SKICSE: Sentence Knowable Information Prompted by LLMs Improves Contrastive Sentence Embeddings

Fangwei Ou, Jinan Xu
Beijing Key Lab of Traffic Data Analysis and Mining,
Beijing Jiaotong University, Beijing, China
{21120384,jaxu}@bjtu.edu.cn

Abstract

Contrastive learning, which utilizes positive pairs and in-batch negatives to optimize the loss objective, has been proven to be an effective method for learning sentence embeddings. However, we argue that the previous methods of constructing positive pairs only through dropout perturbation or entailment relation are limited. Since there is more sentence knowable information (SKI) to be mined, such as sentence external knowledge, semantic analysis, and grammatical description. In this work, we first hand-craft a simple and effective prompt template that is able to obtain the knowable information of input sentences from LLMs (e.g., LLaMA). Then we combine the original sentence and its knowable information to form a positive pair for contrastive learning. We evaluate our method on standard semantic textual similarity (STS) tasks. Experimental results show that our unsupervised and supervised models using $\text{BERT}_{\text{base}}$ achieve an average of 78.65% and 82.45% Spearman's correlation respectively, a 2.40% and 0.88% improvement compared to SimCSE. Our model outperforms the previous state-of-the-art model PromptBERT in both unsupervised and supervised settings and specifically yields a new state-of-the-art performance in supervised setting.

1 Introduction

Learning sentence embeddings is a fundamental task of natural language processing (NLP), which embeds sentences of natural language text into high-dimensional dense vectors containing rich semantic information. High-quality sentence representations find applications across various practical domains, including question answering systems, translation systems, recommendation systems, and retrieval systems.

In recent years, Transformer-based (Vaswani et al., 2017) pre-trained language models such as BERT (Devlin et al., 2018) have achieved remarkable results in NLP. However, Reimers and Gurevych (2019) found that the performance of BERT without fine-tuning is even inferior to GloVe (Pennington et al., 2014) on STS tasks (Agirre et al., 2012, 2013, 2014, 2015, 2016; Cer et al., 2017; Marelli et al., 2014), and proposed to train SBERT through siamese network structures and supervised data such as NLI (Bowman et al., 2015; Williams et al., 2017), STS-B, and MRPC (Dolan et al., 2004). Li et al. (2020) analyzed the dilemma of native BERT from the perspective of anisotropic sentence embedding distribution, and proposed the corresponding improved method BERT-flow. Gao et al. (2021) proposed SimCSE, a simple contrastive sentence embedding framework, which improves the sentence vector space in terms of alignment and uniformity (Wang and Isola, 2020), and has made great progress on STS tasks.

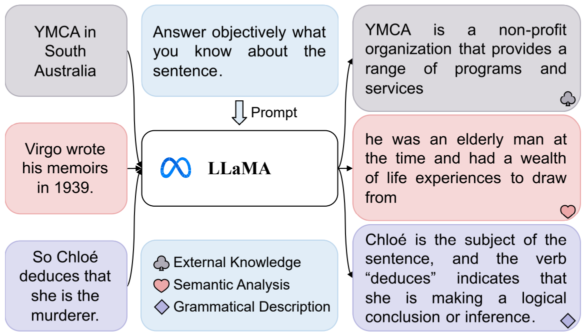


Figure 1: On the left are three training example sentences of SimCSE. Their exclusive SKI on the right is obtained through the prompt template and LLaMA2-7B. Notice that both the template and the SKI are excerpts.

Witnessing the notable success of SimCSE on STS tasks, many variations (Wu et al., 2021; Jiang et al., 2022; Zhang et al., 2022; Chuang et al., 2022; Wu et al., 2022) of SimCSE have been introduced by researchers. Although many of them have novel ideas and methods, few of them can adapt to both

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 141–146

unsupervised and supervised scenarios. Another prevalent issue among them is the way to construct positive pairs, which often relies solely on minimal data augmentation[1] (MDA). We think that more knowable information of sentences can be mined to construct positive pairs to enhance the knowledge, semantics and grammar of sentence representations.

Recently, LLMs such as ChatGPT (Ouyang et al., 2022; OpenAI, 2023) and LLaMA (Touvron et al., 2023) have attracted widespread attention. By leveraging the comprehension and generation capabilities of LLMs, coupled with our effective handcrafted prompt template, we are able to obtain knowable information about input sentences, as shown in Figure 1. We further use input sentences and generated sentences as positive pairs to compute the contrastive loss, and make a trade-off with the original loss through the weighting coefficient.

Our main contributions can be summarized as the following two points:

- We propose to use sentence knowable information mined by LLMs to form positive pairs with original sentences to enhance sentence representations. Our approach to construct positive pairs is an excellent complement to the previous ones that mainly focused on minimal data augmentation.

- Our proposed method works on both unsupervised and supervised settings, weighing our additional contrastive loss against the original ones, resulting in extraordinary improvements. We yield a new state-of-the-art performance on STS tasks in supervised setting based on $\text{BERT}_{\text{base}}$.

2 Related Work

2.1 Contrastive Objective

Contrastive learning can effectively improve the sentence vector space by pulling semantically related vectors closer while pushing apart semantically irrelevant ones.

SimCSE (Gao et al., 2021), by applying the standard dropout twice, obtains two different embeddings as positive pairs. ESimCSE (Wu et al., 2021) proposes word repetition and momentum contrast applied on positive and negative pairs separately to enhance SimCSE. PromptBERT (Jiang et al., 2022) reformulates the output way of sentence embeddings as a fillin-the-blanks problem based on prompt templates. SemCSR (Wang et al., 2023) also uses LLMs as tools, but they generate pseudo-NLI data and filter low-quality data through the evaluation capabilities of LLMs.

2.2 Integrate with Other Objectives

Many researchers inject other learning objectives to conduct a multi-task learning based on the traditional contrastive objective, or transform it.

DiffCSE (Chuang et al., 2022) uses additional generator and discriminator to perform the Replaced Token Detection task with the cross-entropy loss. InfoCSE (Wu et al., 2022) designs an auxiliary network for MLM task to force the representation of `[CLS]` positions to aggregate denser sentence information. ArcCSE (Zhang et al., 2022) models pairwise and triple-wise sentence relations with Additive Angular Margin Contrastive Loss and Triplet Loss. AnglE (Li and Li, 2023) introduces angle optimization which mitigates the adverse effects of the saturation zone in the cosine function.

3 Methodology

3.1 Prompt Template for SKI

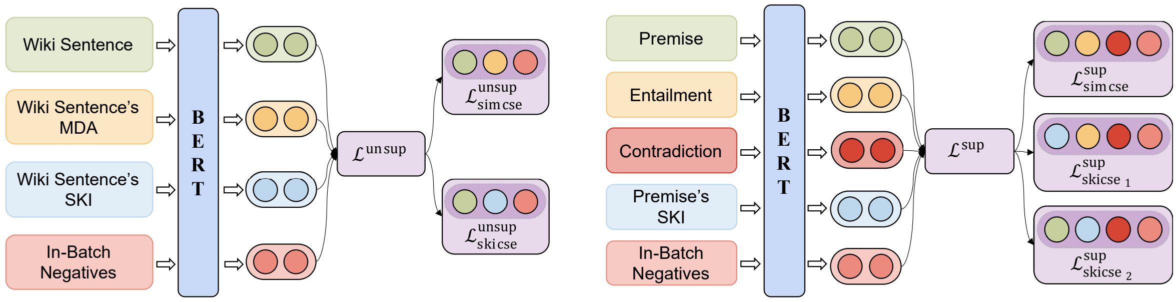
We design the prompt template, "1) *Answer objectively what you know about the sentence.* 2) *Make sure your answers are no more than four sentences and contain important information.*", to obtain the SKI of input sentences.

The first sentence is the core of the prompt template. We find that when we ask LLMs whether they know anything about the sentence we input, they do their best to answer from the three aspects we summarized in Figure 1. If there are entities in the input sentence that contain external knowledge, LLMs will explain and supplement them. Otherwise, LLMs will perform semantic and grammatical analysis of the sentence. The word "objectively" is intended to alleviate hallucinations (Huang et al., 2023) in LLMs. The purpose of the second sentence is to keep LLMs' answers from being overwhelming and to emphasize that the answers should not be irrelevant information.

3.2 Introduce SKICSE

Our SKICSE can be seen as combining the original objective from SimCSE with the additional

[1] This expression originates from SimCSE, where dropout is characterized as a form of minimal data augmentation.



(a) Unsupervised SKICSE

(b) Supervised SKICSE

Figure 2: An illustration for the composition of unsupervised SKICSE loss and supervised ones.

contrastive learning objective which leverages SKI.

3.2.1 Unsupervised SKICSE

Given an unlabeled input sentence x, SKICSE creates a positive example x^{ski} for x by obtaining its SKI. We can constitute a triplet of sentences (x, x', x^{ski}) as shown in Figure 2(a). Here, x and x' have the same text, but different hidden dropout masks. By using the $\text{BERT}_{\text{base}}$ encoder f, we can get a triplet of sentence embeddings $(f(x), f(x'), f(x^{\text{ski}})) = (\mathbf{h}, \mathbf{h}', \mathbf{h}^{\text{ski}})$, and objective functions can be formulated as:

$$\mathcal{L}^{\text{unsup}}_{\text{simcse}} = -\log \frac{e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}'_i\right)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}'_j\right)/\tau}}, \tag{1}$$

$$\mathcal{L}^{\text{unsup}}_{\text{skicse}} = -\log \frac{e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}^{\text{ski}}_i\right)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}^{\text{ski}}_j\right)/\tau}}, \tag{2}$$

where N is the batch size for the input batch $\{x_i\}_{i=1}^{N}$, τ is a temperature hyperparameter, and $\text{sim}(\cdot, \cdot)$ is the cosine similarity function.

Finally, the final objective function of unsupervised SKICSE is the weighted summary of the aforementioned two objectives:

$$\mathcal{L}^{\text{unsup}} = (1-\lambda)\mathcal{L}^{\text{unsup}}_{\text{simcse}} + \lambda \mathcal{L}^{\text{unsup}}_{\text{skicse}}, \tag{3}$$

where the weight λ is a balanced hyperparameter and reflects the importance of SKI.

3.2.2 Supervised SKICSE

In NLI datasets, for each premise x, there are its entailment hypothesis x^+ and an accompanying contradiction x^-. SKICSE creates a positive example x^{ski} for x by obtaining its SKI. Similarly, we can constitute a quadruplet of sentences $(x, x^+, x^-, x^{\text{ski}})$ and pass it through the encoder to get a quadruplet of sentence embeddings $(\mathbf{h}, \mathbf{h}^+, \mathbf{h}^-, \mathbf{h}^{\text{ski}})$ as shown in Figure 2(b). Objective functions can be formulated as:

$$\mathcal{L}^{\text{sup}}_{\text{simcse}} = -\log \frac{e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}^+_i\right)/\tau}}{\sum_{j=1}^{N} \left(e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}^+_j\right)/\tau} + e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}^-_j\right)/\tau}\right)}, \tag{4}$$

$$\mathcal{L}^{\text{sup}}_{\text{skicse}_1} = -\log \frac{e^{\text{sim}\left(\mathbf{h}^{\text{ski}}_i, \mathbf{h}^+_i\right)/\tau}}{\sum_{j=1}^{N} \left(e^{\text{sim}\left(\mathbf{h}^{\text{ski}}_i, \mathbf{h}^+_j\right)/\tau} + e^{\text{sim}\left(\mathbf{h}^{\text{ski}}_i, \mathbf{h}^-_j\right)/\tau}\right)}, \tag{5}$$

$$\mathcal{L}^{\text{sup}}_{\text{skicse}_2} = -\log \frac{e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}^{\text{ski}}_i\right)/\tau}}{\sum_{j=1}^{N} \left(e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}^+_j\right)/\tau} + e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}^-_j\right)/\tau}\right)}, \tag{6}$$

In a similar way, the final objective function of supervised SKICSE becomes:

$$\mathcal{L}^{\text{sup}} = (1-\lambda_1-\lambda_2)\mathcal{L}^{\text{sup}}_{\text{simcse}} + \lambda_1 \mathcal{L}^{\text{sup}}_{\text{skicse}_1} + \lambda_2 \mathcal{L}^{\text{sup}}_{\text{skicse}_2}. \tag{7}$$

4 Experiments

4.1 Setup

Training Details We adapt SimCSE codebase[2] to our experimental settings and start from the pre-trained checkpoint of *bert-base-uncased* from the Huggingface model repository[3]. The LLM we use to generate SKI is LLaMA2-7B. More training details are shown in Appendix A.

Datasets We use the source data provided by SimCSE as training data. We train unsupervised SKICSE on 10^6 randomly sampled sentences from English Wikipedia, and train supervised SKICSE on the combination of MNLI and SNLI datasets. The model with the highest performance on STS-B development set will be chosen. We conduct our experiments on 7 STS tasks, which evaluate

[2] https://github.com/princeton-nlp/SimCSE
[3] https://huggingface.co/models

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
				Unsupervised Models				
ConSERT	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
SimCSE	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
SemCSR	69.63	82.61	76.61	82.67	80.23	80.86	**73.66**	78.04
ArcCSE	72.08	84.27	76.25	82.32	79.54	79.92	72.39	78.11
ESimCSE	**73.40**	83.27	**77.25**	82.66	78.81	80.17	72.30	78.27
DiffCSE	72.28	84.43	76.47	83.90	80.54	80.59	71.23	78.49
PromptBERT	71.56	84.58	76.98	84.47	80.60	81.60	69.87	78.54
SKICSE (Ours)	72.92	84.11	76.81	82.18	80.45	80.69	73.38	78.65
InfoCSE	70.53	**84.59**	76.40	**85.10**	**81.95**	**82.00**	71.37	**78.85**
				Supervised Models				
SBERT	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
ConSERT	74.07	83.93	77.05	83.66	78.76	81.36	76.77	79.37
SimCSE	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
PromptBERT	75.48	85.59	80.57	85.99	81.08	84.56	80.52	81.97
AnglE	75.09	85.56	80.66	**86.44**	**82.47**	**85.16**	**81.23**	82.37
SKICSE (Ours)	**75.79**	**86.14**	**81.47**	86.13	82.05	85.08	80.48	**82.45**

Table 1: Sentence embedding performance on STS tasks. All models use $\text{BERT}_{\text{base}}$ as the backbone and results are from their own papers.

whether the semantic similarity between two sentences predicted by a model is relevant to human judgments. And Spearman's correlation coefficient is used to report the model performance.

Baselines We compare unsupervised and supervised SKICSE to previous state-of-the-art sentence embedding methods on STS tasks. These strong baselines include SBERT (Reimers and Gurevych, 2019), ConSERT (Yan et al., 2021), SimCSE (Gao et al., 2021), ESimCSE (Wu et al., 2021), PromptBERT (Jiang et al., 2022), DiffCSE (Chuang et al., 2022), InfoCSE (Wu et al., 2022), ArcCSE (Zhang et al., 2022), SemCSR (Wang et al., 2023), AnglE (Li and Li, 2023).

4.2 Results

The experimental results of STS tasks are shown in Table 1. It can be seen that few variants of SimCSE can adapt to both unsupervised and supervised scenarios. However, our SKICSE is not only suitable for both scenarios but also achieves great improvement, obtaining a 2.40% and 0.88% absolute increase compared to SimCSE on average Spearman's correlation. It is worth mentioning that such performance is rare, and previously only PromptBERT has come close to reaching our results in both scenarios. To the best of our knowledge, we yield a new state-of-the-art performance in supervised setting with $\text{BERT}_{\text{base}}$ as the backbone.

SemCSR also makes use of LLMs. But what it does is to generate the entailment and contradiction of a given sentence to obtain pseudo-NLI triplets. Our unsupervised results exceed SemCSR by 0.61% absolute point, even though it is actually performing weakly supervised training with pseudo-NLI data. According to SemCSR's paper, the result will drop significantly to 75.59% if the generated pseudo-NLI data is not evaluated and filtered. In contrast, our generated SKI requires no additional processing for the model to produce satisfactory results.

5 Conclusion

In this paper, we propose a novel concept called sentence knowable information (SKI). It is an excellent complement to positive pairs constructed by minimal data augmentation and entailment relation. Owing to the powerful generation capabilities of LLMs and our effectively handcrafted prompt template, we mine SKI whose main content is external knowledge, semantic analysis, and grammatical description. SKI is injected into the model through an additional standard contrastive learning objective to better optimize the sentence vector space. Experimental results on STS tasks show that our method can achieve better performance than almost all sentence embedding strong baselines.

Acknowledgments

The research work described in this paper has been supported by the National Nature Science Foundation of China (No. 62376019, 61976015, 61976016, 61876198 and 61370130).

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. 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)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. 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).

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. 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)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * 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*, pages 32–43.

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

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. *arXiv preprint arXiv:2204.10298*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

William Dolan, Chris Quirk, Chris Brockett, and Bill Dolan. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

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. *arXiv preprint arXiv:2201.04337*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Huiming Wang, Liying Cheng, Zhaodonghui Li, De Wen Soh, and Lidong Bing. 2023. Semantic-aware contrastive sentence representation learning with large language models.

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

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

Xing Wu, Chaochen Gao, Zijia Lin, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. Infocse: Information-aggregated contrastive learning of sentence embeddings. *arXiv preprint arXiv:2210.06432*.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2021. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. *arXiv preprint arXiv:2109.04380*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4892–4903.

A Training Details

For both unsupervised and supervised SKICSE, we set the batch size as 512, learning rate as 1e-4, max sequence length as 128. We keep these parameter settings unchanged and search for weight coefficients. Empirically, we find that $\lambda = 0.15$ for the unsupervised SKICSE and $\lambda_1 = 0.1$, $\lambda_2 = 0.3$ for the supervised SKICSE work well. There are two weight coefficients in supervised setting and we carry out grid-search of $\lambda_1, \lambda_2 \in \{0.1, 0.2, 0.3\}$ on STS-B development set as shown in Table 2.

$\boldsymbol{\lambda_1}$	$\boldsymbol{\lambda_2}$ 0.1	0.2	0.3
0.1	86.3196	86.3368	**86.3412**
0.2	86.2750	86.3102	86.3189
0.3	86.2640	86.3302	86.3384

Table 2: STS-B development results (Spearman's correlation) with different combinations of λ_1 and λ_2.

We run the experiments on a server with 60 vCPU AMD EPYC 7543 32-Core Processor and 4 NVIDIA A40 GPUs. The system operates on Ubuntu 18.04 with Python 3.8, PyTorch torch1.7.1+cu110, and Transformers 4.2.1. The training of unsupervised and supervised SKICSE take approximately 35 and 30 minutes respectively.