Boosting Monolingual Sentence Representation with Large-scale Parallel Translation Datasets

Jue Wang [1 2 *] **Haofan Wang** [2 *] **Xing Wu** [2 3] **Chaochen Gao** [2 3] **Debing Zhang** [2]

Abstract

Although contrastive learning greatly improves sentence representation, its performance is still limited by the size of existing monolingual datasets. So can semantically highly correlated massively parallel translation pairs be used for pre-training of monolingual models? This paper proposes an exploration of this. We leverage parallel translated sentence pairs to learn single-sentence sentence embeddings and demonstrate superior performance in balancing alignment and consistency. We achieve new state-of-the-art performance on the mean score of Standard Semantic Text Similarity (STS), outperforming both SimCSE and Sentence-T5.

1. Introduction

Gao et al. 2021 demonstrates that a contrastive objective can be extremely effective when coupled with pre-trained language models and sentence-pair datasets. However, the generality and capability of the language model are strictly limited by the size of existing sentence-pair datasets (Bowman et al., 2015; Williams et al., 2017). Meanwhile, there have accumulated large-scale parallel translation datasets (100x larger than existing monolingual sentence-pair datasets) in multilingual learning community (Yang et al., 2019a; Feng et al., 2020; Pan et al., 2021), which have not been utilized for learning sentence representations. Furthermore, given parallel translation pairs, previous contrastive learning frameworks (Radford et al., 2021; Gao et al., 2021) cannot well balance the alignment and uniformity (Wang & Isola, 2020) of monolingual sentence embeddings, where alignment calculates the expected distance between positive embeddings and uniformity measures how well the embeddings are uniformly distributed.

In this paper, we build on the top of dual encoder (Radford et al., 2021; Yang et al., 2019b), and adopt a similar strategy as Frozen (Tsimpoukelli et al., 2021), where we freeze the source language encoder and only train the target language encoder for better monolingual sentence embeddings. The source language encoder is fixed and provides consistent embeddings to supervise the target language encoder via contrastive learning. The corresponding source-target translation pairs are regarded as positives.

We conduct a comprehensive evaluation protocol following SimCSE (Gao et al., 2021) on seven standard semantic textual similarity (STS) tasks (Agirre et al., 2012; 2013; Marelli et al., 2014; Agirre et al., 2014; 2015; 2016; Cer et al., 2017). We outperforme SimCSE (Gao et al., 2021) and Sentence-T5 (Ni et al., 2021) by a large margin. On the average score of STS tasks, our pre-trained BERT$_{base}$ with or without fine-tuning surpasses SimCSE-BERT$_{base}$ by 4.39% and 3.25% respectively, and RoBERTa$_{large}$ achieves 85.58 on average. Surprisingly, BERT$_{base}$ with fine-tuning achieves better results than Sentence-T5 (11B) with only 1% parameters in comparison.

In summary, we provide an exploration of utilizing existing large-scale parallel translation pairs for learning monolingual sentence representation, based on cross-lingual contrastive learning framework that well balances alignment and uniformity.Our approach achieves a new state-of-the-art on standard semantic textual similarity (STS), and the best performance in corresponding tracks on transfer tasks evaluated by SentEval[1].

2. Proposed Approach

2.1. Background

Scaling up the size of training dataset (Radford et al., 2021; Jia et al., 2021) has proved to be effective to improve robustness and generalization of representations in contrastive learning framework. However, previous works (Reimers & Gurevych, 2019; Gao et al., 2021) only utilize limited size of monolingual sentence pairs to learn sentence em-

[*] Equal contribution [1] Zhongnan University of Economics and Law [2] Kuaishou Technology [3] University of Chinese Academy of Sciences. Correspondence to: Jue Wang <201821090281@stu.zuel.edu.cn>.

First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1] https://github.com/facebookresearch/SentEval

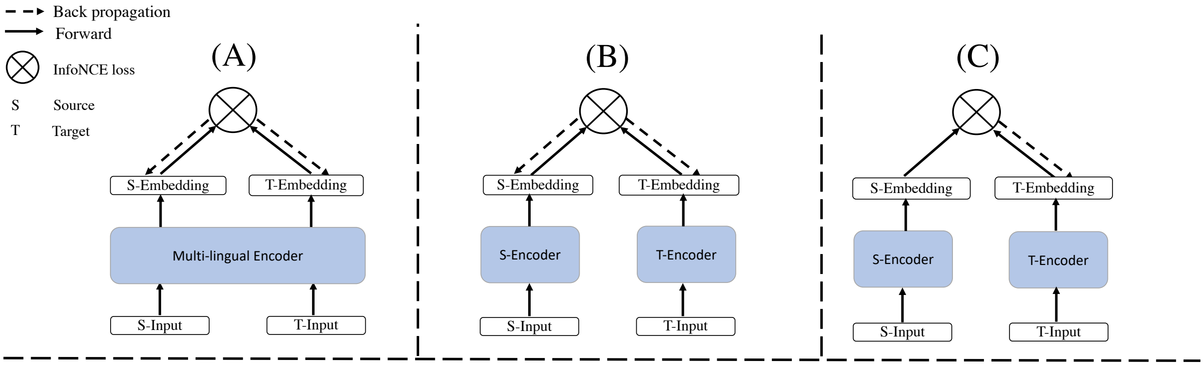

Figure 1. **Comparison of preliminaries and our approach for utilizing parallel translation pairs.** (A), (B) and (C) represent a multilingual encoder, dual encoder and our modified dual encoder, respectively.

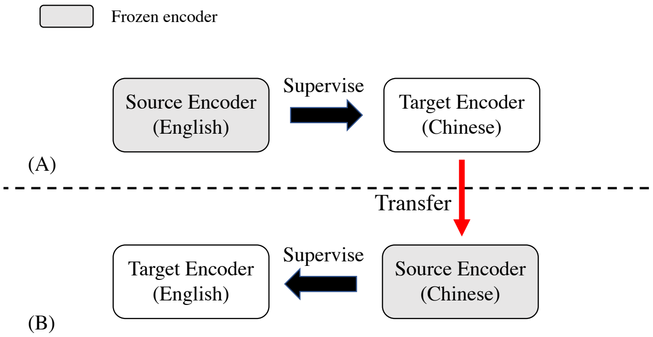

Figure 2. **Training pipeline.** We first obtain a target (Chinese) encoder given a pre-trained SimCSE model as the source encoder. Then, we take the pre-trained Chinese encoder as the source encoder and freeze it to supervise a target (English) encoder. Step (A) and step (B) both follow our proposed framework.

beddings, such as MNLI datasets (Williams et al., 2017) and SNLI (Bowman et al., 2015). In contrast, there have existed large-scale well-annotated parallel translation pairs (100x larger than monolingual paired datasets) in the community of multilingual learning. Instead of training on limited monolingual sentence pairs, utilizing existing parallel translation datasets shows better flexibility and a potential to further improve the performance of sentence embeddings, where a parallel translation pair that is highly correlated in semantic can be treated as a positive sample.

Preliminaries. To utilize paired inputs, single multilingual encoder (Ma et al., 2020; Pan et al., 2021) and dual encoder (He et al., 2020; Radford et al., 2021; Ni et al., 2021) are the most commonly adopted strategies for learning multilingual representations. Multilingual encoder embeds sentences from different languages into a single semantic space using a unified encoder, based on the hypothesis that multilingual learning leads to better multilingual sentence representation. Its architecture is illustrated in A, Figure 1. Dual encoder, also known as two-tower, models the paired data with two independent encoders, and projects the embeddings of paired inputs into the same semantic space through joint training. Its architecture is illustrated in B, Figure 1.

2.2. Method

Although multilingual encoder and dual encoder can use parallel translation pairs straightforwardly, they both suffer from the imbalance between alignment and uniformity, as source language encoder and target language encoder keep updating in the training process. In other words, while they pull the positive samples (source-target translation pairs) closer and the negative samples (source-non target translation pairs) farther away through an explicit contrastive learning objective, the alignment and uniformity of embeddings from monolingual sentence pairs cannot be guaranteed. Specifically, let (s_i, t_i) denote the representation of a parallel translation pair generated by the source language encoder and target language encoder, respectively. We simplify the explicit contrastive objective as Eq 1.

$$L_{explicit} = \alpha_1 * L_p - \alpha_2 * L_n \tag{1}$$

Where L_p and L_n represent the distance for positives and negatives of parallel translation pairs as defined in Eq 2 and Eq 3, α denote the linear weights, D is a distance function, and $i \neq j$. The explicit contrastive objective is to minimize the distance between positives and maximize the distance between negatives.

$$L_p = D(s_i, t_i) + D(s_j, t_j) \tag{2}$$

$$L_n = D(s_i, t_j) + D(s_j, t_i) \tag{3}$$

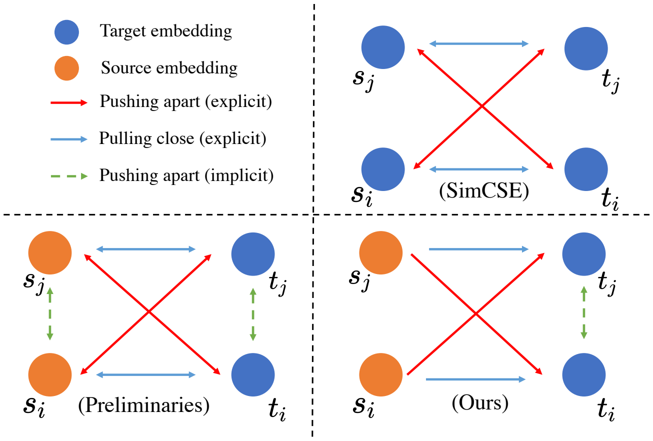

Figure 3. **Illustration of contrastive objectives.** (s_i, t_i) and (s_j, t_j) are two paired samples. In (SimCSE), (s_i, t_i) denotes monolingual pairs, while in (Preliminaries) and (Ours), it denotes parallel translation pairs.

Given parallel translation pairs, we also define the implicit or actual objective that has not been considered into contrastive learning framework in Eq 4, which measures the alignment and uniformity of monolingual sentence embeddings. Although $L_{implicit}$ is not considered in the explicit contrastive objective, we expect to retain good alignment and uniformity of monolingual sentence embeddings from the target encoder, as the actual objective is to learn monolingual sentence embeddings from parallel translation pairs.

$$L_{implicit} = \beta_1 * L_p^{'} - \beta_2 * L_n^{'} \tag{4}$$

Where $L_p^{'}$ and $L_n^{'}$ represent the distance for positives and negatives of monolingual pairs as defined in Eq 5 and Eq 6. s_i^+ and t_i^+ represent the monolingual positive samples for s_i and t_i, respectively. β denote linear weights.

$$L_p^{'} = D(s_i, s_i^+) + D(t_i, t_i^+) \tag{5}$$

$$L_n^{'} = D(s_i, s_j) + D(t_i, t_j) \tag{6}$$

In preliminaries, as shown in (A) and (B), Figure 1, the source language encoder keeps updating in training and can not provide consistent supervision for the target language encoder. The implicit objective for preliminaries is Eq 4, where the alignment and uniformity of source embeddings and target embeddings are both required to be implicitly optimized. However, given two independent implicit objectives, it becomes hard to find a local optimum through Eq 1 without any constraints.

To effectively improve the uniformity and retain the alignment simultaneously, and optimize the implicit objective (4) through an explicit objective (1), we propose to soften the implicit objective for better optimization with our modified architecture, built on the top of regular dual encoder. To be clear, we freeze the side of the source language encoder, so that the alignment and uniformity of source embeddings are frozen in the training. In this case, the implicit objective degrades to Eq 7.

$$L_{implicit} = \beta_1 * D(t_i, t_i^+) - \beta_2 * D(t_i, t_j) \tag{7}$$

As the optimization space shrinks and the implicit objective relaxed, finding the local optimal solution becomes easier and more efficient. We show the differences between our approach (C) and preliminaries (A, B) in Figure 1.

2.3. Visualization of alignment and uniformity

To validate the effectiveness of our approach, we take the checkpoint of our model and preliminaries every 100 steps during training and visualize their alignment and uniformity (Wang & Isola, 2020) on a monolingual sentence-pair dataset and parallel translation dataset in Figure 4, training details can be found in 3.4.2. Figure 4, we show the promising results of implicit objective (the alignment and uniformity of target encoder), given monolingual sentence pairs as input, where we greatly improve uniformity and retain a steady alignment, while others dramatically degrade alignment.

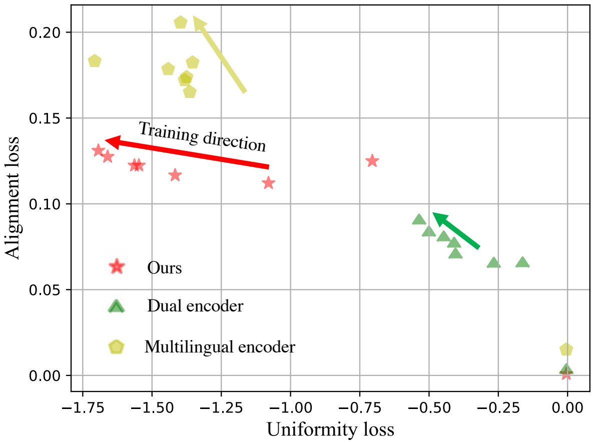

Figure 4. l_{align} and $l_{uniform}$ Plot. Checkpoints are saved every 100 training steps, the arrows indicate the training direction. For both 'align' and 'uniform', lower numbers are better.

3. Experiments

3.1. Training Datasets

We adopt WMT and source-mixed datasets that have parallel translation pairs for cross-lingual contrastive learning, while the Chinese NLI dataset[2] that has monolingual Chinese sentence pairs is only utilized for fine-tuning.

WMT Dataset[3] is a common-used machine translation dataset composed of various sources. We perform an elaborate cleaning process following (Meng et al., 2020) to filter out low-quality pairs. We get 19,442,200 Chinese-English translation parallel pairs after cleaning.

Source-mixed Dataset collects from more open-sourced translation datasets built on the top of WMT dataset, including AIC (Wu et al., 2017), translation2019zh (Xu, 2019), UN Corpus (Ziemski et al., 2016), etc. Finally, we establish a larger-scale dataset including 56,741,808 Chinese-English translation pairs. This dataset is used to show that further scaling up the size of the training set helps improve overall performance.

3.2. Training Details

We elaborate the training details of our pipeline that is shown in Figure 2. We maintain a consistent memory queue (He et al., 2020) of negative embeddings, where the current mini-batch of the source language encoder's embeddings are enqueued and the oldest are dequeued. The pooling method used in the training is [CLS] with an MLP layer following SimCSE. All experiments are conducted on 8 V100 GPUs. The batch size in experiments represents the batch size on each GPU.

3.2.1. Training a Chinese Encoder

As shown in (A), Figure 2, the first step is to train a target language (Chinese) encoder. Specifically, we adopt the pre-trained SimCSE-RoBERTa$_{large}$ model as the source language (English) encoder, and initialize a Chinese RoBERTa$_{large}$ model[4] with pre-trained weights as the target language (Chinese) encoder. We adopt a series of hyperparameters from 3.2.2: learning rate is 5e-5, batch size is 200, dropout is 0.1, and the input sentence length is 50. We freeze the source language (English) encoder and only update the target language (Chinese) model. We evaluate every 250 training steps on the development set of Chinese STS-B and save the best checkpoint.

3.2.2. Training an English Encoder

As shown in B, Figure 2, we train a target language (English) encoder that generates sentence embeddings. Specifically, we reuse the pre-trained Chinese encoder from 3.2.1 as the source language (Chinese) encoder and freeze its parameters. We evaluate every 250 training steps on the development set of STS-B and save the best checkpoint. For BERT$_{base}$ (or RoBERTa$_{base}$), the learning rate is 1e-4, batch size is 400, and the dropout is defaulted set as 0.1. In the term of RoBERTa$_{large}$ (or BERT$_{large}$), we set the learning rate to 5e-5, batch size to 200, all other hyperparameters keep the same as BERT$_{base}$.

3.3. Evaluation Results

Following Gao et al., we evaluate our models on seven transfer and seven STS tasks by SentEval tools. As the main goal of learning sentence embeddings is to cluster semantically similar sentences, we also take STS result as the main metric.

Semantic textual similarity tasks. We evaluate our approach under zero-shot and fine-tuned settings, respectively. To fairly compare with previous works (Gao et al., 2021; Ni et al., 2021), we adopt seven STS tasks including STS 2012–2016 (Agirre et al., 2012; 2013; 2014; 2015; 2016), STS Benchmark (Cer et al., 2017) and SICK-Relatedness (Marelli et al., 2014). STS tasks are widely used in measuring the discriminative power of sentence embeddings.

We compare several well-known benchmarks, include: Sentence-BERT (Reimers & Gurevych, 2019), CT-BERT (Carlsson et al., 2020), SimCSE and Sentence-T5 (Ni et al., 2021) 11B model, which contains 11 billion parameters. Table 1 reports the evaluation results on seven STS tasks. Our approach can substantially improve results on all the datasets with or without extra NLI supervision, greatly outperforming the previous state-of-the-art models. Specifically, our approach outperforms the averaged scores of SimCSE by 1.27-2.65 under a zero-shot setting in all tracks. When using NLI datasets, Ours-BERT$_{base}$ further pushes the state-of-the-art results from 84.94 to 85.15. The gains are more pronounced on RoBERTa encoders, and our method achieves 85.58 with RoBERT$_{large}$.

3.4. Ablation Studies

We investigate the impact of source language encoder and contrastive objectives. We use BERT$_{base}$ (WMT) without fine-tuning as our benchmark.

3.4.1. The effect of source language encoder

To analyze the role of source language encoder, we train a SimCSE-RoBERTa$_{large}$ model on the Chinese NLI dataset directly and use it as the source language (Chinese) encoder. For comparison, we train two RoBERTa$_{large}$ models on the WMT dataset following the steps in 3.2.1 with and without fine-tuning. Then, we train three target language (English) encoders as 3.2.2 given different source language models and evaluate them on the SST-B development set. We report the results in table 2.

[2] https://github.com/pluto-junzeng/CNSD

[3] http://www.statmt.org/wmt20/

[4] https://huggingface.co/hfl/chinese-RoBERTa-wwm-ext-large

Model	Fine-tune data	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg
SBERT$_{base}$	NLI	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT$_{base}$-flow	NLI	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
SBERT$_{base}$-whitening	NLI	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
CT-SBERT$_{base}$	NLI	74.84	83.20	78.07	83.84	77.93	81.46	76.42	79.39
SimCSE-BERT$_{base}$	NLI	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
Ours-BERT$_{base}$(WMT)	-	80.73	85.82	83.20	88.57	82.50	86.60	80.64	84.01
Ours-BERT$_{base}$(SMD)	NLI	80.26	**88.70**	**84.05**	**88.62**	**84.57**	**87.95**	**81.87**	**85.15**
SBERT$_{large}$	NLI	72.27	78.46	74.90	80.90	76.25	79.23	73.75	76.55
SimCSE-BERT$_{large}$	NLI	75.78	86.33	80.44	86.60	80.86	84.87	81.14	82.21
Ours-BERT$_{large}$(WMT)	-	80.71	86.10	83.18	89.13	83.25	86.75	81.43	84.36
Ours-BERT$_{large}$(SMD)	-	79.18	87.75	82.85	88.53	82.60	86.85	81.51	84.18
Ours-BERT$_{large}$(WMT)	NLI	**81.88**	88.78	84.04	88.42	84.94	88.08	81.38	85.36
Ours-BERT$_{large}$(SMD)	NLI	80.86	**89.47**	**84.35**	**88.97**	**85.04**	**88.58**	**81.63**	**85.56**
SRoBERTa$_{base}$-whitening	NLI	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
SimCSE-RoBERTa$_{base}$	NLI	76.53	85.21	80.95	86.03	82.57	85.83	80.50	82.52
Ours-RoBERTa$_{base}$(WMT)	-	**80.59**	85.36	82.16	87.84	82.30	85.96	80.90	83.59
Ours-RoBERTa$_{base}$(SMD)	-	78.60	87.33	83.22	88.64	83.04	86.59	81.15	84.08
Ours-BRoBERTa$_{base}$(WMT)	NLI	80.25	86.97	82.92	87.97	83.78	87.10	81.06	84.29
Ours-RoBERTa$_{base}$(SMD)	NLI	80.02	**87.90**	**83.64**	**88.59**	**85.26**	**87.59**	**81.32**	**84.90**
SRoBERTa$_{large}$	NLI	74.53	77.00	73.18	81.85	76.82	79.10	74.29	76.68
SimCSE-RoBERTa$_{large}$	NLI	77.46	87.27	82.36	86.66	83.93	86.70	81.95	83.76
Ours-RoBERTa$_{large}$(WMT)	-	79.26	87.80	83.76	88.51	83.76	86.94	81.86	84.56
Ours-RoBERTa$_{large}$(SMD)	-	80.86	88.19	84.34	89.20	83.90	87.47	81.26	85.03
Ours-RoBERTa$_{large}$(WMT)	NLI	**81.24**	88.69	84.58	88.59	**85.55**	88.05	82.00	85.53
Ours-RoBERTa$_{large}$(SMD)	NLI	80.07	**89.45**	**84.64**	**88.85**	85.14	**88.60**	**82.28**	**85.58**
ST5-Enc mean (11B)	NLI	77.42	87.50	82.51	87.47	84.88	85.61	80.77	83.74
ST5-EncDec first (11B)	NLI	80.11	88.78	84.33	88.36	85.55	86.82	80.60	84.94
Ours-BERT$_{large}$(SMD)	NLI	**80.86**	**89.47**	84.35	**88.97**	85.04	88.58	81.63	85.56
Ours-RoBERTa$_{large}$(SMD)	NLI	80.07	89.45	**84.64**	88.85	**85.14**	**88.60**	**82.28**	**85.58**

Table 1. **Comparison with previous state-of-the-art works in STS tasks.** All results are from Gao et al., 2021; Ni et al., 2021; Reimers & Gurevych, 2019; WMT and SMD represent the model is trained on WMT dataset and source-mixed dataset, respectively. The pooling methods used for comparison can be found in Appendix **??**, and the Ours-RoBERTa$_{large}$(WMT)'s pooling method is [CLS] with MLP.

Source Encoder	SimCSE$_{CN}$	Ours	Ours+F
STS-B	86.58	86.91	**88.06**

Table 2. Performance of target language encoders given different source language encoders on STS-B development dataset. SimCSE$_{CN}$ represents the Chinese SimCSE-RoBERTa$_{large}$. Ours+F and Ours are RoBERTa$_{large}$ that trained by our strategy with and without fine-tuning, respectively.

Models	Multilingual	Dual	Ours
STS-B	71.02	73.13	**86.82**

Table 3. **The effect of contrastive objectives.** Dual, Multilingual and Ours represent dual encoder, multilingual encoder and our modified dual encoder.

3.4.2. The effect of contrastive objectives

To show the effectiveness of our cross-lingual contrastive learning scheme, we train models with multilingual encoder, dual encoder and our modified dual architecture, respectively, and evaluate their performance on STS-B development set. For dual encoder, we adopt the pre-trained source language (Chinese) encoder from 3.2.1 and a pre-trained RoBERTa$_{base}$, then train it via contrastive learning. For multilingual encoder, we adopt a RoBERTa$_{base}$-xlm (Lample & Conneau, 2019) model that accepts multilingual input. For our modified dual architecture, we use the same source and target encoder as dual encoder, while keeping the source encoder frozen. All models are trained on WMT dataset.

For a fair comparison, we unify the hyperparameters of different objectives: batch size is 128, learning rate is 2e-4. The only difference between dual encoder and ours is whether the source language encoder is frozen in the training. Table 3 shows the effectiveness of our approach.

4. Conclusion

In this work, we provide the first exploration of utilizing existing large-scale parallel translation pairs for learning sentence representation, propose a modified dual architecture that well balances the alignment and uniformity of embeddings. We demonstrated that our method achieves a new state-of-the-art on standard semantic textual similarity (STS), and the best performance on corresponding tracks on transfer tasks, outperforming both SimCSE and Sentence-T5.

References

Agirre, E., Bos, J., Diab, M., Manandhar, S., Marton, Y., and Yuret, D. * sem 2012: The first joint conference on lexical and computational semantics–volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation (semeval 2012). In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 2012.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. * sem 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pp. 32–43, 2013.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 81–91, 2014.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., et al. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pp. 252–263, 2015.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez Agirre, A., Mihalcea, R., Rigau Claramunt, G., and Wiebe, J. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics), 2016.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Carlsson, F., Gyllensten, A. C., Gogoulou, E., Hellqvist, E. Y., and Sahlgren, M. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*, 2020.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, 2020.

Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

Lample, G. and Conneau, A. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Ma, S., Yang, J., Huang, H., Chi, Z., Dong, L., Zhang, D., Awadalla, H. H., Muzio, A., Eriguchi, A., Singhal, S., et al. Xlm-t: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders. *arXiv preprint arXiv:2012.15547*, 2020.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., et al. A sick cure for the evaluation of compositional distributional semantic models. In *Lrec*, pp. 216–223. Reykjavik, 2014.

Meng, F., Yan, J., Liu, Y., Gao, Y., Zeng, X., Zeng, Q., Li, P., Chen, M., Zhou, J., Liu, S., et al. Wechat neural machine translation systems for wmt20. *arXiv preprint arXiv:2010.00247*, 2020.

Ni, J., Constant, N., Ma, J., Hall, K. B., Cer, D., Yang, Y., et al. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*, 2021.

Pan, X., Wang, M., Wu, L., and Li, L. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. Multimodal few-shot learning with frozen language models. *arXiv preprint arXiv:2106.13884*, 2021.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Williams, A., Nangia, N., and Bowman, S. R. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017.

Xu, B. Nlp chinese corpus: Large scale chinese corpus for nlp, September 2019. URL https://doi.org/10.5281/zenodo.3402023.

Yang, Y., Abrego, G. H., Yuan, S., Guo, M., Shen, Q., Cer, D., Sung, Y.-H., Strope, B., and Kurzweil, R. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *arXiv preprint arXiv:1902.08564*, 2019a.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., et al. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*, 2019b.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3530–3534, 2016.