tuned embedding model per architecture: mBERT-base and LaBSE (encoder), and Llama3.1-8B and

[12] We use the prompt "`This sentence : "[X]" means [MASK] .`" and extract the embedding of the final-layer hidden state at the position of the `[MASK]` token. Here, [X] denotes a placeholder for the input sentence.

[13] We use the instruction "`Retrieve semantically similar text`" for the STS task.

[14] We input "`This sentence : "[X]" means in one word: "`" and use the embedding of the final-layer hidden state at the position of the closing quotation mark at the end of the sentence.

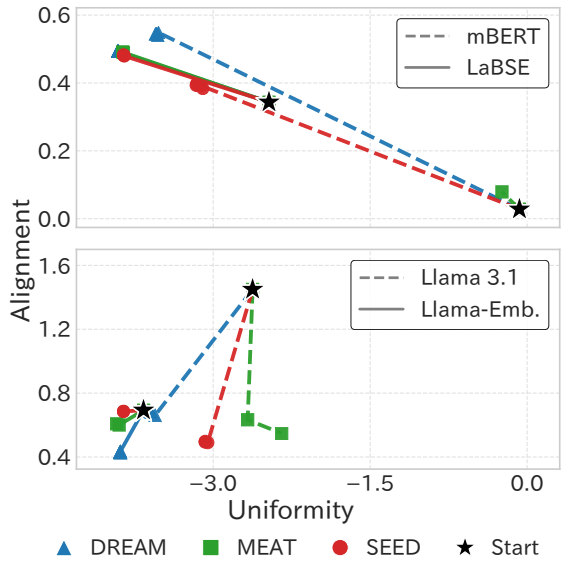

Figure 2: Changes in Alignment and Uniformity every 3 epochs during disentanglement training.

Llama-Embed-Nemotron-8B (decoder). For pretrained models, we obtain sentence embeddings using only the input sentence, denoting the L2-normalized meaning embedding of an input x as $\tilde{h}(x)$.

4.1 Alignment and Uniformity

Alignment, which measures how close the meaning embeddings of parallel sentences are in the normalized embedding space, is defined as follows, where x and y are sentences in languages X and Y, and p_{pos} is the distribution of parallel pairs (x, y):

$$L_{\text{align}} = \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} \left\| \tilde{h}(x) - \tilde{h}(y) \right\|^2. \quad (1)$$

Smaller values indicate better cross-lingual alignment, with parallel sentences closer together.

Uniformity, which measures how evenly normalized embeddings are distributed across the entire space, is defined as follows using the marginal distribution p_{data} across all language pairs:

$$L_{\text{uniform}} = \log \mathop{\mathbb{E}}_{x,z \overset{\text{i.i.d.}}{\sim} p_{\text{data}}} e^{-2\left\| \tilde{h}(x) - \tilde{h}(z) \right\|^2}. \quad (2)$$

Smaller values indicate that embeddings are uniformly dispersed across the hypersphere, utilizing the entire space rather than collapsing locally.

4.2 Analysis Results

Following the settings in Section 3.1, we measured Alignment and Uniformity every three epochs during training using 1,000 randomly sampled validation instances per language pair.

As shown in Figure 2, pretrained models exhibit larger geometric improvements (decreased metrics) after disentanglement compared to embedding models, providing geometric evidence that fine-tuning dictates the degree of performance improvement. While mBERT-base initially shows a deceptively low Alignment, its extremely high Uniformity indicates that embeddings are collapsed into a limited region. Therefore, the drastic improvement in Uniformity during training more than compensates for the slight deterioration in Alignment, as evidenced by its significant downstream performance gains after applying disentanglement. Overall, the presence or absence of embedding model training determines the degree of geometric improvement, which directly translates to downstream task performance.

5 Conclusion

We comprehensively investigated the effectiveness of disentanglement methods across diverse multilingual sentence embeddings. Experiments on the WMT20 QE task and geometric analyses demonstrate that these methods are broadly effective, and notably, that embedding model fine-tuning significantly dictates the magnitude of performance improvement.

Acknowledgments

This work was supported by JST BOOST Program Japan Grant Number JPMJBY24036821.

Limitations

We evaluate disentanglement methods only on the WMT20 QE task. While this task is suitable for assessing cross-lingual semantic representations and has been widely adopted in prior work, the effectiveness of disentanglement methods on other tasks, such as retrieval and classification, has not been sufficiently investigated.

In addition, we disentangle sentence embeddings into only two components: meaning and language. Prior work has explored disentanglement along different factors, such as separating semantics from syntax (Chen et al., 2019; Huang et al., 2021). Our work focuses specifically on the language factor and does not consider other disentanglement axes. Furthermore, sentence embeddings may contain additional attributes, such as domain (Kondo et al., 2025) and style. Our approach does not extend to finer-grained disentanglement of these factors, which we leave for future work.

References

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-Embed-Nemotron-8B: A Universal Text Embedding Model for Multilingual and Cross-Lingual Tasks. *arXiv:2511.07025*.

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.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A Multi-Task Approach for Disentangling Syntax and Semantics in Sentence Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2453–2464.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218.

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.

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.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. MMTEB: Massive Multilingual Text Embedding Benchmark. *arXiv:2502.13595*.

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.

Keita Fukushima, Tomoyuki Kajiwara, and Takashi Ninomiya. 2025. Reversible Disentanglement of Meaning and Language Representations from Multilingual Sentence Encoders. In *Proceedings of the 5th Workshop on Multilingual Representation Learning*, pages 265–270.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation Degeneration Problem in Training Natural Language Generation Models. In *International Conference on Learning Representations*.

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.

James Y. Huang, Kuan-Hao Huang, and Kai-Wei Chang. 2021. Disentangling Semantics and Syntax in Sentence Embeddings with Pre-trained Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1372–1379.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024. Scaling Sentence Embeddings with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. PromptBERT: Improving BERT Sentence Embeddings with Prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837.

Dayeon Ki, Cheonbok Park, and Hyunjoong Kim. 2024. Mitigating Semantic Leakage in Cross-lingual Embeddings via Orthogonality Constraint. In *Proceedings of the 9th Workshop on Representation Learning for NLP*, pages 256–273.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Risa Kondo, Hiroki Yamauchi, Tomoyuki Kajiwara, Marie Katsurai, and Takashi Ninomiya. 2025. Domain Knowledge Distillation for Multilingual Sentence Encoders in Cross-lingual Sentence Similarity Estimation. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing*, pages 572–577.

Yuto Kuroda, Tomoyuki Kajiwara, Yuki Arase, and Takashi Ninomiya. 2022. Adversarial Training on Disentangling Meaning and Language Representations for Unsupervised Quality Estimation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5240–5245.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the Sentence Embeddings from Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9119–9130.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving In-context Learning of Multilingual Generative Language Models with Cross-lingual Alignment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 8058–8076.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the Language Neutrality of Pre-trained Multilingual Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674.

Llama Team. 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.

Kanade Nonomura, Keita Fukushima, Risa Kondo, and Tomoyuki Kajiwara. 2026. Mitigating Language Bias in Multilingual Sentence Embeddings for Cross-Lingual Similarity Estimation. In *Proceedings of the 15th Joint Conference on Lexical and Computational Semantics*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 48–53.

Qwen Team. 2025. Qwen3 Technical Report. *arXiv:2505.09388*.

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.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 Shared Task on Quality Estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-agnostic Representation from Multilingual Sentence Encoders for Cross-lingual Similarity Estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv:2402.05672*.

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.

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, and 3 others. 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.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv:2506.05176*.

A Per-Language Pair Results

To confirm that the average improvements in Table 3 are not skewed by specific languages, Table 4 reports the Pearson correlation scores for all six language pairs.

Model (Extraction)	Method	en–de	en–zh	ro–en	et–en	ne–en	si–en	Avg..
mBERT (Input only)	Baseline	0.094	0.040	0.223	0.112	0.064	–	0.107
	DREAM	0.143	**0.124**	**0.676**	**0.378**	**0.406**	–	**0.345**
	MEAT	0.136	0.060	0.419	0.269	0.037	–	0.184
	SEED	**0.164**	0.108	0.644	0.371	0.370	–	0.331
XLM-R (Input only)	Baseline	-0.071	-0.022	0.347	0.032	0.059	0.201	0.091
	DREAM	**0.122**	**0.197**	**0.693**	**0.421**	**0.463**	**0.406**	**0.384**
	MEAT	0.074	0.103	0.498	0.212	0.178	0.237	0.217
	SEED	0.110	0.190	0.667	0.417	0.441	**0.406**	0.372
mBERT (Prompt)	Baseline	-0.002	0.043	0.307	0.172	0.169	–	0.138
	DREAM	0.031	**0.113**	**0.653**	0.369	**0.365**	–	**0.306**
	MEAT	**0.088**	0.103	0.362	0.182	0.086	–	0.164
	SEED	0.027	0.109	0.638	**0.375**	0.330	–	0.296
XLM-R (Prompt)	Baseline	-0.024	-0.026	0.457	0.158	0.227	0.199	0.165
	DREAM	0.106	**0.178**	**0.691**	**0.456**	**0.492**	**0.408**	**0.389**
	MEAT	0.091	0.065	0.316	0.123	0.162	0.278	0.173
	SEED	**0.119**	0.154	0.661	0.447	0.453	0.395	0.372
LaBSE (Input only)	Baseline	0.084	0.036	0.705	0.550	0.547	0.455	0.396
	DREAM	0.151	0.156	0.711	0.549	0.627	0.552	0.458
	MEAT	**0.215**	**0.222**	0.717	**0.587**	**0.634**	**0.571**	**0.491**
	SEED	0.192	0.192	**0.725**	0.583	0.633	0.565	0.482
mE5 (Input only)	Baseline	0.020	0.100	0.734	0.556	0.538	0.493	0.407
	DREAM	0.172	**0.257**	**0.783**	0.629	0.584	0.541	0.494
	MEAT	**0.184**	0.243	**0.783**	0.629	0.566	0.530	0.489
	SEED	0.176	0.248	0.781	**0.635**	**0.591**	**0.543**	**0.496**
Llama3.1 (Input only)	Baseline	-0.088	-0.127	0.122	-0.279	0.293	0.155	0.013
	DREAM	0.067	0.139	**0.657**	**0.379**	**0.351**	**0.271**	**0.311**
	MEAT	**0.084**	**0.174**	0.576	0.300	0.316	0.059	0.252
	SEED	0.064	0.135	0.636	0.377	0.045	-0.034	0.204
Qwen3 (Input only)	Baseline	-0.068	-0.040	0.095	-0.018	0.018	0.068	0.010
	DREAM	0.071	0.124	0.534	0.290	**0.215**	**0.143**	**0.230**
	MEAT	0.063	0.079	0.460	0.134	0.081	0.085	0.150
	SEED	**0.157**	**0.156**	**0.600**	**0.329**	0.154	-0.041	0.226
Llama3.1 (Prompt)	Baseline	0.009	0.152	0.477	0.240	0.403	0.343	0.271
	DREAM	0.102	0.211	0.685	0.412	**0.569**	**0.499**	0.413
	MEAT	**0.141**	**0.220**	0.664	0.367	0.515	0.449	0.393
	SEED	0.107	0.215	**0.687**	**0.423**	0.562	0.487	**0.414**
Qwen3 (Prompt)	Baseline	-0.007	0.092	0.572	0.132	0.098	0.049	0.156
	DREAM	0.126	0.201	0.560	0.376	0.449	0.272	0.331
	MEAT	0.080	0.241	0.678	0.409	0.410	0.201	0.337
	SEED	**0.170**	**0.258**	**0.701**	**0.469**	**0.548**	**0.323**	**0.412**
Llama-Emb. (Prompt)	Baseline	0.203	0.283	**0.817**	0.602	**0.711**	0.528	0.524
	DREAM	0.217	0.274	0.773	0.562	0.620	0.482	0.488
	MEAT	**0.232**	**0.309**	0.806	**0.610**	0.695	0.536	**0.531**
	SEED	0.209	0.288	0.814	0.602	0.709	**0.538**	0.527
Qwen3-Emb. (Prompt)	Baseline	0.186	0.261	0.755	0.476	**0.626**	0.431	0.456
	DREAM	0.203	0.230	0.623	0.359	0.468	0.332	0.369
	MEAT	0.183	0.268	**0.766**	0.502	0.619	0.413	0.459
	SEED	**0.222**	**0.293**	0.757	**0.518**	0.625	**0.454**	**0.478**

Table 4: Detailed Pearson correlation scores for each language pair on the WMT20 QE task. mBERT does not support Sinhala(si), hence the missing scores.