# Appendicies for Submission 7971: DenoSent: A Denoising Objective for Self-Supervised Sentence Representation Learning

**Xinghao Wang, Junliang He, Pengyu Wang, Yunhua Zhou, Tianxiang Sun, Xipeng Qiu***

School of Computer Science, Fudan University

{xinghaowang22, jlhe22, pywang22}@m.fudan.edu.cn, {zhouyh20, txsun19, xpqiu}@fudan.edu.cn

## Comparison Between Discrete Perturbation Strategies

In the main paper, we propose to use two discrete perturbation approaches to modify the expression or syntax of training sentences while retaining their inherent semantics, including utilizing back-translation or leveraging an instruction-tuned large-scale language model. Here we provide some more details.

For the back-translation method, we use pre-trained machine translation models available on Huggingface to translate the Wiki1m dataset (proposed in SimCSE) into Chinese[1] then back into English[2].

When using an LLM for discrete perturbations, we use the `gpt-3.5-turbo` API[3]. The following prompt is used to generate sentences with similar semantics to the original sentences in the Wiki1m dataset:

```
You are a data generator that is aware of sentence-level semantics. Your task is to generate sentences based on the user's input sentence. Here are the requirements:
1. Produce a sentence as "sentence1", based on the given input, that retains the same semantic information as the input sentence.
2. Produce a sentence as "sentence2", based on the given input, that has the contrary semantic information as the input sentence or is completely unrelated to the input sentence.
3. The sentences you generated should be diverse, coherent and grammatically correct.
4. The sentences you generated should be varied in length, syntax and expression from the input sentence.
```

We generated only 126,014 samples using the Large Language Model (LLM) owing to resource constraints. This limitation is likely the primary factor contributing to the underperformance of this approach compared to the back-translation method.

We include the detailed evaluation results for both discrete perturbation strategies in the subsequent sections.

## MTEB STS Results

Table 1 illustrates the evaluation performance on 7 STS tasks using the MTEB toolkit. These findings align closely with those presented in the main paper.

## Detailed results of Reranking & Retrieval Tasks.

Table 2 illustrates the detailed evaluation results for each of the reranking and retrieval tasks.

## Detailed results of Classification Tasks

Table 3 illustrates the detailed evaluation results for each dataset.

## Datasets Details

We use the unsurprised dataset adopted in the SimCSE paper, which contains 1000000 samples. For datasets for evaluation, the number of testing samples for each dataset is listed in Table 4. Note that we use all the available sentence-level datasets for reranking, retrieval and classification in MTEB.

*Corresponding author



[1] https://huggingface.co/Helsinki-NLP/opus-mt-en-zh

[2] https://huggingface.co/Helsinki-NLP/opus-mt-zh-en

[3] https://platform.openai.com/docs/models/gpt-3-5

| Model | STS12 | STS13 | STS14 | STS15 | STS16 | STS-B | SICK-R | Avg. |
|---|---|---|---|---|---|---|---|---|
| SimCSE | 68.4 | 82.41 | 73.81 | 80.91 | 78.56 | 76.39 | 72.55 | 76.15 |
| PaSeR | 69.05 | 82.83 | 71.5 | 82.87 | 76.3 | 77.68 | 63.79 | 74.86 |
| PromptBERT | 72.38 | **83.83** | 76.1 | **84.32** | **81.38** | **81.7** | 69.72 | 78.49 |
| SNCSE | 69.54 | 83.02 | 75.04 | 81.96 | 80.00 | 80.88 | 73.67 | 77.73 |
| **DenoSent**$_{backtrans}$ | 75.57 | 83.77 | **77.25** | 84.30 | 79.5 | 80.81 | **74.46** | **79.38** |
| **DenoSent**$_{LLM}$ | **76.15** | 82.86 | 76.69 | 82.68 | 81.18 | 81.12 | 72.36 | 79.01 |

Table 1: MTEB STS Results

| Model | AskUbuntuDupQuestions | | MindSmallReranking | | SciDocsRR | | StackOverflowDupQuestions | | QuoraRetrieval | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAP | MRR | MAP | MRR | MAP | MRR | MAP | MRR | MAP | MRR |
| SimCSE | 51.88 | 64.47 | 28.68 | 29.49 | 67.87 | 88.26 | 39.57 | 39.91 | 60.99 | 70.4 |
| PaSeR | 51.78 | 66.13 | 29.59 | 30.55 | 67.62 | 87.84 | 40.97 | 41.21 | 63.47 | 72.89 |
| PromptBERT | **53.63** | **67.57** | 27.39 | 27.99 | 65.65 | 86.84 | 40.8 | 41 | 64.27 | **74.15** |
| SNCSE | 52.79 | 65.83 | 28.44 | 29.22 | 69.98 | **89.83** | **41.84** | **42.22** | 63.07 | 72.72 |
| **DenoSent**$_{backtrans}$ | 52.81 | 65.92 | **29.81** | **30.76** | 68.17 | 88.27 | 41.75 | 41.86 | **64.36** | 74.06 |
| **DenoSent**$_{LLM}$ | 53.01 | 65.28 | 28.19 | 28.87 | **70.01** | 89.69 | 41.42 | 41.76 | 63.44 | 73.02 |

Table 2: Detailed results of Reranking & Retrieval Tasks.

| Model | ACC | ARC | BC | EC | MIC | MSC | MTOPDC | MTOPIC | TCC | TSEC | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Glove$^{\heartsuit}$ | 56.91 | 29.67 | 67.69 | 36.93 | 56.19 | 66.03 | 79.11 | 55.85 | 65.40 | 50.80 | 56.42 |
| BERT (CLS)$^{\heartsuit}$ | **74.25** | 33.56 | 63.41 | 35.28 | 59.88 | 64.28 | 82.63 | **68.14** | **70.00** | 51.81 | 60.32 |
| SimCSE | 68.54 | 34.60 | 74.41 | 43.27 | 61.32 | 67.86 | 83.92 | 61.28 | 69.35 | 54.27 | 62.73 |
| PaSeR | 68.00 | 33.02 | 76.89 | 43.53 | 62.59 | 67.53 | 85.86 | 63.33 | 68.33 | 52.46 | 63.23 |
| PromptBERT | 63.67 | 35.09 | **80.05** | **46.15** | 62.94 | 68.92 | 85.20 | 63.39 | 68.63 | 56.19 | 63.78 |
| SNCSE | 68.69 | **36.02** | 75.84 | 43.26 | 62.34 | 68.38 | 82.46 | 60.42 | 67.94 | 54.09 | 62.82 |
| **DenoSent**$_{backtrans}$ | 65.70 | 35.97 | 76.82 | 45.10 | 65.31 | 70.54 | 86.38 | 64.64 | 69.66 | **56.52** | **64.46** |
| **DenoSent**$_{LLM}$ | 67.87 | 33.34 | 76.84 | 42.55 | **65.85** | **71.62** | **86.52** | 63.87 | 66.13 | 54.67 | 62.93 |

Table 3: Detailed results of Classification Tasks. ♡: results from the MTEB paper. Abbreviations: ACC, ARC, BC, EC, MIC, MSC, MTOPDC, MTOPIC, TCC and TSEC denotes AmazonCounterfactual, AmazonReviews, Banking77, Emotion, MassiveIntent, MassiceScenario, MTOPDomain, MTOPIntent, ToxicConverstions and TweetSentimentExtraction, respectively.

| Dataset | Number of test samples |
|---|---|
| *Semantic Textual Similarity* | |
| STS12 | 3108 |
| STS13 | 1500 |
| STS14 | 3750 |
| STS15 | 3000 |
| STS16 | 1186 |
| STS-Benchmark | 1379 |
| SICK-Relatedness | 4927 |
| *Reranking & Retrieval* | |
| AskUbuntuDupQuestions | 2255 |
| MindSmallReranking | 107968 |
| SciDocsRR | 19599 |
| StackOverflowDupQuestions | 3467 |
| QuoraRetrieval | 532931 |
| *Classification* | |
| AmazonCounterfactualClassification | 670 |
| AmazonReviewsClassification | 30000 |
| Banking77Classification | 3080 |
| EmotionClassification | 2000 |
| MassiveIntentClassification | 2974 |
| MassiveScenarioClassification | 2974 |
| MTOPDomainClassification | 4386 |
| MTOPIntentClassification | 4386 |
| ToxicConversationsClassification | 50000 |
| TweetSentimentExtractionClassification | 3534 |

Table 4: Number of samples for each dataset used in the experiments.