Reversible Disentanglement of Meaning and Language Representations from Multilingual Sentence Encoders

Keita Fukushima[†] **Tomoyuki Kajiwara**[†‡] **Takashi Ninomiya**[†]
[†] Graduate School of Science and Engineering, Ehime University, Japan
[‡] D3 Center, The University of Osaka, Japan
{fukushima@ai.cs., kajiwara@cs., ninomiya.takashi.mk@} ehime-u.ac.jp

Abstract

We propose an unsupervised method to disentangle sentence embeddings from multilingual sentence encoders into language-specific and language-agnostic representations.[1] Such language-agnostic representations distilled by our method can estimate cross-lingual semantic sentence similarity by cosine similarity. Previous studies have trained individual extractors to distill each language-specific and -agnostic representation. This approach suffers from missing information resulting in the original sentence embedding not being fully reconstructed from both language-specific and -agnostic representations; this leads to performance degradation in estimating cross-lingual sentence similarity. We only train the extractor for language-agnostic representations and treat language-specific representations as differences from the original sentence embedding; in this way, there is no missing information. Experimental results for both tasks, quality estimation of machine translation and cross-lingual sentence similarity estimation, show that our proposed method outperforms existing unsupervised methods.

1 Introduction

Estimating semantic textual similarity (STS) (Cer et al., 2017) is one of the fundamental techniques in natural language processing (NLP). This technology has many potential applications, including information retrieval (Bajaj et al., 2016) and automatic evaluation of NLP-generated sentences (Shimanaka et al., 2018). In recent years, this task has commonly been based on Transformer-based sentence encoders (Reimers and Gurevych, 2019; Wang et al., 2022) that are pre-trained in objectives such as masked language modeling (Devlin et al., 2019) and contrastive learning (Gao et al., 2021). These techniques are generalized across languages (K et al., 2020), and multilingual sentence encoders (Reimers and Gurevych, 2020; Feng et al., 2022; Wang et al., 2024) pre-trained in various languages are also being actively developed for applications such as quality estimation (QE) of machine translation (Specia et al., 2018) and cross-lingual information retrieval (Nie, 2010).

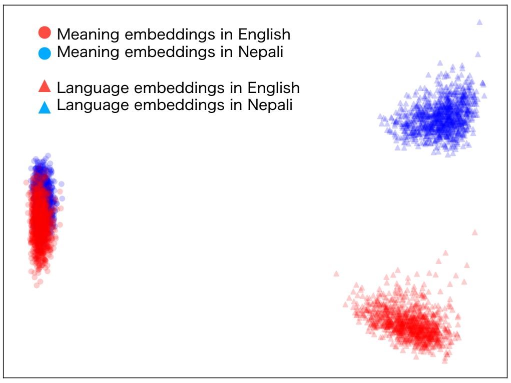

Figure 1: Visualization of embeddings in QE task by principal component analysis. Two colors represent the source and target languages, and two markers represent meaning and language embeddings. Our proposed method forms one cluster of language-agnostic meaning embeddings on the left side and two clusters of language-specific embeddings on the right side.

However, since sentence embeddings from multilingual sentence encoders are dominated by language rather than meaning (Tiyajamorn et al., 2021), they suffer from accurate estimation of cross-lingual sentence similarity. Previous studies (Tiyajamorn et al., 2021; Kuroda et al., 2022; Ki et al., 2024) have disentangled sentence embeddings from multilingual sentence encoders into embeddings that represent language-specific information (language embedding) and language-agnostic information (meaning embedding), and used the latter meaning embeddings for cross-lingual sentence similarity estimation. They have disentangled sentence embeddings using both extractors for language embeddings and for meaning embeddings, however, this approach may result in information missing during the disentanglement process.

[1] https://github.com/EhimeNLP/SEED

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 265–270
November 8-9, 2025 ©2025 Association for Computational Linguistics

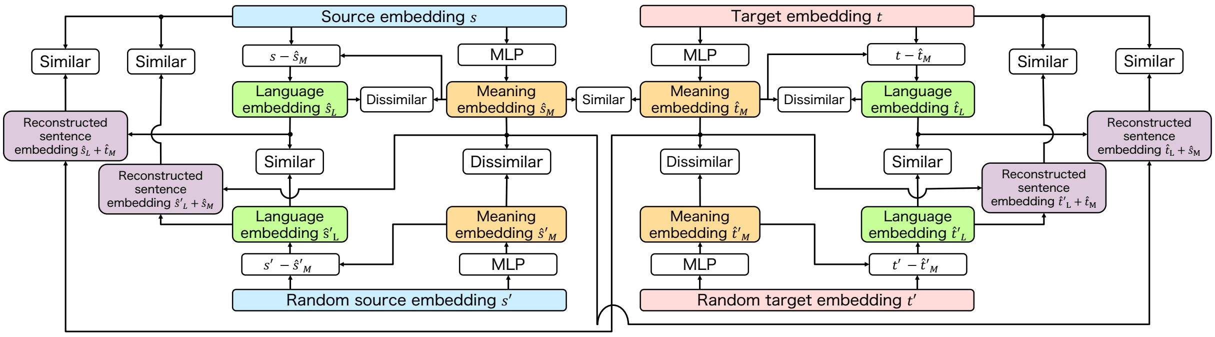
Source embedding s
Similar
Similar
$s - \hat{s}_M$
MLP
Language embedding $\hat{s}_L$
Dissimilar
Meaning embedding $\hat{s}_M$
Similar
Reconstructed sentence embedding $\hat{s}_L + \hat{t}_M$
Similar
Dissimilar
Reconstructed sentence embedding $\hat{s}'_L + \hat{s}_M$
Language embedding $\hat{s}'_L$
Meaning embedding $\hat{s}'_M$
$s' - \hat{s}'_M$
MLP
Random source embedding s'
Target embedding t
MLP
$t - \hat{t}_M$
Similar
Similar
Meaning embedding $\hat{t}_M$
Dissimilar
Language embedding $\hat{t}_L$
Reconstructed sentence embedding $\hat{t}_L + \hat{s}_M$
Dissimilar
Similar
Reconstructed sentence embedding $\hat{t}'_L + \hat{t}_M$
Meaning embedding $\hat{t}'_M$
Language embedding $\hat{t}'_L$
MLP
$t' - \hat{t}'_M$
Random target embedding t'

Figure 2: Distilling meaning embeddings from multilingual sentence embeddings. The four MLPs share weights.

To address this issue, we train only an extractor for meaning embeddings and treat language embeddings as the difference between original sentence embeddings and meaning embeddings. Since missing information cannot happen in this architecture, we expect to more accurately distill meaning embeddings from multilingual sentence encoders. Experimental results on QE in WMT20 (Specia et al., 2020) and cross-lingual STS in SemEval-2017 (Cer et al., 2017) show that the proposed method outperforms previous unsupervised methods (Tiyajamorn et al., 2021; Kuroda et al., 2022) and more accurately distills language-agnostic embeddings.

2 Proposed Method

Our proposed method disentangles sentence embeddings $\boldsymbol{e} \in \mathbb{R}^d$ from a multilingual sentence encoder into language embeddings representing language-specific information and meaning embeddings representing language-agnostic information using a multi-layer perceptron (MLP). Note that d is the dimension of sentence embeddings.

Previous studies (Tiyajamorn et al., 2021; Kuroda et al., 2022) used two MLPs (MLP_M and MLP_L) to distill meaning embeddings $\hat{\boldsymbol{e}}_M = \mathrm{MLP}_M(\boldsymbol{e})$ and language embeddings $\hat{\boldsymbol{e}}_L = \mathrm{MLP}_L(\boldsymbol{e})$ independently. MLPs are trained by adding these embeddings together to reconstruct the original sentence embeddings, however, complete reconstruction from independently extracted embeddings is difficult, and some information is lost. In contrast, the proposed method treats language embeddings as the difference between original sentence embeddings and meaning embeddings, allowing the addition of language and meaning embeddings to reconstruct the original sentence embeddings completely. We use only one MLP to extract meaning embeddings, as follows.

$$\hat{\boldsymbol{e}}_M = \mathrm{MLP}(\boldsymbol{e}) \tag{1}$$

$$\hat{\boldsymbol{e}}_L = \boldsymbol{e} - \hat{\boldsymbol{e}}_M \tag{2}$$

As shown in Figure 2, our extractor for meaning embeddings is trained in a multi-task learning manner based on the following three loss functions.

$$L = L_M + L_L + L_C \tag{3}$$

We train the MLP with bilingual parallel corpora. Figure 2 shows how the MLP is trained by combining loss functions based on cosine similarity between embeddings. Meaning embedding $\hat{\boldsymbol{s}}_M$, language embedding $\hat{\boldsymbol{s}}_L$, and others are disentangled from the sentence embeddings $\boldsymbol{s}$ and $\boldsymbol{t}$ from the multilingual sentence encoder for bilingual sentences consisting of a sentence in the source language s and a sentence in the target language t. As in previous study (Tiyajamorn et al., 2021), sentences s' and t', randomly selected from the source and target languages, respectively, are also used for training supplementally.

2.1 Loss for Language-agnostic Embeddings

Between bilingual sentences (s, t), the meaning embeddings $\hat{\boldsymbol{s}}_M$ and $\hat{\boldsymbol{t}}_M$ should be similar, and between randomly selected sentences (s, s') and (t, t'), they should be dissimilar. To train them, we define the following loss functions.

$$\begin{aligned} L_M = {} & 2\,(1 - \cos(\hat{\boldsymbol{s}}_M, \hat{\boldsymbol{t}}_M)) \\ & + \max(0, \cos(\hat{\boldsymbol{s}}_M, \hat{\boldsymbol{s}}'_M)) \\ & + \max(0, \cos(\hat{\boldsymbol{t}}_M, \hat{\boldsymbol{t}}'_M)) \end{aligned} \tag{4}$$

Note that the first term is weighted to balance the positive and negative terms.

	Model	en-de	en-zh	ro-en	et-en	ne-en	si-en	Avg.
mE5-base	Baseline	0.003	0.074	0.674	0.443	0.486	0.463	0.357
	Mean Centering	0.079	0.141	0.729	0.445	0.544	0.507	0.408
	DREAM	**0.120**	**0.213**	0.738	0.499	0.527	0.515	0.435
	MEAT	0.119	0.209	0.735	0.500	0.533	0.514	0.435
	Ours	0.116	0.190	**0.741**	**0.513**	**0.543**	**0.525**	**0.438**
mE5-large	Baseline	0.020	0.100	0.734	0.556	0.538	0.493	0.407
	Mean Centering	0.151	0.184	0.779	0.583	0.592	0.544	0.472
	DREAM	0.172	**0.257**	**0.783**	0.629	0.584	0.541	0.494
	MEAT	0.117	0.186	0.751	0.610	0.541	0.499	0.451
	Ours	**0.175**	0.249	0.782	**0.636**	**0.591**	**0.544**	**0.496**
mE5-large-instruct	Baseline	0.143	0.203	**0.767**	0.590	0.549	0.422	0.446
	Mean Centering	0.212	0.261	0.766	0.576	0.589	0.505	0.485
	DREAM	0.212	**0.290**	0.765	0.595	0.585	0.499	0.491
	MEAT	**0.215**	0.283	0.757	0.607	0.563	0.476	0.484
	Ours	**0.215**	0.284	0.762	**0.611**	**0.598**	**0.515**	**0.498**

Table 1: Pearson correlation coefficients evaluated on WMT20 QE task.

2.2 Loss for Language-specific Embeddings

Language embeddings $(\hat{\boldsymbol{s}}_L, \hat{\boldsymbol{t}}_L)$ should be similar for (s, s') and (t, t') within the same language. To train them, we define the following loss functions.

$$L_L = (1 - \cos(\hat{\boldsymbol{s}}_L, \hat{\boldsymbol{s}}'_L)) + (1 - \cos(\hat{\boldsymbol{t}}_L, \hat{\boldsymbol{t}}'_L)) \quad (5)$$

2.3 Loss for Both Language-specific and Language-agnostic Embeddings

Since the purpose of this method is to disentangle original sentence embeddings into meaning and language embeddings, it is desirable that these embeddings are not similar. In addition, the original sentence embedding should be reconstructed by adding meaning and language embeddings. Therefore, when language embeddings are swapped between (s, s') and (t, t') within the same language, or when meaning embeddings are swapped between (s, t) in the bilingual sentence, we want to reconstruct the original sentence embedding by adding meaning and language embeddings. To train them, we define the following loss functions.

$$\begin{aligned} L_C = &\max(0, \cos(\hat{\boldsymbol{s}}_M, \hat{\boldsymbol{s}}_L)) + \max(0, \cos(\hat{\boldsymbol{t}}_M, \hat{\boldsymbol{t}}_L)) \\ &+ 2 - \cos(\hat{\boldsymbol{s}}, \hat{\boldsymbol{s}}_M + \hat{\boldsymbol{s}}'_L) - \cos(\hat{\boldsymbol{t}}, \hat{\boldsymbol{t}}_M + \hat{\boldsymbol{t}}'_L) \\ &+ 2 - \cos(\hat{\boldsymbol{s}}, \hat{\boldsymbol{t}}_M + \hat{\boldsymbol{s}}_L) - \cos(\hat{\boldsymbol{t}}, \hat{\boldsymbol{s}}_M + \hat{\boldsymbol{t}}_L) \end{aligned} \quad (6)$$

3 Evaluation

We evaluate the performance of the proposed method on the QE task in WMT20 (Specia et al., 2020) and on the cross-lingual STS task in SemEval-2017 (Cer et al., 2017). Both tasks estimate the similarity between sentences, the former between an input sentence in the source language and a machine-translated sentence in the target language, and the latter between two sentences in different languages. Following the official evaluation metrics, we used Pearson correlation.

QE		STS	
en-de, en-zh	1,000 k	en-it, en-tr	500 k
ro-en, et-en	200 k	en-de, en-es, en-fr	200 k
ne-en, si-en	50 k	en-ar, en-nl	30 k

Table 2: Number of sentence pairs for each language pair in the training dataset. From each of these, 10% of the sentence pairs are used for validation.

	QE	STS
LaBSE	0.396	0.734
LaBSE + Ours	**0.482**	**0.753**
mE5	0.446	0.826
mE5 + Ours	**0.498**	**0.832**

Table 3: Summary of experimental results.

3.1 Setting

Data The WMT20 QE task includes six language pairs. English to German (en-de), English to Chinese (en-zh), Romanian to English (ro-en),

	Model	en-ar	en-de	en-tr	en-es	en-fr	en-it	en-nl	Avg.
mE5-base	Baseline	0.726	**0.809**	0.687	**0.772**	**0.802**	**0.811**	**0.799**	0.772
	Mean Centering	0.688	0.788	0.652	0.730	0.764	0.797	0.777	0.742
	DREAM	0.727	0.741	0.707	0.731	0.763	0.787	0.763	0.746
	MEAT	0.693	0.773	0.698	0.727	0.781	0.790	0.787	0.750
	Ours	**0.749**	0.786	**0.724**	0.754	0.793	0.810	0.795	**0.773**
mE5-large	Baseline	0.774	0.846	0.783	**0.806**	0.834	0.836	0.835	0.816
	Mean Centering	0.757	0.830	0.759	0.790	0.831	0.826	0.821	0.802
	DREAM	**0.803**	0.839	0.796	0.798	0.826	0.840	0.841	0.820
	MEAT	0.773	0.849	0.772	0.784	0.831	0.838	**0.855**	0.815
	Ours	0.797	**0.854**	**0.800**	0.801	**0.835**	**0.847**	0.853	**0.827**
mE5-large-instruct	Baseline	0.788	**0.847**	0.782	**0.840**	**0.835**	0.846	0.842	0.826
	Mean Centering	0.760	0.824	0.759	0.827	0.806	0.804	0.809	0.798
	DREAM	0.823	0.834	0.789	0.818	0.824	0.839	0.836	0.823
	MEAT	0.813	0.839	0.776	0.800	0.830	0.843	0.838	0.820
	Ours	**0.825**	0.846	**0.795**	0.827	**0.835**	**0.851**	**0.845**	**0.832**

Table 4: Pearson correlation coefficient evaluated on SemEval-2017 cross-lingual STS task.

Estonian to English (et-en), Nepali to English (ne-en), and Sinhalese to English (si-en), respectively, and for each language pair, $1,000$ sentence pairs of machine translation input/output and human evaluation scores are available for evaluation. The target machine translation model is a Transformer (Vaswani et al., 2017) trained with the fairseq toolkit (Ott et al., 2019).

The SemEval-2017 cross-lingual STS task includes seven language pairs in English and other languages. They are Arabic (en-ar), German (en-de), Turkish (en-tr), Spanish (en-es), French (en-fr), Italian (en-it), and Dutch (en-nl), respectively, with 250 sentence pairs and human evaluation scores available for each language pair.

Model Our MLP is a single-layer feed-forward neural network. LaBSE[2] (Feng et al., 2022) and three types of multilingual E5 (mE5)[3] (Wang et al., 2024) were used for multilingual sentence encoders. Only MLP is trained on bilingual corpora, and multilingual sentence encoders are frozen.

We used a batch size of 512, Adam (Kingma and Ba, 2015) optimizer with a learning rate of 10^{-4}. We employed early stopping for training with a patience of 5 epochs using a validation loss of Equation (3). As in previous studies (Tiyajamorn et al., 2021; Kuroda et al., 2022), we used part of the bilingual corpus available in WMT20[4] for the QE task and Tatoeba[5] for the STS task, respectively, for training. Table 2 shows our training data sizes.

L_M	L_L	L_C	Pearson
✓	✓		0.493
✓		✓	0.487
	✓	✓	0.451
✓	✓	✓	0.498

Table 5: Ablation on QE task. All loss functions are useful, and losing any of them degrades performance.

Comparison We compare the proposed method with DREAM (Tiyajamorn et al., 2021) and MEAT (Kuroda et al., 2022), previous studies that disentangle sentence embeddings into meaning and language embeddings. There are two baselines, one using sentence embeddings from the multilingual sentence encoder as is. The other is a simple disentangling method; the average embedding for the target language in the training corpus is subtracted from the embedding of each sentence to obtain the meaning embedding. The Cosine similarity of these sentence or meaning embeddings estimates translation quality or semantic similarity.

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

[3] https://huggingface.co/intfloat/multilingual-e5-{base,large,large-instruct}

[4] https://www.statmt.org/

[5] https://tatoeba.org/

3.2 Result

Table 3 provides a summary of the experimental results. For both multilingual encoders, LaBSE and mE5 (large-instruct), the proposed method improved the performance of both QE and STS tasks. For the mE5, which achieved higher performance, Tables 1 and 5 show detailed results for each task.

Experimental results for QE in Table 1 and STS in Table 4 show that the proposed method consistently achieves the best average performance for all multilingual sentence encoders. Figure 1 also reveals that disentangling sentence embeddings has been successful.

4 Conclusion

We disentangled sentence embeddings from multilingual sentence encoders into language-specific and language-agnostic embeddings, and applied the latter to cross-lingual sentence similarity estimation. The model architecture of our method has the advantage that there is no missing information during disentangling embeddings. Experimental results on QE and cross-lingual STS tasks in an unsupervised manner revealed the effectiveness of the proposed method for both state-of-the-art multilingual sentence encoders, LaBSE and mE5.

Limitations

Our method is based on pre-trained multilingual sentence encoders and is not applicable to languages not covered by the original encoders. Nonetheless, for example, LaBSE is available in as many as 109 languages.

Our model needs training on GPUs. However, the computation time is not very long: about 12 minutes per epoch on a single GPU of TITAN RTX, and about 5 to 9 hour for the entire training.

Acknowledgments

This work was supported by JST BOOST Program Japan Grant Number JPMJBY24036821 and JSPS KAKENHI Grant Number JP23K24907.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the 30th Conference on Neural Information Processing Systems*.

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.

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.

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.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *Proceedings of the Eighth International Conference on Learning Representations*.

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*.

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.

Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*. Morgan & Claypool Publishers.

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 (Demonstrations)*, pages 48–53.

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.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758.

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.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. *Quality Estimation for Machine Translation*. Morgan & Claypool Publishers.

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 5998–6008.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv:2212.03533*.

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