# Generate, Discriminate and Contrast: A Semi-Supervised Sentence Representation Learning Framework

**Yiming Chen**[†] **Yan Zhang**[†] **Bin Wang**[†] **Zuozhu Liu**[‡⋆*] **Haizhou Li**[♮,†,§]
[†]National University of Singapore [‡]Zhejiang University [⋆]Angelalign Inc., China
[♮]The Chinese University of Hong Kong, Shenzhen, China [§]Kriston AI Lab, China
yiming.chen@u.nus.edu, zuozhuliu@intl.zju.edu.cn
{haizhou.li,eleyanz,bin.wang}@nus.edu.sg

## Abstract

Most sentence embedding techniques heavily rely on expensive human-annotated sentence pairs as the supervised signals. Despite the use of large-scale unlabeled data, the performance of unsupervised methods typically lags far behind that of the supervised counterparts in most downstream tasks. In this work, we propose a semi-supervised sentence embedding framework, GenSE, that effectively leverages large-scale unlabeled data. Our method include three parts: 1) Generate: A generator/discriminator model is jointly trained to synthesize sentence pairs from open-domain unlabeled corpus; 2) Discriminate: Noisy sentence pairs are filtered out by the discriminator to acquire high-quality positive and negative sentence pairs; 3) Contrast: A prompt-based contrastive approach is presented for sentence representation learning with both annotated and synthesized data. Comprehensive experiments show that GenSE achieves an average correlation score of 85.19 on the STS datasets and consistent performance improvement on four domain adaptation tasks, significantly surpassing the state-of-the-art methods and convincingly corroborating its effectiveness and generalization ability.[1]

## 1 Introduction

Sentence representation learning has recently attracted outspread research attention. It learns vector representations for sentences, which can be subsequently utilized on a wide range of downstream tasks, including information retrieval (Thakur et al., 2021b; Wang et al., 2022; Misra et al., 2016), language understanding and evaluation (Cer et al., 2018; Conneau and Kiela, 2018; Perone et al., 2018; Zhang et al., 2021a). For sentence representation learning, the contrastive learning-based approaches, including supervised and unsupervised ones, have demonstrated to be the most efficient and effective (Zhang et al., 2020; Carlsson et al., 2021; Giorgi et al., 2021; Gao et al., 2021). In contrastive learning, the quality of positive and negative pairs has a large impact on the overall performance (Chen et al., 2020; Gao et al., 2021). In particular, previous supervised contrastive methods usually construct sentence pairs using human-annotated natural language inference (NLI) data (Bowman et al., 2015; Williams et al., 2018), and outperform unsupervised approaches by a large margin (Gao et al., 2021). However, the state-of-the-art supervised methods usually rely on multi-source labeled data to generalize to various downstream tasks (Reimers and Gurevych, 2019), while such large-scale annotated data across domains are not always available.

Recent works also attempt to leverage the resources of unlabeled sentences for better sentence representation learning. A straightforward choice is to adopt unsupervised or self-supervised paradigms, such as BERT-flow (Li et al., 2020), SimCSE (Gao et al., 2021). However, the performance of these methods is still far behind the supervised counterparts. Another stream of work employs retrieval strategies to obtain close sentences as potential entailment pairs, and then trains a discriminator to re-label the retrieved sentence pairs (Thakur et al., 2021a). They attain satisfactory results but with two main limitations: 1) The retrieved sentences may not be the ideal entailment to the input sentences due to the limited numbers of sentences in the banking corpus, leading to low-quality synthetic sentence pairs; 2) The retrieval-based methods can only obtain entailment pairs but not contradiction ones, which can be an important source of hard negative samples.

In this paper, we propose a novel semi-supervised sentence representation learning framework leveraging both annotated and unlabeled corpus to address the aforementioned issues. Our

* Corresponding author.
[1]Code, Synthetic data and Models available at https://github.com/MatthewCYM/GenSE



*Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8150–8161
December 7-11, 2022 

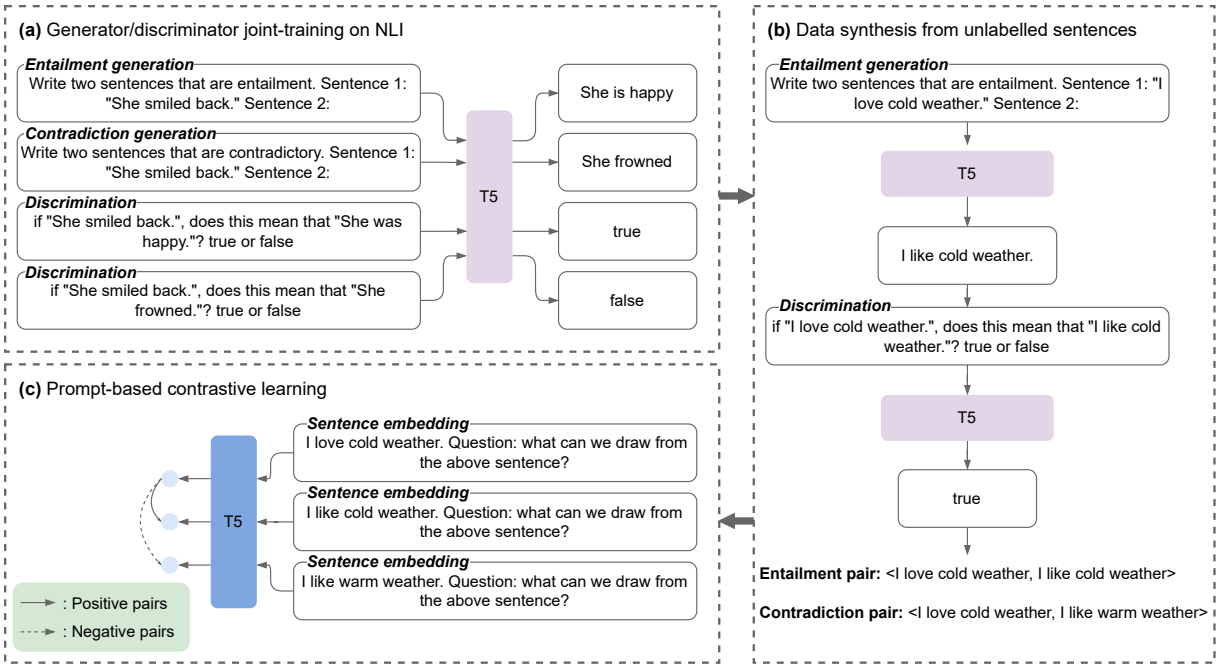


Figure 1: GenSE framework includes three steps: jointly training of generator/discriminator; data synthesis from unlabelled data; prompted-based contrastive learning. All models are initialized from same pre-trained weights. The model with the same colour shares the same weights. The upper-left of each sentence block refers to the prompt name listed in Table 1.

method, GenSE, is built upon pre-trained text-to-text models. It integrates three tasks, i.e., generation, discrimination, and contrastive learning, into a single framework. Specifically, we first train a unified generator/discriminator model from NLI data, which is responsible for sentence pair generation and noisy pair discrimination. Afterwards, the open-domain unlabeled sentences are taken as inputs to the generator to synthesize sentence pairs, which are further discriminated to obtain high-quality positive and negative pairs. Finally, a prompt-based encoder-decoder model is trained in a contrastive manner to learn sentence embeddings from both human-annotated and our synthesized sentence pairs.

We evaluate GenSE on the standard semantic textual similarity (STS) benchmark and four downstream domain adaptation tasks. On the STS benchmark, GenSE achieves an averaged Spearman's correlation of 84.78, which is further boosted to 85.19 by integrating additional QA data, significantly outperforming the state-of-the-art baselines. On domain adaptation tasks, we consider two different settings, i.e. direct transfer and domain adaptation. Our GenSE achieves an average 1.9% improvement over baselines for direct transfer. Through further domain adaptation with sentence pairs synthesized from unlabeled in-domain data, GenSE achieves an extra 1.7% improvement, confirming its adaptability for various downstream tasks. Comprehensive ablation studies on different components of GenSE, extensive comparisons between data synthesis strategies, and in-depth analysis on the uniformity/alignment and quality of the synthetic data, convincingly validate the effectiveness and generalization ability of our GenSE framework.

## 2 Methodology

The GenSE comprises two neural models, including a unified generator/discriminator model for sentence pair generation and quality check, and a sentence embedding model optimized with a contrastive objective. The learning schema of these two models consists of three consecutive steps, as illustrated in Figure 1.

Firstly, we perform joint-training of the generator/discriminator model based on NLI dataset. Afterwards, we take the large-scale unlabeled sentences as inputs to generate entailment and contradiction pairs with the trained generator, which are further filtered by the discriminator to keep high-quality positive and negative pairs. Finally, we train the contrastive learning model based on the synthesized and human-annotated sentence pairs.

| Tasks | Input templates | Output templates |
|---|---|---|
| $T_e$: Entailment generation | Write two sentences that are entailment. Sentence 1: "**[X1]**" Sentence 2: | **[X2]** |
| $T_c$: Contradiction generation | Write two sentences that are contradictory. Sentence 1: "**[X1]**" Sentence 2: | **[X2]** |
| $T_d$: Discrimination | if "**[X1]**", does this mean that "**[X2]**"? true or false | true/false |
| $T_s$: Sentence embedding | **[X]** Question: what can we draw from the above sentence? | sentence embedding vector |

Table 1: Prompt used in GenSE. [X] refers to the placeholder for input/output sentences.

We formulate all tasks into a text-to-text format with a prompt as the task signal. Therefore, we can build our models from the initialization of the same pre-trained text-to-text model, which leads to high modeling simplicity. Table 1 lists all the input, output templates used in GenSE, including entailment generation template $T_e$, contradiction generation template $T_c$, discrimination template $T_d$, and sentence embedding template $T_s$. Below we'd like to elaborate more.

### 2.1 Generate and Discriminate

The training process of the generator/discriminator model is illustrated in Figure 1.(a). The model is trained to perform two tasks, i.e., generation and discrimination, simultaneously with NLI data. For the generation task, we first obtain two training instances for each NLI triplet $\{x_{ori}, x_{entail}, x_{contra}\}$. In particular, as for entailment, we place $x_{ori}$ into the pre-defined entailment input templates $T_e$ to obtain model input $T_e(x_{ori})$, and set the output label as the corresponding entailment hypothesis $x_{entail}$. Similarly, we use $T_c$ in the contradiction generation to obtain the training instance: $\{T_c(x_{ori}), x_{contra}\}$. The generator is trained to predict the output labels given the input prompts and sentences, as shown in the top of Figure 1.(a).

For the discrimination task, we apply the prompt template $T_d$ on the concatenated sentence pairs to obtain model inputs, and the output is either true or false, leading to two training instances: $\{T_d(x_{ori}, x_{entail}), \text{true}\}$ and $\{T_d(x_{ori}, x_{contra}), \text{false}\}$. The output is mapped from the annotated labels, i.e., entailment $\rightarrow$ true, and contradiction $\rightarrow$ false.

By adding the prompts, both the generation and discrimination tasks can be transformed into conditional generation tasks, which largely reduce the model complexity. The model can be trained with a standard conditional generation loss to maximize the probability of the output sequence $y_{1,...,M}$ given the input sequence $X$:

$$P(y_1, y_2, ..., y_M|X) = \prod_{m=1}^{M} P(y_m|y_0, ..., y_{m-1}, X), \quad (1)$$

where $y_0$ is the decoder start token.

To better balance the performance between the generation and discrimination tasks, we equally mix the generation and discrimination training instances, and employ a weighted sum of the generation perplexity and discrimination accuracy on the development set for model selection.

The data synthesis and quality check process is shown in Figure 1.(b). Given an unlabelled sentence $u$, we first augment it with the generation prompts to obtain model inputs $T_e(u)$ and $T_c(u)$, which are fed into the trained generator to get entailment/contradiction prediction $u_{entail}$ and $u_{contra}$. The generated triplets $\{T_d(u_{ori}, u_{entail}), \text{true}\}$ and $\{T_d(u_{ori}, u_{contra}), \text{false}\}$ are evaluated by the discriminator for quality check, where the output probabilities of 'true' and 'false' are compared with a predefined threshold $\alpha$ to filter out the low-quality triplets. Finally, we can get a clean synthetic corpus with both positive and negative pairs for sentence representation learning.

### 2.2 Prompt-based Contrastive Learning

Contrastive learning has been widely used in sentence representation learning, achieving state-of-the-art performance on the STS benchmark (Gao et al., 2021; Ni et al., 2021). Recently, PromptBERT (Jiang et al., 2022) utilizes prompts to obtain sentence embeddings of much higher quality from encoder-based models. Inspired by it, we propose an improved prompt-based contrastive sentence representation model for text-to-text models.

Our prompt-based contrastive learning is shown in Figure 1.(c). we first augment the input sentence [X] with prompt $T_s$ shown in Table 1, and feed the "[PAD]" token into the decoder. We take the first decoder output as the sentence embedding as in (Ni et al., 2021). During training, we regard the

entailment pairs as positive pairs, and the contradictory sentences as hard negative pairs. We also use in-batch negative sampling to include more negative samples during training. For a batch with $N$ triplets, the prompt-based contrastive learning loss function is defined as:

$$\mathcal{L} = \frac{e^{\text{sim}(h_i,h_i^+)/\tau}}{\sum_{j=1}^{N}(e^{\text{sim}(h_i,h_i^+)/\tau} + e^{\text{sim}(h_i,h_i^-)/\tau})}, \quad (2)$$

where $i, j$ is the sentence index, $\tau$ is the temperature hyper-parameter, $\text{sim}(\cdot)$ is the cosine similarity function, and $\{h, h^+, h^-\}$ are representations of the premise, and corresponding entailment and contradiction hypotheses.

### 2.3 Training Settings

We consider three different settings, including universal sentence embedding, domain adaptation and QA training.

**Universal sentence embedding:** We consider the general case where we have large-scale open domain unlabeled sentences and restricted amount of human-labeled NLI data. In this case, the GenSE is trained with a two-stage schema. In particular, we first train the sentence embedding model on the synthetic open-domain triplets from the generator/discriminator model, which is subsequently finetuned on the labeled NLI datasets.

**Domain adaptation:** The GenSE can also be used for domain adaptation. In this case, we assume the GenSE would have access to extra in-domain unlabeled data from the downstream tasks, which is regarded as a standard setting in unsupervised sentence embedding (Wang et al., 2021). The open-domain corpus are available as well. We follow the same procedure in Section 2.1 to obtain in-domain sentence triplets by taking the in-domain sentences as the input to the trained generator/discriminator model. Afterwards, the prompt-based sentence embedding model is trained in three stages: 1) Pretraining on open-domain synthetic pairs; 2) Adapting the sentence embedding to a specific domain using in-domain synthetic pairs; 3) Finetuning on labeled NLI data.

**QA training:** It is also not uncommon to use additional QA data to improve sentence representation learning. Here, we also develop a setting to demonstrate whether GenSE can further boost the performance along with labeled QA data. Since hard negatives are not available in QA data, we remove them from the loss function:

$$\mathcal{L}' = \frac{e^{\text{sim}(h_i,h_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(h_i,h_i^+)/\tau}}. \quad (3)$$

We consider two ways to train the prompt-based contrastive learning model to demonstrate the effectiveness of GenSE: 1) GenSE-QA: the model is first trained on QA data and then finetuned on NLI data. 2) GenSE+: the model is first pretrained on the open-domain synthetic data generated by the generator/discriminator model, and then trained on the QA data, and finally finetuned on the NLI data.

## 3 Experiments

### 3.1 Experiment Setup

We evaluate GenSE on seven popular STS tasks for universal sentence embedding, and on four datasets from various domains, including AskUbuntu (Lei et al., 2016), CQADupStack (Hoogeveen et al., 2015), Twitter (Xu et al., 2015; Lan et al., 2017), and BIOSSES (Soğancıoğlu et al., 2017). For all the experiments, we assume the only available labeled sentence pairs are NLI (MNLI+SNLI) (Bowman et al., 2015; Williams et al., 2018). For open-domain data synthesis, we sample sentences from C4 news-like and English partitions (Raffel et al., 2020) [2], and obtain around 61M synthetic triplets. In the domain adaptation setting, we follow TSDAE (Wang et al., 2021) to use unlabelled training set as in-domain sentences for AskUbuntu, CQADupStack, and Twitter. Since no training set is available for BIOSSES, we use PubMed subset in the Pile (Gao et al., 2020) as in-domain sentences, and remove the sentences existing in the test set.

As for the QA training, we utilize public available QA data for training, since CommQA used in (Ni et al., 2021) is not released. We choose datasets that are sampled from web sources, and have a sentence as both input and output. We also remove datasets that are closely related to downstream tasks, e.g., Stack Exchange, as well as the ones that are manually annotated, e.g., MS MARCO (Bajaj et al., 2016), for fair comparison. Finally, we obtain 4M QA pairs, including ELI5 (Fan et al., 2019), GOOAQ (Khashabi et al., 2021), and Yahoo (Zhang et al., 2015).

We build GenSE upon the widely-used T5 Encoder-Decoder models (Raffel et al., 2020). For

[2] huggingface.co/datasets/c4

| Model | # Params | STS12 | STS13 | STS14 | STS15 | STS16 | STSb | SICK-R | Avg |
|---|---|---|---|---|---|---|---|---|---|
| *Large-scale sentence embedding models* | | | | | | | | | |
| SimCSE-RoBERTa-Large | 354M | 77.46 | 87.27 | 82.36 | 86.66 | 83.93 | 86.70 | 81.95 | 83.76 |
| ST5-Enc-Large | 335M | 76.52 | 85.75 | 81.01 | 87.13 | 83.26 | 85.45 | 79.85 | 82.71 |
| ST5-EncDec-Large | 335M | 79.15 | 87.42 | 83.61 | 87.64 | 83.92 | 86.35 | 80.64 | 84.11 |
| ST5-Enc-Large-CommQA | 770M | 79.10 | 87.32 | 83.17 | 88.27 | 84.36 | 86.73 | 79.84 | 84.11 |
| ST5-EncDec-3B | 3B | 79.24 | 87.80 | 83.95 | 87.75 | 84.60 | 86.62 | 80.91 | 84.41 |
| ST5-Enc-3B-CommQA | 1.24B | 79.02 | 88.80 | 84.33 | 88.89 | 85.31 | 86.25 | 79.51 | 84.59 |
| *Base-size sentence embedding models* | | | | | | | | | |
| SBERT-Base | 110M | 70.97 | 76.53 | 73.19 | 79.09 | 74.30 | 77.03 | 72.91 | 74.89 |
| SimCSE-RoBERTa-Base | 110M | 76.53 | 85.21 | 80.95 | 86.03 | 82.57 | 85.83 | **80.50** | 82.52 |
| ST5-Enc-Base | 110M | 77.37 | 83.65 | 80.41 | 86.04 | 81.70 | 84.49 | 79.79 | 81.92 |
| ST5-EncDec-Base | 220M | 77.90 | 85.62 | 82.24 | 86.81 | 82.13 | 84.98 | 79.97 | 82.81 |
| ST5-Enc-Base-CommQA | 110M | 78.05 | 85.84 | 82.19 | 87.46 | 84.03 | 86.04 | 79.75 | 83.34 |
| GenSE | 220M | 80.72 | 87.43 | 83.96 | 88.63 | 85.19 | 87.65 | 79.87 | 84.78 |
| GenSE-QA | 220M | **80.84** | 87.52 | 83.19 | 87.48 | 84.35 | 86.42 | 79.73 | 84.22 |
| GenSE+ | 220M | 80.65 | **88.18** | **84.69** | **89.03** | **85.82** | **87.88** | 80.10 | **85.19** |

Table 2: Results of sentence embedding on STS tasks. Spearman's correlation is reported. The first block shows previous state-of-the-art large-scale embedding models, and the second block shows results from base-size models.

generator/discriminator training, we set the learning rate to 5e-5 and batch size to 256. The generator/discriminator is trained for 10 epochs with evaluation step set as 500, and $(accuracy - 10 \times ppl)$ as validation metric for model selection. For data synthesis, we use nucleus sampling (Holtzman et al., 2019) with $p = 0.9$, and set the confidence threshold $\alpha = 0.9$. For contrastive sentence representation learning, we set the learning rate to 5e-5 and batch size to 512 on NLI, and learning rate of 5e-5 with batch size of 1024 on the synthetic data. The temperature $\tau$ is set to 0.01 for QA training, and 0.05 for others.

### 3.2 STS Results

We compare GenSE with T5-Base to previous state-of-the-art sentence embedding methods, including encoder-based and encoder-decoder based models. All baselines are trained on the labeled NLI corpus. We also include large-scale models and models trained with additional large-scale semi-structured CommQA data for comparison. The results are reported in Table 2.

Overall, GenSE can greatly outperform previous state-of-the-art models. Specifically, GenSE achieves an average Spearman's correlation of 84.78, significantly outperforming Base-size sentence embedding models, and even surpassing methods with much larger model sizes, e.g., ST5-Enc-3B and ST5-Enc-3B-CommQA. GenSE attains new state-of-the-art base-model results on 6 out of 7 STS datasets, i.e., except the SICK-R tasks, demonstrating that our synthetic data can greatly improve sentence embedding quality with GenSE.

We also report the performance of GenSE+, which achieves even higher performance with an average Spearman's correlation of 85.19 by integrating additional QA data. The results suggest that our synthetic data is complementary to semi-structured data like question-answer pairs. But compared to semi-structured data which requires lots of efforts for data mining and cleaning, our GenSE only need unlabelled sentences that are much easier to collect for various domains, exhibiting better practical applicability.

### 3.3 Transfer Tasks

**Direct transfer:** We first consider the direct transfer scenario where unlabeled in-domain data are not used for sentence embedding training, with results in the top of Table 3. We can see that GenSE shows 1.9 % improvements over strong ST5 baselines, and GenSE-QA works even slightly better. Combining QA and synthetic data, i.e., GenSE+, leads to a substantial improvement of 3.2%, demonstrating the effectiveness of synthetic data and QA pairs.

**Domain adaptation:** We also evaluate the model for domain adaptation, where in-domain synthetic data is also included in contrastive learning. We can observe that the average performance is improve by 1.7% over the best TSDAE-BERT-Base

[3] github.com/princeton-nlp/SimCSE
github.com/UKPLab/sentence-transformers

| Model | Fine-tune Data | AskU. | CQADup. | TwitterP. | | | BIOSSES | Avg. |
|---|---|---|---|---|---|---|---|---|
| | | | | TURL. | PIT | Avg. | | |
| *Direct transfer* | | | | | | | | |
| SimCSE-BERT-Base | NLI | 53.5 | 12.4 | 75.6 | 66.9 | 71.2 | 68.4 | 55.4 |
| SimCSE-RoBERTa-Base | NLI | 54.6 | 11.7 | 74.4 | 68.5 | 71.5 | 67.7 | 55.4 |
| ST5-Enc-Base | NLI | 56.6 | 14.3 | 73.9 | 72.5 | 73.2 | 70.2 | 57.5 |
| ST5-EncDec-Base | NLI | 56.1 | 13.7 | 73.3 | 75.0 | 74.1 | 71.4 | 57.9 |
| GenSE | Open-domain→NLI | 58.2 | 15.3 | 76.3 | 75.9 | 76.1 | 73.1 | 59.8 |
| GenSE-QA | QA→NLI | 57.4 | 15.6 | 75.8 | 75.8 | 75.8 | 75.1 | 59.9 |
| GenSE+ | Open-domain→QA→NLI | **58.4** | **16.8** | **76.4** | **77.0** | **76.7** | **76.7** | **61.1** |
| *Domain adaptation* | | | | | | | | |
| SimCSE-BERT-Base | In-domain | 55.9 | 12.4 | 74.5 | 62.5 | 68.5 | 76.8 | 56.4 |
| SimCSE-BERT-Base | In-domain→NLI | 56.2 | 13.1 | 75.5 | 67.3 | 71.4 | 76.9 | 57.8 |
| TSDAE-BERT-Base | In-domain | 59.4 | 14.5 | 76.8 | 69.2 | 73.0 | 47.4 | 53.5 |
| TSDAE-BERT-Base | In-domain→NLI | 59.4 | 14.4 | 75.8 | 73.1 | 74.5 | 76.5 | 59.8 |
| GenSE | Open-domain→In-domain | **60.3** | **16.2** | 75.0 | **77.3** | 76.2 | 77.8 | 61.3 |
| GenSE | Open-domain→In-domain→NLI | **60.3** | 16.0 | **76.7** | 76.7 | **76.7** | **77.9** | **61.5** |

Table 3: Performance on four transfer downstream tasks from various domains. Average precision (AP) is reported for AskUbuntu, CQADupStack, and Twitter. Spearman's correlation is reported for BIOSSES. The first block shows the result of out-of-the-box supervised sentence embeddings. The second block shows the result using different domain adaptation approaches. For ASkUbuntu, CQADupStack, and Twitter, the baseline results are from (Wang et al., 2021). For BIOSSES, we obtain the baseline results using open-source repo [3].

| Model | STS-B | Trasnfer |
|---|---|---|
| ST5-EncDec-Small | 86.0 | 55.9 |
| +Prompt | 86.5 | 56.6 |
| ST5-EncDec-Base | 87.2 | 57.9 |
| +Prompt | 87.7 | 58.2 |

Table 4: Ablation study on prompt: performance on STS-B dev set, and average performance on four transfer tasks are shown.

model through using in-domain data. The consistent and significant improvements over the four tasks also demonstrate the great generalization ability across various domains of GenSE. Furthermore, compared to direct transfer, we can also achieve an improvement of 0.4% even without labeled QA data, which convincingly demonstrates that GenSE can make full use of unlabeled sentences for better sentence embedding.

### 3.4 Ablation Studies

In this section, we present ablation studies on each component in GenSE, including the prompt learning, model scale, amount of labeled or synthesized data, and synthesis strategy.

**Prompt learning:** It has been shown that adding prompt can improve the encoder-based sentence embedding (Jiang et al., 2022), but the impact of prompt in text-to-text sentence embedding remains unknown. We compare the performance of models trained on NLI with or without prompt. As shown in Table 4, adding prompt consistently improve the performance for both STS-B and transfer tasks across different model sizes. Ablation experiments on T5-Small show that stacking a single decoder cannot further improve the performance. Yet, adding prompt can further boost the performance. We hypothesize that prompt helps to elicit knowledge contained in the decoder, leading to performance improvement.

**Model scale:** We investigate whether GenSE works across different model scales. Due to resource limitation, we conduct experiments on T5-Small. Specifically, we first train a unified data generator from T5-Small. Then we use the model for data synthesis, which results in 34M training pairs. Finally, we fine-tune a T5-Small sentence embedding model with synthetic and NLI data. Table 5 shows the result. GenSE outperforms ST5-Small by 1.6% on average for STS tasks, and 1.1% for transfer tasks. Yet, compared to T5-Base, the performance gain for T5-Small is less significant. It's reasonable since the small data augmenter is less expressive than base one, which produces less synthetic data with lower quality from same amount of unlabelled sentences. We also find that although adding decoder improves the sentence embedding for base and large size models, it brings no benefit for T5-small model. One possibility is that the T5-small decoder architecture is too shallow, which

acts like a mean pooling in encoder-only model.

| Model | # Params | STS | Transfer |
|---|---|---|---|
| ST5-Enc-Small | 30M | 80.9 | 56.0 |
| ST5-EncDec-Small | 60M | 80.9 | 55.9 |
| GenSE-Small | 60M | **82.5** | **57.1** |

Table 5: Performance of models using T5-Small: we report the average performance on STS and transfer benchmarks.

**Synthetic data amount:** We also study how the amount of synthetic data influence the final performance. As shown in Table 6, by adding 20% data, GenSE already achieves substantial improvement. GenSE achieves the best performance on STS-B dev set (88.9) with more data, which even outperforms fine-tuned cross-attention T5-Base (88.02). For the transfer tasks, the average performance continues to improve as more data are used. The experiments validate the idea of using open-domain sentences to improve model generalization ability.

**Supervised data amount:** We investigate whether synthetic data can help reduce the annotation burden. We train another GenSE model with 50% randomly-sampled NLI data, with results in Table 7. Together with Table 4 we can find that the performance of baselines degrades significantly, i.e., 0.6% on STS-B and 1.4% on transfer tasks. However, using synthetic data can greatly alleviate the performance drop, and even outperforms baseline models trained on full NLI data.

**Data synthesis strategy:** We demonstrate the superiority of our data synthesis strategy by comparing to previous methods that utilize synthetic data for sentence embedding under the same semi-supervised setting, i.e.,back-translation (Wieting and Gimpel, 2018; Zhang et al., 2021b) and retrieval (Thakur et al., 2021a). The results are in Table 8. We can see that both retrieval-based method and back-translation can improve the performance, while our generator/discriminator based method can further significantly boost the performance on both tasks. Further experiments for GenSE without hard negatives demonstrates that synthesizing negative pairs plays an important role for the performance improvement, as it provides more information for contrastive learning.

| Data Amount | 0% | 20% | 40% | 60% | 100% |
|---|---|---|---|---|---|
| **STS-B** | 87.7 | 88.7 | 88.8 | 88.7 | 88.9 |
| **Transfer** | 58.2 | 59.3 | 59.5 | 59.7 | 59.8 |

Table 6: Performance of GenSE with different amount of synthetic data. We report STS-B dev set and average transfer performance. 0% refers to directly apply prompt-based contrastive learning on NLI.

| Model | STS-B | Trasnfer |
|---|---|---|
| ST5-EncDec-Base + Prompt | 87.1 | 56.8 |
| GenSE | 88.6 | 59.1 |

Table 7: Performance of models trained with 50% of supervised NLI data. Performance on STS-B dev set and transfer tasks are shown.

| Model | STS-B | Transfer |
|---|---|---|
| ST5-EncDec-Base + Prompt | 87.7 | 58.2 |
| Retrieval | 88.3 | 58.9 |
| Back Translation | 88.0 | 59.6 |
| GenSE | 88.9 | 61.5 |
| w/o synthetic negatives | 88.6 | 59.5 |

Table 8: Comparison of different synthesis strategies on STS-B dev set and transfer tasks

### 3.5 Analysis

**Uniformity and alignment:** We investigate the uniformity and alignment of GenSE, which capture the quality of produced sentence embedding:

$$\mathcal{L}_{align} = - \mathop{\mathbb{E}}_{s,s^+ \sim p_{pos}} \|f(s) - f(s^+)\|, \quad (4)$$

$$\mathcal{L}_{uniform} = \log \mathop{\mathbb{E}}_{s,u \overset{\text{iid}}{\sim} p_{data}} e^{-2\|f(s)-f(u)\|}, \quad (5)$$

where $p_{data}$ is the distribution of positive pairs, and $p_{pos}$ refers to all entailment pairs. Smaller $\mathcal{L}_{align}$ indicates shorter distance between sentence embeddings of positive pairs, and smaller $\mathcal{L}_{uniform}$ means that the embedding space is more uniform.

We follow the setting in (Ni et al., 2021) to measure $\mathcal{L}_{uniform}$ on the whole STS-B test set, and $\mathcal{L}_{align}$ on sentence pairs in STS-B test set with correlation scores higher than 4. As shown in Figure 2, GenSE shows better alignment and uniformity than the SimCSE model based on RoBERTa-Base. Compared to the supervised ST5 baseline models, GenSE achieves much lower uniformity loss but larger alignment loss. We conjecture that GenSE obtains lower uniformity by learning from more synthesized unlabeled data, which might bring some noisy pairs and result in large alignment loss.

| Method | Entailment | Contradiction |
|---|---|---|
| | ***Input:*** *A young man getting ready to release a red kite.* | |
| DINO | A young man releasing a red kite.<br>A red kite releasing a red kite. | A man getting ready to release a red kite.<br>It was a big deal to him and he didn't know how he would explain it to his parents. |
| Retrieval | A man getting ready to fly a kite.<br>A man in a yard getting ready to play with a kite | N/A<br>N/A |
| BackTrans | A young man prepares to release a red kite.<br>A young man is about to release a red kite. | N/A<br>N/A |
| GenSE | The man is prepared to fly the kite.<br>A man is planning to fly a kite. | A man is playing basketball.<br>The woman is flying a kite. |
| | ***Input:*** *One of the hotel's rooms* | |
| DINO | The hotel room<br>One of the hotel rooms | I have no idea what that is.<br>The other one is on fire |
| Retrieval | One of the 300 modular hotel rooms<br>The grand hotel birmingham one of the hotel rooms | N/A<br>N/A |
| BackTrans | One of the hotel rooms<br>One of the rooms of the hotel | N/A<br>N/A |
| GenSE | A room inside a hotel.<br>A hotel room. | There is no room at the hotel.<br>It's not the hotel's room. |

Table 9: Comparison of different data synthesis approaches. For retrieval-based method, only entailment pairs can be obtained. We bold the correct samples through human judgement.

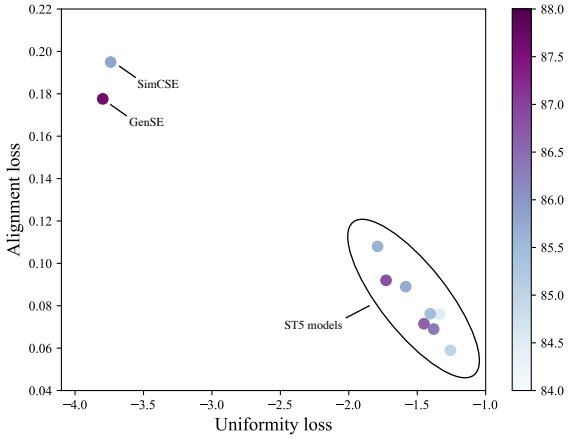


Figure 2: Alignment and uniformity losses plot: The color of dots refer to models' performance on STS-B test split. For ST5 models, we reports the result from the original paper. (Ni et al., 2021)

**Quality of synthetic data:** Several methods have been implemented to synthesize the sentence pairs, including generation-based DINO (Schick and Schütze, 2021), retrieval-based approach (Thakur et al., 2021a), and back-translation, with samples shown in Table 9. For each input sentences, we give two samples using over-generation. We run the experiment on the image caption data CC12M (Changpinyo et al., 2021). For DINO, we use the official repo [4]. For retrieval, we follow (Thakur et al., 2021a) to use BM25 (Amati, 2009) to produce possible pairs, and use a cross-encoder trained on NLI to further label the pair. For back-translation, we use google translator to produce English-French and English-German pairs.

For entailment pair generation, unsupervised DINO (Schick and Schütze, 2021) can generate some meaningful pairs with a large-scale generative model. It also produces many incorrect pairs, and cannot be filtered out since no supervision signal is available. BackTrans, Retrieval and GenSE can all produce entailments efficiently. However, BackTrans usually produces entailment pairs with very high lexical overlap, which fail to give high-quality supervision signals for contrastive representation learning. Retrieval can produce noisy pairs even after filtering due to the limited size of the banking corpus.

For contradiction pair generation, DINO hardly generate correct or related sentences, which cannot serve as hard negatives in sentence representation learning. Although retrieval and back-translation can produce entailment pairs, they cannot generate contradiction pairs. In contrast, GenSE can produce both entailment and contradiction pairs, which are important for contrastive sentence representation learning as demonstrated in Table 8.

## 4 Related Work

Prior approaches for sentence embedding include two main categories: (1) supervised learning with labeled sentences, and (2) unsupervised sentence embedding with unlabeled sentences, while a few early approaches leverage on both of them.

[4] https://github.com/timoschick/dino

Supervised sentence representation learning relies on human-annotated sentence pairs, e.g., NLI data. Early works learn sentence embedding through fine-tuning the model on NLI with classification objectives (Conneau et al., 2017; Cer et al., 2018). Recent works find that contrastive objectives can help learn better sentence representation (Gao et al., 2021; Carlsson et al., 2020). There have been several works exploring the effect of additional supervised training pairs on sentence representation learning (Ni et al., 2021). ST5 (Ni et al., 2021) utilizes question-answer pairs for pre-training before fine-tuning the model on the NLI corpus, leading to better generalization performance.

Many approaches attempt to develop unsupervised objectives for sentence embedding. Early works train the model to predict surrounding sentences (Kiros et al., 2015; Hill et al., 2016; Logeswaran and Lee, 2018). Recent works start to adopt contrastive learning through maximizing the similarity between different view of the same sentences (Zhang et al., 2020; Carlsson et al., 2021; Giorgi et al., 2021; Gao et al., 2021). Recently, Jiang et al. (2022) utilizes the prompt to extract embeddings from encoder models, which inspires the prompt-based contrastive objective in our GenSE. Despite the promising results from unsupervised approaches, there's still a large performance gap between unsupervised and supervised approaches.

In this work, we aim to combine the supervised and unsupervised approach. Similar to our motivation, USE (Cer et al., 2018) uses the SkipThought-like (Kiros et al., 2015) loss for unlabeled sentences, and a classification loss for NLI. However, the performance is unsatisfactory. Recent works mainly focus on using unlabeled data for domain adaptation. Thakur et al. (2021a) first adopts sampling strategies, e.g. BM25 (Amati, 2009) and semantic search, to obtain weakly-labelled pairs, and then uses cross-encoders trained on NLI for re-labelling. Wang et al. (2021) proposes an auto-encoder loss for unsupervised domain adaptation of supervised sentence encoder. Different from these approaches, we utilize data synthesis model (Shorten and Khoshgoftaar, 2019; Tjandra et al., 2020; Feng et al., 2021; Gao et al., 2022a,b) to convert large-scale unlabeled sentences into sentence pairs towards better sentence embeddings.

## 5 Conclusion and Future Work

In this work, we propose a novel semi-supervised framework, GenSE, for sentence representation learning. We first train a unified model for generation and discrimination, which can effectively obtain high-quality synthetic positive and negative sentence pairs from open-domain unlabelled corpus. Afterwards, we train a prompt-based text-to-text sentence embedding model with contrastive learning on both synthesized and labeled NLI data. Extensive results on STS and transfer tasks validate that GenSE can achieve significantly better performance than current state-of-the-arts and exhibit better generalization ability. Future work includes better synthesizing strategies to generate better sentence pairs and advanced designs of semi-supervised sentence representation learning frameworks based on more diverse open-domain data.

## 6 Limitations

Firstly, our generator/discriminator is only trained on NLI data, which makes the generator/discriminator less expressive. As demonstrated in ST5 and GenSE, using semi-structured data, e.g., QA pairs, in contrastive sentence representation learning leads to a significant performance improvement, especially on transfer tasks. We hypothesize that QA pairs contain rich information, and capture very different semantic relationships compared to NLI pairs, which could lead to much stronger generalization ability. Therefore, we plan to include semi-structured data into generator/discriminator training as future work to produce more diverse synthetic pairs.

Secondly, due to insufficient GPU resources, we are unable to scale our model to T5-Large, or using more unlabelled data for training. Therefore, we cannot fully evaluate the potential of synthetic data. In addition, we cannot include large-scale QA data in training, which will lead to a more universal sentence embedding.

Thirdly, multi-stage training on the synthetic data leads to higher computational cost. Training on 61M synthetic triplets takes around 88 GPU hours. As shown in the ablation study on the amount of synthetic data , GenSE continues to improve with more data. To achieve a better trade-off between performance and cost, we leave representative data selection and efficient sentence embedding training as future research directions.

## Acknowledgements

We would like to thank all the reviewers for their constructive comments. This work is supported by Science and Engineering Research Council, Agency of Science, Technology and Research (A*STAR), Singapore, through the National Robotics Program under Human-Robot Interaction Phase 1 (Grant No. 192 25 00054); Human Robot Collaborative AI under its AME Programmatic Funding Scheme (Project No. A18A2b0046); This work is partially supported by the Internal Project Fund from Shenzhen Research Institute of Big Data under Grant T00120220002; This work is also supported by the National Natural Science Foundation of China (Grant No: 62106222).

## References

Giambattista Amati. 2009. *BM25*, pages 257–260. Springer US, Boston, MA.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

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.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

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.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

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.

Xiaoxue Gao, Chitralekha Gupta, and Haizhou Li. 2022a. Automatic lyrics transcription of polyphonic music with lyrics-chord multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2280–2294.

Xiaoxue Gao, Chitralekha Gupta, and Haizhou Li. 2022b. Genre-conditioned acoustic models for automatic lyrics transcription of polyphonic music. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 791–795. IEEE.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. 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 879–895, Online. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Doris Hoogeveen, Karin M Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian document computing symposium*, pages 1–8.

Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2022. Promptbert: Improving BERT sentence embeddings with prompts. *CoRR*, abs/2201.04337.

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. GooAQ: Open question answering with diverse answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 421–433, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluís Màrquez. 2016. Semi-supervised question retrieval with gated convolutions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1279–1289, San Diego, California. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *CoRR*, abs/2108.08877.

Christian S Perone, Roberto Silveira, and Thomas S Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

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

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021a. Augmented

SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021b. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Machine speech chain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:976–989.

Bin Wang, C.-C. Kuo, and Haizhou Li. 2022. Just Rank: Rethinking evaluation with word and sentence similarities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6060–6077, Dublin, Ireland. Association for Computational Linguistics.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

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.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021a. DynaEval: Unifying turn and dialogue level evaluation. 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 5676–5689, Online. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yan Zhang, Ruidan He, Zuozhu Liu, Lidong Bing, and Haizhou Li. 2021b. Bootstrapped unsupervised sentence representation learning. 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 5168–5180, Online. Association for Computational Linguistics.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610, Online. Association for Computational Linguistics.