Investigating Data Augmentations in Unsupervised Sentence Embeddings for Biomedical Text

Anonymous ACL submission

Abstract

Unsupervised sentence representation learning is crucial in NLP, with contrastive learning showing notable success. This study concentrates on sentence embeddings in the biomedical domain, employing Bert-base-uncased and Chinese-bert-wwm-ext for English and Chinese text, respectively. We assess our models using BIOSSES and ChineseBLUE benchmarks, marking the first investigation into data augmentation methods for enhancing contrastive learning in biomedical NLP. Our findings reveal that general-purpose natural language pretrained Bert-base models excel in biomedical tasks when fine-tuned with domain-specific texts. By applying various data augmentation techniques, we enhance the contrastive learning of biomedical sentence embeddings. Results show a 4.34% increase in BIOSSES's unsup-SimCSE average Spearman's correlation, and improvements in ChineseBLUE tasks, surpassing state-of-the-art unsup-SimCSE scores. We also establish that augmentation methods preserving sentence constituents, like Punctuation insertion and MixCSE-Instance weighting, yield superior outcomes.

1 Introduction

Recently, the volume of biomedical literature has grown rapidly, and reports containing valuable information on discoveries and new insights continue to be added to an already large body of literature. Therefore, there is increasingly more demand for accurate biomedical text mining tools for extracting information from the biomedical literature.

The advancements of deep learning techniques in natural language processing (NLP) made it possible to develop biomedical text mining models. For example, BERT (Devlin et al., 2019), ERNIE (Sun et al., 2019), XLNet (Yang et al., 2019) and RoBERTa (Liu et al., 2019) with training language models have achieved remarkable success in modeling contextualized word representations using large amounts of training text.

However, like most deep learning architectures, it requires a large amount of labeled data to train whereas task-specific labels in most realistic scenarios are often of limited size (e.g., in the case of medical imaging for example acquiring samples is difficult and in order to create labels professionals have to spend a lot of time and effort to manually classify and segment the images.). Simultaneously, several studies have found that the sentence representations derived by Pretrain Language Models (PLMs) are not uniformly distributed with respect to directions, but instead occupy a narrow cone in the vector space (Ethayarajh, 2019), which largely limits their expressiveness.

To address this issue, researchers use contrastive learning to learn better unsupervised sentence embedding. It aims to learn effective sentence embeddings based on the assumption that effective sentence embeddings should bring similar sentences closer while pushing away dissimilar ones. But, as contrastive learning word representation models such as ConSERT (Yan et al., 2021), SimCLR (Chen et al., 2020) and SimCSE (Gao et al., 2021) are trained and tested mainly on datasets containing general domain texts or English datasets, it is difficult to estimate their performance on datasets containing biomedical texts, especially Chinese medical texts. Also, in the learning process, both positive and negative examples are involved in contrast with the original sentence. For positive examples, previous works apply data augmentation strategies (Yan et al., 2021) on the original sentence to generate highly similar variations. While, negative examples are commonly sampled from the batch or training data (e.g., in-batch negatives (Gao et al., 2021)) at random, due to the lack of ground-truth annotations for negatives. It is likely to hurt the semantics of the sentence representations by simply pushing apart these sampled negatives.

Therefore, in this paper, we aim to tackle the aforementioned challenges in the context of the

biomedical domain. Our goal is to improve unsupervised sentence representation by infusing domain knowledge into the augmentation and contrast schemes. We propose to leverage biomedical domain corpora to assist contrastive learning on biomedical sentence embedding. Simultaneously, we explore five data augmentation schemes to assist contrastive learning, and We propose a simple but effective data augmentation, MixCSE-Instance weighting, which to help the model better capture language knowledge and semantic information, yield superior outcomes. In summary, the main contributions of our paper are the following:

- Our approach is the first attempt to improve contrastive learning of unsupervised sentence representations using multiple data augmentation strategies in multiple tasks of biomedical natural language processing.
- We explore various effective data augmentation strategies to generate views for contrastive learning and analyze their effects on unsupervised sentence representation. Specifically, we propose a simple but effective MixCSE-Instance weighting data augmentation methods.
- We conduct extensive experiments on the semantic text similarity (BIOSSES) (Soğancıoğlu et al., 2017) task and the Chinese Biomedical Language Understanding Evaluation benchmark (Chinese-BLUE) (Zhang et al., 2020). Experimental results show that our approach achieves good results compared to the state-of-the-art unsup-SimCSE, respectively.

2 Related Work

2.1 Sentence Representation Learning

Supervised Approaches. Several works are a well studied area with dozens of proposed methods. Previous work (Conneau et al., 2017) finds the supervised Natural Language Inference (NLI) task is useful to train good sentence representation. Stanford NLI (SNLI) (Bowman et al., 2015) and Multi-Genre NLI (MNLI) (Williams et al., 2018) train a Siamese network with max-pooling over the output. SBERT (Reimers and Gurevych, 2019) proposes a siamese architecture with a shared BERT encoder and is also trained on SNLI and MNLI datasets.

Unsupervised Approaches. RAE (Socher et al., 2011) proposes to learn sentence representations based on the internal structure of each sentence.Skip (Kiros et al., 2015) predicts the surrounding sentences for a given sentence based on the distributional assumptions. Sent2Vec (Pagliardini et al., 2018) allows sentence embeddings to be composed using word vectors and n-gram embeddings. BERT-flow (Li et al., 2020) improves the performance of semantic textual similarity (STS) tasks by converting anisotropic sentence embedding distributions into smooth and anisotropic Gaussian distributions. BERT-whitening (Su et al., 2021) uses traditional whitening methods to obtain a smooth distribution of sentence embeddings, and reduce the dimensionality of sentence embeddings.

2.2 Contrastive Learning

Contrastive learning originates in the fields of computer vision (Hadsell et al., 2006; He et al., 2020) and has been widely applied in NLP tasks. For example, the SimCSE (Gao et al., 2021). This core idea aims to learn effective representation by pulling semantically close neighbors (i.e., positive examples) together and pushing apart non-neighbors (i.e., negative examples). One critical question in contrastive learning is how to construct positive example pairs. CERT (Fang et al., 2020) employs back translation for data augmentation. ConSERT (Yan et al., 2021) uses token shuffling to augment positive examples. However, they use to pre-train the language model, which trains in general domain corpus and lack of biomedical domain knowledge. Data augmentation methods reducing sentence composition cause semantic changes. It reduces the effects of contrastive learning.

3 Approach

3.1 General Framework

We use the SimCSE as our general framework. Specifically, given a set of paired sentences $\left\{x_i, x_i^+\right\}_{i=1}^m$, and using $x_i^+ = x_i$. The key element is therefore to construct positive pairs by applying different dropout masks z_i and z_i^+ feeding the same input x_i to the encoder twice and outputting two separate sentence embeddings: $h_i = f_\theta\left(x_i, z_i\right)$ and $h_i^+ = f_\theta\left(x_i, z_i^+\right)$, using h_i and h_i^+ for each sentence in a mini-batch with batch size *N*. The

training objective for contrastive learning is:

$$\ell_i = -\log \frac{e^{\text{sim}(h_i,h_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(h_i,h_j^+)/\tau}} \quad (1)$$

where τ is the temperature hyperparameter and $\text{sim}\,(h_i, h_i')$ is the similarity metric, which is typically the cosine similarity function as follows:

$$\text{sim}\left(h_i, h_i^+\right) = \frac{h_i^\top h_i^+}{\|h_i\| \cdot \left\|h_i^+\right\|} \quad (2)$$

3.2 MixCSE-Instance weighting

We observe that SimCSE simply pushes apart these sampled negatives, which is likely to hurt the semantics of the sentence representations. Therefore, we propose a MixCSE-Instance weighting method to alleviate the problem. This method continuously injects artificial hard negative features into the training process so as to maintain a strong gradient signal throughout training. Simultaneously, we utilize a complementary model to produce the weights for each negative and we use the weights to punish the false negatives.

Given a sentence feature h_i, we construct a negative feature $\widetilde{h'}_{i,j}$ by mixing the positive feature h'_i and a random negative feature h'_j:

$$\widetilde{h'}_{i,j} = \frac{\lambda h'_i + (1-\lambda)h'_j}{\|\lambda h'_i + (1-\lambda)h'_j\|^2} \quad (3)$$

where λ is an hyperparameter to control the degree of mixing. Then, for a negative representation h'_j from the representation of the original sentence h_i, we utilize the complementary model to produce the weight as:

$$\alpha_{h'_j} = \begin{cases} 0 & \text{sim}_{\text{C}}(\text{h}_\text{i}, \text{h}'_\text{j}) \geq \phi \\ 1 & \text{sim}_{\text{C}}(\text{h}_\text{i}, \text{h}'_\text{j}) < \phi \end{cases} \quad (4)$$

where ϕ is a hyper-parameter of the instance weighting threshold, and $\text{sim}_{\text{C}}(\text{h}_\text{i}, \text{h}'_\text{j})$ is the cosine similarity score evaluated by the complementary model. In this way, the negative that has a higher semantic similarity with the representation of the original sentence will be regarded as a false negative and will be punished by assigning the weight 0. Based on the weights, we optimize the sentence representations with a debiased cross-entropy contrastive learning loss function as

$$L = -\log \frac{e^{\text{sim}(h_i,h'_i)/\tau}}{C + \sum_{h'_j \in h'_j} \alpha_{h'_j} e^{\text{sim}\left(h_i, SG(\widetilde{h'}_{i,j})\right)/\tau}} \quad (5)$$

Corpus	Words/Dialogs	Domain
PubMed abstract	$>$ 4,000M	Biomedical
CMCQA	1294753	45 Departments
MedDialog-CN	3407494	29 Departments
Baikemy-Medicine	90785	Medicine

Table 1: Statistics of training corpus.

Method	Text
None	During the 1970s, initial clinical experience with bioprostheses determined their worldwide use.
RC	During the 1970s, initial clinical experience[ ] [ ] [ ] [ ] worldwise use.
WD	During the 1970s, [ ] clinical with bioprostheses determined their worldwide [ ].
RS	During with 1970s, initial bioprostheses with experience the clinical determined their worldwide use.
SI	During the 1970s, initial the clinical experience at with bioprostheses determined their worldwide use in.
PI	During : the 1970s, initial clinical ! experience with ? bioprostheses determined their worldwide use.

Table 2: Sentences generated using different data augmentation methods. RC: random crop. WD: words deletion. RS: random swap. SI: stopwords insertion. PI: punctuations insertion.

where τ is a temperature hyper-parameter. SG(·) denotes a 'stop gradient' operator (Paszke et al., 2019) which ensures that back-propagation does not go through the mixed negative $\widetilde{h'}_{i,j}$

4 Experimental Settings

4.1 Training Data

We use three Chinese datasets: CMCQA (Xia et al., 2022), MedDialog-CN (Zeng et al., 2020), and Baikemy-Medicine[1] and a English dataset: PubMed abstracts (Fiorini et al., 2018), which we randomly select 1 million sentences. The details of the training corpus are shown in Table 1.

4.2 Data Augmentation Strategies

We explore five different data augmentation strategies to construct positive examples for contrastive learning, whose examples are shown in Table 2:

[1]See https://https://www.baikemy.com.

Dataset	Train	Dev	Test	Task	Metrics	Domain
cMedIC	1683	123	84	Intent Classification	F1	Medical
cMedQQ	16071	1793	1935	Paraphrase Identification	F1	Medical
cMedQA	49719	5475	6149	Question Answering	F1	Medical
cMedQNLI	80950	9065	9969	Question Answering	F1	Medical
BIOSSES	64	16	20	Sentence Similarity	Spearman corr.	Biomedical

Table 3: Statistics of BIOSSES and ChineseBLUE.

- **Random Crop (RC)** aims to introduce variability and perturbations into the training data, encouraging the model to learn more robust and generalized representations by processing partially masked or altered inputs. Random crop randomly selects and removes sections of text from sentences or paragraphs and replaces them with a placeholder or mask token.

- **Words Deletion (WD)** aims to simulate scenarios where words are missing or omitted, thereby encouraging the model to learn more robust representations and improve its ability to handle incomplete or altered text inputs. Words within sentences or text are randomly removed or deleted.

- **Random Swap (RS)** aims to help models understand various styles and contexts by different word orderings and sentence structures. In this way, two words within a sentence are randomly selected and swapped with each other. The sentence structure is altered slightly, introducing variability into the dataset.

- **Stopwords Insertion (SI)** aims to diversify the language patterns and structures in the data for training NLP models, enhancing their robustness and ability to handle different language styles and contexts. In this way, stopwords are strategically added to the text. For example, *and*, *the*, *of*, *etc.* are inserted into sentences to create variations in the dataset.

- **Punctuations Insertion (PI)** is the strategic addition of punctuation marks (e.g., *!*, *?*, *etc.*) within the text to generate variations and expand the training dataset. This method aims to enhance the robustness of NLP models by introducing diverse sentence structures and patterns, thereby improving their ability to comprehend and process different writing styles and contexts.

4.3 Evaluation Task

To verify the effectiveness of our proposed approach, we use one Chinese dataset ChineseBLUE (Zhang et al., 2020), where the ChineseBLUE experiments four subtasks (cMedIC, cMedQQ, cMedQA, and cMedQNLI) and one English dataset: BIOSSES (Soğancıoğlu et al., 2017). The detailed statistics are shown in Table 3:

- **cMedIC (Clinical Medical Information Corpus)** contains textual clinical medical information covering diseases, symptoms, diagnoses, treatment plans, and other medical content. It's designed to train models to understand and process clinical medical text.

- **cMedQQ (Clinical Medical Question and Question)** comprises medical-related questions and their corresponding answers, involving medical knowledge, diagnoses, treatments, and more. Through this dataset, models can learn to answer medical questions and interpret medical information.

- **cMedQA (Clinical Medical Question Answering)** dataset is also focused on medical domain question answering but might encompass a wider range of medical topics and question types.

- **cMedQNLI (Clinical Medical Question Natural Language Inference)** primarily focuses on natural language inference, containing pairs of medical domain questions and sentences that require the model to infer logical relationships between these question-sentence pairs, such as entailment or contradiction.

- **BIOSSES** is a semantic text similarity task. This provides a collection of 100 similar sentence pairs manually annotated in the biomedical domain. We use the training-testing split of BLURB (Gu et al., 2021), where 64 pairs are used for training, 16 pairs for validation, and the remaining 20 pairs for testing.

Model	cMedIC	cMedQQ	cMedQA	cMedQNLI	Avg.
Chinese-bert-wwm-ext	$87.26_{\pm 0.80}$	$77.44_{\pm 0.99}$	$84.44_{\pm 0.35}$	$88.52_{\pm 0.32}$	$84.42_{\pm 0.28}$
SimCSE	$92.25_{\pm 1.34}$	$80.22_{\pm 1.11}$	$84.87_{\pm 0.42}$	$91.66_{\pm 0.08}$	$87.25_{\pm 0.41}$
+Words deletion 10%	$89.86_{\pm 0.63}$	$81.08_{\pm 0.43}$	$84.84_{\pm 0.53}$	$91.88_{\pm 0.20}$	$86.92_{\pm 0.28}$
+Words deletion 20%	$90.79_{\pm 0.90}$	$81.39_{\pm 0.57}$	$84.84_{\pm 0.38}$	$91.81_{\pm 0.37}$	$87.21_{\pm 0.28}$
+Words deletion 30%	$89.54_{\pm 1.28}$	$81.18_{\pm 0.42}$	$85.12_{\pm 0.25}$	$91.36_{\pm 0.43}$	$86.80_{\pm 0.32}$
+Random crop 10%	$90.42_{\pm 0.80}$	$81.52_{\pm 0.91}$	$85.01_{\pm 0.33}$	$91.50_{\pm 0.34}$	$87.11_{\pm 0.31}$
+Random crop 20%	$91.02_{\pm 0.60}$	$81.31_{\pm 0.52}$	$84.68_{\pm 0.79}$	$91.68_{\pm 0.22}$	$87.17_{\pm 0.33}$
+Random crop 30%	$89.86_{\pm 0.51}$	$80.70_{\pm 0.48}$	$85.14_{\pm 0.26}$	$91.73_{\pm 0.31}$	$86.86_{\pm 0.26}$
+Stopwords insertion	$89.01_{\pm 1.52}$	$81.69_{\pm 0.29}$	$84.21_{\pm 0.34}$	$91.62_{\pm 0.09}$	$86.63_{\pm 0.57}$
+Random swap	$88.80_{\pm 0.57}$	$81.08_{\pm 0.45}$	$84.28_{\pm 0.59}$	$91.44_{\pm 0.39}$	$86.40_{\pm 0.39}$
+Punctuations insertion	$91.57_{\pm 1.16}$	$\mathbf{81.73_{\pm 0.60}}$	$84.59_{\pm 0.52}$	$92.20_{\pm 0.53}$	$87.52_{\pm 0.19}$
+MixCSE-Instance weighting	$\mathbf{93.26_{\pm 0.58}}$	$80.62_{\pm 0.94}$	$\mathbf{85.21_{\pm 0.24}}$	$\mathbf{93.08_{\pm 0.08}}$	$\mathbf{88.04_{\pm 0.13}}$

Table 4: Results on cMedIC, cMedQQ, cMedQA and cMedQNLI test sets. Metric, weighted-averaged F1 for cMedIC and cMedQA and F1 for cMedQQ and cMedQNLI.

Data augmentation	BIOSSES
SimCSE	$81.02_{\pm 3.02}$
+Words deletion 10%	$80.39_{\pm 3.74}$
+Words deletion 20%	$79.13_{\pm 2.63}$
+Words deletion 30%	$74.62_{\pm 3.40}$
+Random crop 10%	$\mathbf{85.36_{\pm 3.27}}$
+Random crop 20%	$82.36_{\pm 2.48}$
+Random crop 30%	$74.13_{\pm 2.41}$
+Punctuations insertion	$84.72_{\pm 1.36}$
+Stopwords insertion	$84.93_{\pm 0.97}$
+Random swap	$85.29_{\pm 3.05}$
+MixCSE-Instance weighting	$82.43_{\pm 2.20}$

Table 5: Comparison results of different data augmentation methods on the BIOSSES dataset (Spearman's correlation). All results use 5 random seeds to train the model, and report the mean and standard deviation.

4.4 Evaluation Protocols

When evaluating the trained model, We use two methods to evaluate the model:

- **Spearman correlation** We use Spearman correlation to measure the correlation between the ranks of predicted similarities and the ground truth. For a set of size *n*, the n raw scores X_i, Y_i are converted to the corresponding ranks $\mathrm{rg}_{X_i}, \mathrm{rg}_{Y_i}$, then the Spearman correlation is defined as follows:

$$r_s = \frac{\mathrm{cov}\left(\mathrm{rg}_X, \mathrm{rg}_Y\right)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}} \tag{6}$$

  where $\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)$ is the covariance of the rank variable, and σ_{rg_X} and σ_{rg_Y} are the standard deviation of the rank variable.

- **SentEval** We use SentEval (Conneau and Kiela, 2018) to evaluate the quality of sentence embeddings. This uses its generated sentence representations to train a classifier on the downstream task and verifies the quality of the sentence representations by means of F1 scores.

4.5 Implementation Details

We start with the pre-training checkpoints of BERT-base or BERT-wwm, and add the MLP layer at the top of the [CLS] representation to obtain sentence embeddings. The MLP layer is discarded during testing and only the [CLS] output is used, like in the unsup-SimCSE setup. During training, we use the Adam optimizer (Kingma and Ba, 2014). Training our model at temperature $\tau = 0.05$ for one epoch. For Bert-base-uncased and Chinese-bert-wwm-ext, the batch size is 128, and the learning rate is 3e-5 and 1e-5. In using the negative sample optimization strategy, we set the instance weight threshold to 0.9 for both models. For all experiments, we take five different random seeds and average the results with standard deviation. Our code is implemented in Python 3.7, using Pytorch 1.13, and the experiments are conducted on a single 24G NVIDIA 3090 GPU.

5 Experiments Results

5.1 Main Results

Table 4 presents the evaluation results on the ChineseBLUE dataset. We observe that the generic model has a much lower performance compared to a biomedical corpus for contrastive learning. On

Methods	cMedIC	cMedQQ	cMedQA	cMedQNLI	Avg.
SimCSE	$92.25_{\pm 1.34}$	$80.22_{\pm 1.11}$	$84.87_{\pm 0.42}$	$91.66_{\pm 0.08}$	$87.25_{\pm 0.41}$
MixCSE-Instance weighting	$\mathbf{93.26_{\pm 0.58}}$	$80.62_{\pm 0.94}$	$\mathbf{85.21_{\pm 0.24}}$	$\mathbf{93.08_{\pm 0.08}}$	$\mathbf{88.04_{\pm 0.13}}$
–MixCSE	$92.85_{\pm 0.87}$	$80.29_{\pm 0.33}$	$84.95_{\pm 0.40}$	$92.26_{\pm 0.10}$	$87.59_{\pm 0.16}$
–Instance weighting	$88.52_{\pm 0.50}$	$80.12_{\pm 0.15}$	$84.29_{\pm 0.27}$	$91.26_{\pm 0.33}$	$86.05_{\pm 0.10}$
Punctuations insertion	$91.57_{\pm 1.16}$	$\mathbf{81.73_{\pm 0.60}}$	$84.59_{\pm 0.52}$	$92.20_{\pm 0.53}$	$87.52_{\pm 0.19}$
P+M	$91.05_{\pm 0.57}$	$81.18_{\pm 0.29}$	$84.25_{\pm 0.38}$	$92.51_{\pm 0.36}$	$87.25_{\pm 0.10}$

Table 6: Results of the Ablation Study. P+M: Punctuations insertion and MixCSE-Instance weighting.

the cMedQQ data, the model effects of applying different data augmentation methods are all well improved, when using the Punctuations insertion method gains the most for unsup-SimCSE (Gao et al., 2021), the F1 score improves from 80.22% to 81.73%. MixCSE-Instance weighting method improves on all four test sets by an average of 0.79%, based on the weighted-averaged F1 scores, which improve by 1.01% and 0.34% over unsup-SimCSE on the cMedIC and cMedQA test sets, while on the cMedQQ and cMedQNLI test sets the F1 scores improve by 0.40% and 1.42% over unsup-SimCSE, respectively. It has a relatively low standard deviation, which indicates the validity and stability of our method.

Table 5 presents the evaluation results on the BIOSSES dataset. We find that the results of the models improved greatly after changing to different data augmentation methods. Among the five methods we propose, when the Random crop 10% method is used as the data augmentation strategy, it improves the most for the current state-of-the-art SimCSE results, and the Spearman's correlation coefficient improved from 81.02% to 85.36%. However, the model performance gradually decreases as the cropping ratio increases. By observing the effects of each data augmentation method, we find that for both the Words deletion and Random crop methods increasing the proportion of deletion or cropping resulted in poorer contrastive learning, which may be attributed to the fact that too much reduction of sentence components can cause the original semantics to be drastically altered.

6 Ablation Study

We investigate the impact of MixCSE and Instance on the SimCSE in the Chinese datasets. The ablation results are shown in Table 6, where removing each component leads to performance degradation. This suggests that both MixCSE mixing positive and negative samples as a hard-negative example method and instance weighting are important in improving the contrastive learning results. In addition, eliminating the instance weighting method leads to larger performance degradation. The reason may be that false negatives have a greater impact on sentence representation learning.

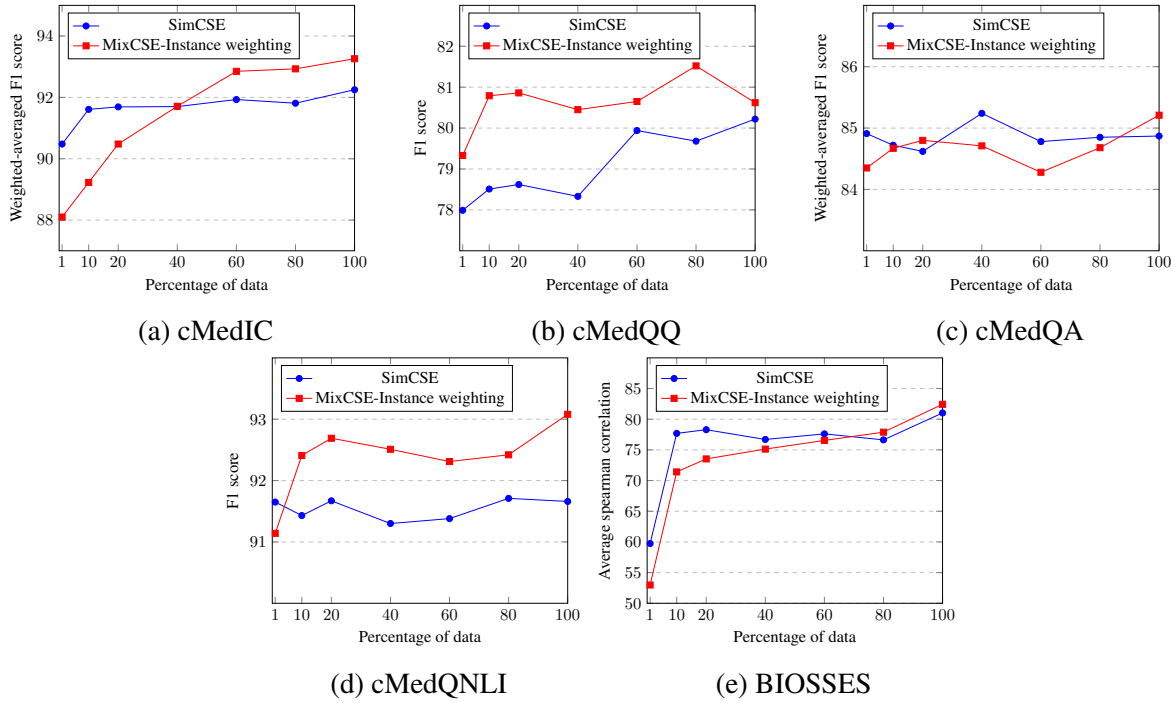
We also study both the positive and negative samples separately for the data augmentation strategies. In order to further compare the effects of different data augmentation methods on improving contrastive learning, we select the two methods with the best average performance in the Chinese downstream task for evaluation: Punctuations insertion and MixCSE-Instance weighting, where both positive sample augmentation and negative sample optimization are applied simultaneously for training. As can be seen in Table 6, this method is only slightly better than the Punctuations insertion method in the cMedQNLI task and the other results are not as good as separating the positive and negative sample augmentation independently.

7 Analysis

7.1 Effect of Training Set Sizes

To validate the reliability and the robustness of the MixCSE-Instance weighting methods under the data scarcity scenarios, we conduct the few-shot experiments. We limit the number of unlabeled texts to 1%, 10%, 20%, 40%, 60%, 80%, and compare their performance with the full dataset.

Figure 1 presents the results. We optimize the model for the same number of training steps as for the full set of settings. In all five tasks, our data augmentation approach achieves good gains compared to baseline SimCSE. In particular, it shows good performance on the two datasets BIOSSES and cMedQQ, which judge semantic similarity. The results reveal the robustness and effectiveness of our approach under the data scarcity scenarios, which are common in reality. With only a small amount of unlabeled texts drawn from the target data distribu-



(a) cMedIC (b) cMedQQ (c) cMedQA

(d) cMedQNLI (e) BIOSSES

Figure 1: Performance of different training data scales. All scores are the average of 5 experiments.

SimCSE	**MixCSE-Instance weighting**
Query: 什么是肾实质性高血压?	
(What is renal parenchymal hypertension?)	
#1 | 是否是肾上腺引起的高血压?	| 肾性高血压是怎么产生的?
(Is the adrenal gland causing high blood pressure?)	(How does renal hypertension occur?)
#2 | 什么是糖尿病肾病?	| 肾性高血压是什么意思?
(What is diabetic nephropathy?)	(What does renal hypertension mean?)
Query: HIV抗体检测方法有哪些?	
(What are the HIV antibody testing methods?)	
#1 | 乙型肝炎抗原检查注意事项。	| 艾滋病检查方法有哪些?
(Hepatitis B antigen test precautions.)	(What are the methods of HIV testing?)
#2 | 艾滋病的检查及费用。	| 有关hiv抗体检测的咨询。
(HIV testing and costs.)	(Consultation about hiv antibody testing.)

Table 7: Retrieved examples from cMedQQ test set.

tion, our approach can also tune the representation space and benefit the downstream tasks.

7.2 Sentence Retrieval

As shown in Table 7, we sampled predictions from the model to see the effect of our approach on downstream tasks. Given an input sentence, the nearest neighbor will be retrieved based on cosine similarity. Sentences retrieved using the MixCSE-Instance weighting data augmentation method have a higher quality compared to those retrieved by SimCSE.

8 Conclusion and Future Work

In this paper, we propose MixCSE-Instance weighting, a data augmentation strategy for contrastive learning framework in biomedical natural language processing domain tasks. Furthermore, few-shot experiments suggest that our method is robust in data scarcity scenarios. We also compare multiple combinations of data augmentation strategies and provide fine-grained analysis for interpreting how our approach works. Experiments show that our approach improves the performance of down-

stream tasks. In the future, we will explore how to improve the generalization ability of data augmentation techniques in biomedical contrastive learning tasks and validate their effectiveness on more contrastive learning methods.

Acknowledgements

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Nicolas Fiorini, Robert Leaman, David J Lipman, and Zhiyong Lu. 2018. How user intelligence is improving pubmed. *Nature biotechnology*, 36(10):937–945.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

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*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

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 (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Richard Socher, Eric Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Advances in neural information processing systems*, 24.

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Fei Xia, Bin Li, Yixuan Weng, Shizhu He, Kang Liu, Bin Sun, Shutao Li, and Jun Zhao. 2022. Lingyi: Medical conversational question answering system based on multi-modal knowledge graphs. *arXiv preprint arXiv:2204.09220*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020. Conceptualized representation learning for chinese biomedical text mining. *arXiv preprint arXiv:2008.10813*.