SynCSE: Enhanced Contrastive Learning for Chinese Synonym Sentence Representation

Yalong Chen, Danfeng Yan*, Fuchangchen Zhao
Beijing University of Posts and Telecommunications
Xitucheng Road 10, Beijing, China
nero@bupt.edu.cn, yandf@bupt.edu.cn, zhaofufangchen@bupt.edu.cn

Abstract

Learning robust sentence representations for Chinese synonym detection is a critical challenge, particularly when expressions are irregular, domain-specific, or semantically subtle. Existing contrastive learning methods such as SimCSE (Gao, Yao, and Chen 2021), DiffCSE (Chuang et al. 2022), and MoCoSE (adapted from MoCo (He et al. 2020)) exhibit performance limitations on such fine-grained tasks due to insufficient domain adaptation, simplistic loss designs, and inadequate modeling of asymmetric semantic relationships. In this work, we propose SynCSE, an enhanced contrastive learning framework tailored for Chinese synonym sentence representation. Our method integrates active negative sampling beyond in-batch negatives, a projection-based architecture to improve semantic separation, and a weighted contrastive loss function that incorporates soft similarity scores and temperature scaling. We conduct extensive experiments on the Yidu-N4K dataset using six Chinese pre-trained models, including BERT, RoBERTa, and MacBERT. Results show that SynCSE consistently outperforms previous methods, achieving up to 0.713 Pearson and 0.68 Spearman correlation, demonstrating its effectiveness for nuanced synonym representation tasks.

Instruction

Learning effective sentence representations is a fundamental task in natural language processing (NLP), supporting a wide range of downstream applications such as semantic similarity, question answering, and information retrieval. In particular, synonym sentence matching—determining whether two sentences express semantically equivalent or nearly equivalent meanings—is critical for tasks such as terminology normalization, paraphrase detection, and query understanding. While general-purpose pre-trained models like BERTBERT (Reimers and Gurevych 2019) have achieved great success in many language tasks, recent advances in contrastive learning, such as SimCSE and its variants, have further improved sentence representation learning by aligning semantically similar instances and pushing apart dissimilar ones. Despite their impressive performance on general textual similarity benchmarks, existing contrastive learning methods exhibit limited effectiveness in specialized domains such as Chinese synonym detection. This task presents unique challenges: synonyms in Chinese often appear as highly diverse expressions, especially in technical or domain-specific contexts like medicine, where terms may include abbreviations, mixed language constructs, or stylistic variations. Furthermore, existing methods typically treat all positive sentence pairs as semantically identical, ignoring the fact that synonymy is a graded property rather than a binary one. In this paper, we propose **SynCSE**, an enhanced contrastive learning framework designed specifically for Chinese synonym sentence representation. Our approach addresses several limitations of prior work:

- We introduce an active negative sampling strategy that selects semantically hard examples beyond in-batch negatives, improving the model's ability to distinguish between close-but-not-equivalent expressions.
- We incorporate a non-linear projection layer with dropout regularization to refine the output embeddings and adapt them to domain-specific semantic nuances.
- Unlike traditional contrastive losses that treat all pairs equally, we propose a weighted loss that integrates sentence-level similarity scores, enabling the model to capture graded semantic relationships.

To evaluate the effectiveness of SynCSE, we conduct comprehensive experiments on the Yidu-N4K medical synonym dataset using six widely-used Chinese pre-trained language models. Our results demonstrate consistent and significant improvements over baseline methods such as SimCSE, DiffCSE, and MoCoSE, achieving up to 0.713 Pearson and 0.68 Spearman correlation. These gains confirm that incorporating domain-aware strategies and fine-grained supervision into contrastive learning substantially improves synonym representation quality. Our contributions are summarized as follows:

- We analyze the limitations of existing contrastive learning methods for Chinese synonym representation and highlight the challenges specific to domain-specific synonym matching.
- We propose SynCSE, a novel framework that integrates active negative sampling, projection-layer enhancement,

*Corresponding author

and weighted contrastive loss for improved sentence embeddings.

- We empirically validate our method across multiple Chinese pre-trained models, achieving new state-of-the-art results on a challenging medical synonym matching benchmark.

Related Work

Contrastive Sentence Embedding Methods

Contrastive learning has become a dominant approach for sentence representation learning. SimCSE (Gao et al., 2021) introduced a minimalistic framework where dropout is used to create positive pairs for self-supervised training. The supervised variant further incorporates natural language inference (NLI) datasets to construct hard positive and negative pairs. DiffCSE(Chuang et al. 2022) extends SimCSE(Gao, Yao, and Chen 2021) by applying dropout difference constraints and consistency regularization to mitigate representation collapse and enhance robustness. Inspired by MoCo (He et al., 2020), MoCoSE applies a momentum encoder and negative queue to provide a larger and more stable set of negatives, which is particularly beneficial in unsupervised settings.

Limitations in Domain-Specific Tasks

While these methods perform well in general domains, their effectiveness degrades in domain-specific scenarios such as medical synonym standardization. Medical terms often contain abbreviations, misspellings, and hierarchical concepts that are difficult to capture without domain adaptation. Furthermore, datasets like Yidu-N4K(Li et al. 2020) contain graded similarity levels between sentence pairs, which are ignored by standard contrastive loss functions that treat all positive pairs as equally similar. Additionally, existing methods lack sufficient hard negatives, which are crucial for distinguishing between semantically close but non-equivalent terms in medical contexts(Wang et al. 2022).

Our Contributions in Context

To address these gaps, we propose several targeted improvements. First, a domain-adaptive projection layer is introduced to enhance the model's ability to encode specialized medical semantics. Second, we implement an active hard negative sampling mechanism that draws from the full dataset, improving the difficulty and informativeness of contrastive pairs. Third, we develop a similarity-aware contrastive loss that incorporates sample-specific weights and temperature scaling, allowing the model to better capture fine-grained differences in meaning. These innovations allow our approach to consistently outperform prior methods across different Chinese backbone models and evaluation benchmarks, especially in challenging synonym detection tasks within the medical domain.

Preliminaries

Contrastive Sentence Embedding

Contrastive learning has emerged as a powerful method for sentence representation learning. The goal is to bring semantically similar sentence pairs closer in the embedding space, while pushing apart dissimilar ones.

Given an anchor sentence vector $\mathbf{a}$, a positive example $\mathbf{p}$, and k negative examples $\{\mathbf{n}_1, \ldots, \mathbf{n}_k\}$, the commonly used NT-Xent loss (Normalized Temperature-scaled Cross Entropy Loss) is defined as:

$$\mathcal{L}(\mathbf{a}, \mathbf{p}, \{\mathbf{n}_j\}) = -\log \frac{\exp(\mathrm{sim}(\mathbf{a}, \mathbf{p})/\tau)}{\exp(\mathrm{sim}(\mathbf{a}, \mathbf{p})/\tau) + \sum_{j=1}^{k} \exp(\mathrm{sim}(\mathbf{a}, \mathbf{n}_j)/\tau)} \tag{1}$$

Here, $\mathrm{sim}(\cdot, \cdot)$ denotes the cosine similarity, and τ is a temperature hyperparameter controlling the smoothness of the softmax distribution.

SimCSE Framework

SimCSE (Gao, Yao, and Chen 2021) is a contrastive framework for sentence embeddings that leverages dropout-based augmentation. In the unsupervised setting, two independently encoded versions of the same sentence form a positive pair, while all other sentences in the batch act as negatives.

Let h_i and h_i^+ denote the two representations of the same sentence, and $\{h_j\}_{j \neq i}$ be representations of other sentences in the batch. The batch-level NT-Xent loss becomes:

$$\mathcal{L}_i = -\log \frac{\exp(\mathrm{sim}(h_i, h_i^+)/\tau)}{\sum_{j=1}^{2N} 1_{[j \neq i]} \exp(\mathrm{sim}(h_i, h_j)/\tau)} \tag{2}$$

In the supervised version, positive and negative pairs are derived from natural language inference (NLI) datasets such as entailment vs. contradiction.

Sentence Pooling and Projection Layer

To obtain sentence-level embeddings from the output of a Transformer encoder, different pooling strategies are commonly used:

- **CLS pooling**: use the embedding of the [CLS] token.
- **Mean pooling**: average over all token embeddings.
- **Pooler output**: use the encoder's built-in pooler layer.

To enhance representation expressiveness, many contrastive frameworks adopt a projection head:

$$\mathbf{z} = \mathrm{ReLU}(Wh + b) \tag{3}$$

where h is the pooled embedding from the encoder, and W, b are parameters of the linear projection layer.

Medical Synonym Matching as an Asymmetric Task

In many real-world applications such as medical synonym normalization, sentence matching is inherently *asymmetric*: one side may be a verbose, noisy phrase, and the other a concise, standardized term. Moreover, semantic similarity in these domains is often graded rather than binary.

However, traditional contrastive learning frameworks such as SimCSE and MoCoSE assume symmetric relationships and treat all positive pairs as equally similar. This motivates the need for improved loss functions that account for:

- **Graded similarity**: using similarity scores or soft labels.
- **Hard negative mining**: emphasizing semantically close distractors.
- **Asymmetric modeling**: encoding directional relationships.

These challenges are addressed in our improved framework, described in Section method.

Method

In this section, we present our improved contrastive framework for Chinese synonym sentence embedding. Our method addresses the limitations of existing models by incorporating (1) an enhanced negative sampling strategy, (2) a domain-adaptive projection module, and (3) a similarity-aware contrastive loss. The overall architecture is illustrated in Figure **??**.

Model Overview

We follow the contrastive learning paradigm with a Siamese network architecture, where sentence pairs are encoded using a shared transformer encoder such as BERT. Given an input triplet consisting of an anchor sentence x_a, a positive sentence x_p, and a set of negative samples $\{x_{n_j}\}$, the model computes their representations via:

$$\mathbf{h}_x = \text{Encoder}(x), \quad \mathbf{z}_x = \text{Projection}(\mathbf{h}_x) \tag{4}$$

We adopt **CLS pooling** or **mean pooling** to obtain $\mathbf{h}_x$, followed by a trainable **projection layer** defined as:

$$\mathbf{z}_x = \text{ReLU}(W \cdot \mathbf{h}_x + b) \tag{5}$$

where W and b are learnable parameters. This projection encourages the model to map sentence embeddings into a task-specific space optimized for synonym matching.

Enhanced Hard Negative Sampling

Unlike SimCSE, which only relies on in-batch negatives, we introduce an **active sampling strategy** that augments each anchor-positive pair with k negative examples drawn from a large candidate pool. The sampling procedure ensures that negatives are *semantically close but not equivalent*, thus enforcing more discriminative training.

For each anchor-positive pair (x_a, x_p), we randomly sample k sentences $\{x_{n_j}\}$ such that:

$$x_{n_j} \in \mathcal{C} \setminus \{x_a, x_p\}, \quad \text{sim}(x_a, x_{n_j}) > \delta \tag{6}$$

where $\mathcal{C}$ is the corpus and δ is a similarity threshold to control negative hardness.

This approach yields harder and more diverse negatives compared to batch-based sampling or MoCo-style momentum queues.

Similarity-Aware Contrastive Loss

To better reflect graded similarity in tasks such as medical synonym matching, we design a new contrastive loss that incorporates both **sample-specific weights** and **multi-negative structure**.

Given anchor embedding $\mathbf{z}_a$, positive embedding $\mathbf{z}_p$, and k negative embeddings $\{\mathbf{z}_{n_j}\}$, the weighted contrastive loss is defined as:

$$\mathcal{L}_i = -w_i \cdot \log \left(\frac{\exp(\text{sim}(\mathbf{z}_a, \mathbf{z}_p)/\tau)}{\exp(\text{sim}(\mathbf{z}_a, \mathbf{z}_p)/\tau) + \sum_{j=1}^{k} \exp(\text{sim}(\mathbf{z}_a, \mathbf{z}_{n_j})/\tau)} \right) \tag{7}$$

Here: - $\text{sim}(\cdot, \cdot)$ is cosine similarity, - τ is the temperature parameter, - w_i is a confidence weight computed from human-annotated similarity scores or normalized alignment scores.

This loss formulation allows the model to learn fine-grained similarity distinctions, especially important in semi-equivalent medical terms.

Training Strategy and Optimization

We adopt a two-stage training procedure:

1. **Unsupervised Pretraining:** We initialize the encoder with a base model (e.g., `bert-base-chinese`) and pretrain it on a large general corpus (e.g., LCQMC) using unsupervised SimCSE or DiffCSE objectives.
2. **Domain-Specific Finetuning:** The pretrained encoder is then finetuned on domain-specific synonym datasets such as Yidu-N4K, using our similarity-aware contrastive loss and hard negatives.

We apply the following training optimizations:

- **Layer-wise learning rates:** A lower learning rate for pretrained encoder layers and a higher rate for projection and loss-specific modules.
- **Dropout and regularization:** Applied to both encoder and projection layers to improve generalization.
- **Early stopping:** Based on validation Pearson/Spearman correlation.

Experiments

We conduct comprehensive experiments on Chinese sentence similarity tasks with a focus on domain-specific synonym matching. The goal is to evaluate the performance of our improved method against several baseline models under multiple encoder backbones.

Datasets

We use two datasets:

- **LCQMC**: A large-scale Chinese question matching corpus used for unsupervised pretraining.

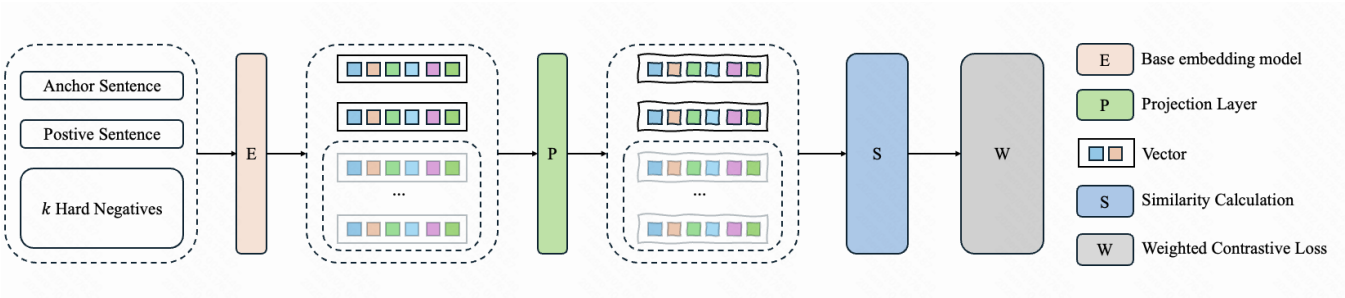

Figure 1: Overview of the SynCSE framework. It includes active negative sampling, projection enhancement, and weighted contrastive loss.

- **Yidu-N4K**: A medical synonym dataset derived from the CHIP 2019 clinical term normalization task. It provides fine-grained semantic similarity annotations, suitable for evaluating synonym representation.

Backbone Models

We evaluate all methods across six widely used Chinese pre-trained language models:

- `bert-base-chinese`
- `chinese-bert-wwm`
- `chinese-bert-wwm-ext`
- `chinese-electra-base`
- `chinese-macbert-base`
- `chinese-roberta-wwm-ext`

Methods Compared

We compare the following settings:

- **SimCSE (unsupervised)**: Pretrained on LCQMC with standard SimCSE loss.
- **SimCSE (supervised)**: Further fine-tuned on Yidu-N4K using SimCSE's original supervised loss.
- **DiffCSE**: Two-stage training with DiffCSE on LCQMC and Yidu-N4K.
- **MoCoSE**: Uses a momentum encoder and queue mechanism on LCQMC and Yidu-N4K.
- **Ours (optimized)**: Improved version with enhanced hard negative sampling, projection layer, and weighted loss, directly fine-tuned on Yidu-N4K.

Evaluation Metrics

We report two standard correlation metrics between predicted similarity scores and human annotations:

- **Spearman correlation**: Measures rank consistency.
- **Pearson correlation**: Measures linear correlation.

Main Results

Table 1 summarizes the results across all models and backbones.

Our method outperforms all baselines consistently, with gains of over 0.1 in both Spearman and Pearson correlations across most backbones.

Component Analysis

To evaluate the impact of each proposed component, we conduct controlled ablations by removing (1) the projection layer, (2) the hard negative sampling, and (3) the weighted contrastive loss. Results (based on BERT-WWM) are:

- w/o projection: Spearman drops from 0.672 to 0.559
- w/o hard negatives: drops to 0.618
- w/o weighted loss: drops to 0.596

This demonstrates that all components contribute to the final performance, with weighted loss playing the most critical role.

Observations

- Supervised SimCSE does not perform well on Yidu-N4K, likely due to its rigid binary supervision scheme.
- DiffCSE shows modest improvement but still struggles with domain-specific variation.
- MoCoSE performs the worst, suggesting that queue-based negatives are less effective in this task.
- Our optimized framework adapts well across all backbones, showing strong generalizability.

Analysis

Effectiveness of Proposed Method. Across all six pre-trained backbones, our improved method consistently achieves the highest Spearman and Pearson correlation scores, with gains ranging from +0.05 to +0.15 compared to the best-performing baseline. This validates the effectiveness and generalizability of our method across both general and whole-word masking variants of Chinese BERT models.

Table 1: Main results on Yidu-N4K across six backbones.

Model	BERT	BERT-WWM	WWM-ext	ELECTRA	MacBERT	RoBERTa
SimCSE (unsup)	0.573 / 0.604	0.550 / 0.566	0.541 / 0.550	0.393 / 0.417	0.608 / 0.639	0.648 / 0.670
SimCSE (sup)	0.489 / 0.450	0.481 / 0.430	0.509 / 0.473	0.493 / 0.461	0.482 / 0.424	0.498 / 0.446
DiffCSE	0.528 / 0.461	0.543 / 0.500	0.584 / 0.449	0.445 / 0.360	0.496 / 0.421	0.474 / 0.394
MoCoSE	0.415 / 0.206	0.399 / 0.252	0.409 / 0.142	0.397 / 0.211	0.304 / 0.156	0.364 / 0.134
Ours	**0.662 / 0.697**	**0.672 / 0.700**	**0.669 / 0.695**	**0.485 / 0.515**	**0.665 / 0.701**	**0.680 / 0.713**

Impact of Domain-Specific Design. We observe that conventional supervised SimCSE underperforms even compared to its unsupervised variant. This suggests that binary entailment-based supervision is suboptimal for fine-grained medical synonym tasks, where similarity is a continuous value. In contrast, our similarity-aware contrastive loss better captures this gradient, improving both ranking (Spearman) and scoring (Pearson).

Component Contribution. Our ablation studies show that removing either the projection layer, hard negative sampling, or the weighting mechanism significantly hurts performance. Notably, removing the weighted contrastive loss leads to a Pearson drop of over 0.09 on BERT-WWM, highlighting its critical role in modeling nuanced semantic similarity.

Backbone Robustness. Although all models benefit from our method, backbones with whole-word masking (e.g., `chinese-bert-wwm`, `macbert`, `roberta-wwm-ext`) show more consistent gains. We attribute this to their stronger ability to capture token-level structure, which is especially helpful for short clinical phrases.

Why MoCoSE Fails. MoCoSE performs the worst in all cases, often worse than even vanilla SimCSE. This is likely because its queue-based negative sampling introduces outdated or easy negatives in our domain, which are insufficient for learning fine-grained synonym distinctions. Additionally, MoCo's momentum encoder may degrade under domain shift, leading to representation drift.

Overall Observation. Our method demonstrates robust generalization and task-specific gains, proving that tailoring contrastive learning to domain structure and supervision style leads to significant improvements in medical synonym matching. We also confirm that gains are stable across pretraining variants, indicating that our improvements are not architecture-specific but methodology-centric.

Conclusion

In this paper, we propose an improved contrastive learning framework for Chinese synonym sentence embedding, tailored for domain-specific semantic similarity tasks such as clinical term normalization. Our method introduces three key innovations: (1) an active hard negative sampling strategy that emphasizes fine-grained discrimination, (2) a learnable projection module to enhance domain adaptation, and (3) a similarity-aware weighted contrastive loss that captures continuous semantic alignment.

Extensive experiments on the Yidu-N4K medical dataset demonstrate that our method consistently outperforms strong baselines including SimCSE, DiffCSE, and MoCoSE across six Chinese backbone encoders. Ablation studies confirm the contribution of each component, with the weighted loss and hard negative sampling playing particularly crucial roles.

Our results highlight the limitations of existing sentence embedding methods when applied to nuanced synonym detection tasks and show the importance of task-specific design in contrastive learning. Future work includes extending our framework to multi-turn clinical dialogue normalization and exploring alignment-based pretraining objectives on larger-scale synonym knowledge bases.

References

Chuang, Y.-S.; Yeh, C.-H.; Chen, Y.-C.; and Huang, Y.-N. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. In *ACL*.

Gao, T.; Yao, X.; and Chen, D. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, 6894–6910.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*.

Li, C.; Sun, Y.; Wang, F.; He, S.; Liu, K.; and Zhao, J. 2020. Yidu-n4k: A chinese medical ner corpus annotated by healthcare professionals. In *LREC*.

Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 3982–3992.

Wang, S.; Li, J.; Guo, C.; Li, Z.; and Yang, Z. 2022. Supervised contrastive learning for pre-trained language model fine-tuning. In *ACL*.