Augmented Bio-SBERT: Improving Performance For Pairwise Sentence Tasks in Bio-medical Domain

Sonam Pankaj
Rasa
s.pankaj@rasa.com

Amit Gautam
Saama Technologies
amit.gautam@saama.com

Abstract

One of the modern challenges in AI is the access to high-quality and annotated data, especially in NLP; that's why augmentation is gaining importance. In computer vision, where image data augmentation is standard, text data augmentation in NLP is complex due to the high complexity of language. And we have seen advantages of augmentation where there are fewer data available, and it can play a massive role in improving the model's accuracy and performance. We have implemented Augmentation in Pairwise sentence scoring in the biomedical domain.

By experimenting with our approach to downstream tasks on biomedical data, we have looked into the solution to improve Bi-encoders' sentence transformer performance using augmented data-set generated by cross-encoders fine-tuned on Biosses and MedNLI on pretrained Bio-BERT model. It has significantly improved the results with respect to the only the model only trained on Gold data for the respective tasks.

1 Introduction

Language models are data hungry; they consume massive amounts of data to identify patterns. For many niches, low-resource domains like that of Bio domain NLP, manually finding or annotating a substantial dataset is complicated. Bio-domain language models comparison (Peng et al., 2019). Fortunately, we don't need to label this new data; we can automatically generate or label data using one or more data augmentation techniques. Pairwise sentence scoring tasks are used widely in NLP Applications. (Thakur et al., 2020) like information retrieval, question answering, duplicate question detection, or clustering. Pre-trained transformers have led to remarkable progress in several tasks, especially BERT (Devlin et al., 2019) is an approach that sets new state-of-the-art performance for many tasks, including pairwise sentence scoring. For tasks that make pairwise comparisons between sequences, matching a given input with a corresponding label, two approaches are common: Cross-encoders and Bi-encoders. We pre-trained Cross-encoders on the gold dataset, then outputs of different pairs from random sampling and semantic sampling were fed to the cross-encoder. The silver dataset produced was then provided to Bi-encoder. There is a new approach like Poly-encoder, which mostly fits tasks around conversational AI. We have used BioBERT, a biomedical language representation model designed for biomedical text mining tasks such as biomedical named entity recognition, relation extraction, question answering, etc. cross- and bi- encoders, details on pretrained model architecture have been discussed in section 4.2 and 4.3. We worked on two tasks for pairwise sentences—semantic similarity and Language Inferences based on medical data such as Biosses and MedNLI. We have evaluated the results on the test set of each data set. In the end, we have compared our model results with results of textual similarity and inference tasks on a blue benchmark and have included the Table 1.

2 Related Work

There have been many NLP-augmentation methods based on paraphrasing models and non- paraphrasing models. A Survey of Data Augmentation Approaches for NLP highlights techniques used for popular NLP applications and tasks, (Feng et al., 2021) like mitigating biases and fixing class imbalance. Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks have worked on sentence pair scor- ing through cross-encoder and Bi-encoder on general language(English). But available language evaluation doesn't fit bio-medical uses, given it can't relate to the biomedical domain.

Proceedings of the 29th International Conference on Computational Linguistic, pages 43–47
October 12–17, 2022.

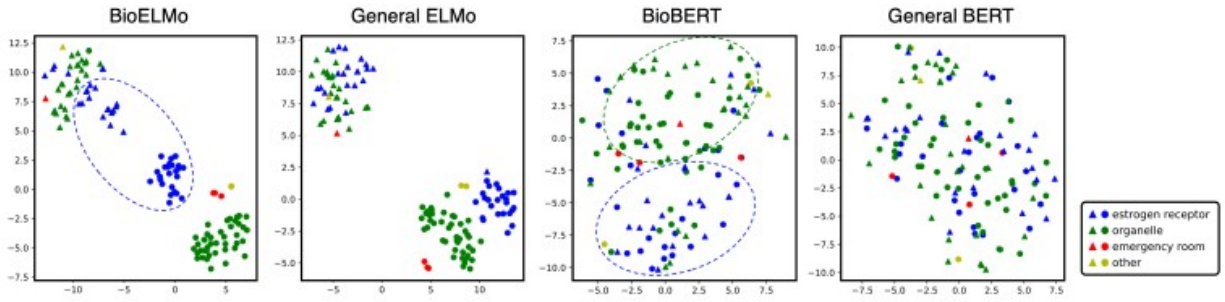

Figure 1: Comparison of BioELMO, General Elmo, BioBERT and General BERT

Probing Biomedical Embeddings from Language Models , (jin, 2019) this paper deals with the contrast in general, English and biomedical on any down- streaming tasks, and the results show that even pre-trained by in-domain corpus as a fixed feature extractor, BioBERT still cannot effectively encode biomedical relations compared to BERT. BioELMo is significantly better than ELMo in representing same relations closer to each other. Augmentation with general language BERT will perform poorly, as it has been compared in Figure 1.

3 Downstreaming Tasks

We have used two tasks for pairwise sentences: natural language inference and semantic textual similarity. The dataset for training and testing have been described in section 5.1.

Natural language inference (NLI): is the task (Romanov and Shivade, 2018) of determining whether a given hypothesis can be inferred from a given premise. We are using, MedNLI - a publicly available, expertly annotated dataset for NLI in the clinical domain.

Semantic textual similarity: Semantic textual similarity deals with determining how similar two pieces of texts are. Related tasks are paraphrase or duplicate identification. We have used the Biosses (Gizem Soganciogl u, 2017) dataset for the same.

4 Methods

In this section, we will describe the pre-trained model used, the fine-tuning, cross- and bi- encoders, and the step-by-step method of getting the predictions from the bi-encoder sentence transformer after fine-tuning it on a silver dataset. We will also discuss the sampling techniques used for creating a silver dataset.

1. Cross-Encoders are trained on gold label dataset of Biosses and MedNLI
2. Sample pairs of sentences by using methods of Random Sampling and Semantic Search Sampling
3. Predicting similarity on trained cross-encoder that we trained on a gold dataset. This makes the silver dataset
4. Training the Bi-Encoder SBERT based on the silver Dataset
5. Predicting the results and comparing the mix of (Gold and Silver dataset) to Gold Dataset

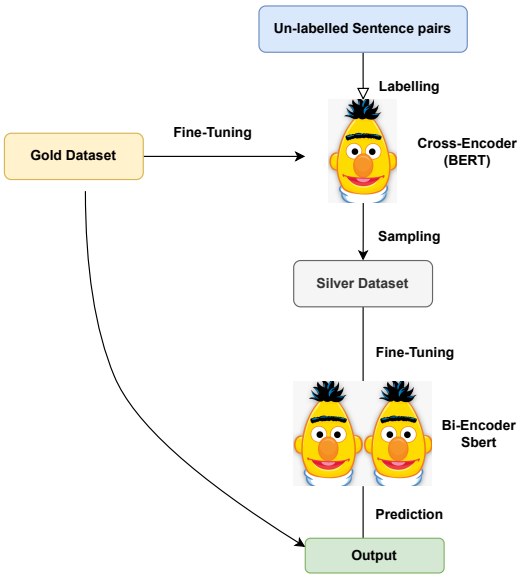

Figure 2: Architecture of training bi-encoders on silver dataset

4.1 Model

Pretrained Model: BioBERT largely outperforms BERT and previous state-of-the-art models

BLUE sentence-pair tasks dataset			
Corpus	gold	gold and silver	test-datset
Biosses	80	880	20
MedNLI	11232	67377	1422

Table 1: dataset for training and testing

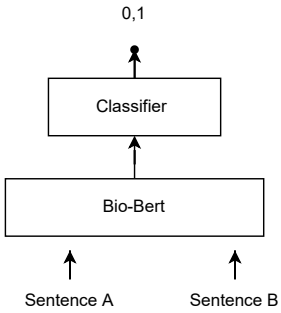

Figure 3: Scoring of Cross-encoder Architecture

in a variety of biomedical text mining tasks when pre-trained on biomedical corpora. Due to this, we have decided to use this sentence transformer based on this model.

SBERT: Given a pre-trained, well-performing cross-encoder, we sample sentence pairs according to a specific sampling strategy (discussed later) and label these using the cross-encoder. We call these weakly labeled examples the silver dataset, and will merge both with the gold training dataset. We then train the bi-encoder on this extended training dataset. We refer to this model as Augmented SBERT (AugSBERT). In Figure 2. we have illustrated the process of Augmented SBERT

Fine-tuning Model: Fine-tuning sentence transformer models requires pairs of labeled data, cross-encoder model fine-tuned on gold dataset, and the bi-encoder fine-tuned on gold and silver data [1].

4.2 Cross-encoder

In a cross-encoder, both sentences are passed to the network, and attention is applied across all tokens of the inputs. This approach is in Figure 3, where both sentences are simultaneously passed to the network. (Gizem Soganciogl u, 2017). It is a single Bio-BERT inference step that takes both sentences as a single input and outputs a similarity score.

Pretrained Model: We have used pretrained Bi-encoder dmis-lab/biobert-v1.1 from DMIS-Labs via hugging face, and finetuned it on gold data.

4.3 Bi-encoder

For a given sentence, bi-encoder produce a sentence embedding. We independently pass to a Bio-BERT the sentences A and B, which result in the sentence embedding u and v. These sentence embeddings can then be compared using cosine similarity, and thus we get similarity score as shown in Figure 4.

Pretrained Model: We have used pretrained Bi-encoder dmis-lab/biobert-v1.1 from DMIS-Labs via hugging face, and finetuned it on gold+silver data.

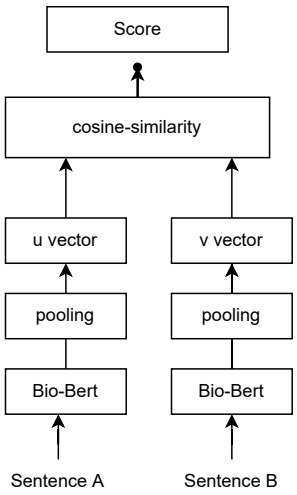

Figure 4: Bi-encoder Architecture

4.4 Sampling Techniques

We used random and semantic search sampling to make new pairwise data.

Random Sampling (RS) : We randomly sample a sentence pair and weakly label it with the cross-encoder. Randomly selecting two sentences usually leads to a dissimilar (negative) pair; positive pairs are extremely rare. This skews the label distribution of the silver dataset heavily towards negative pairs.

Semantic Search Sampling (SS) : We train a bi-encoder (SBERT) on the gold training set (Reimers and Gurevych) and use it to sample further similar sentence pairs. We use cosine-similarity and

[1] codes available in supplementary

Dataset Sampling	Biosses	MedNLI
Gold (baseline)	71.4	12.7
Gold and Silver (Random Sampling)	74.5	57.9
Gold and Silver (Semantic Sampling)	88.5	74

Table 2: Comparison with our model comparison of Gold with respect to Gold + Silver on Biosses and MedNLI

retrieve the top five most similar sentences in our collection on every sentence. We used pretrained-model which is BioBERT fine-tuned on the SNLI and the MultiNLI datasets using the sentence-transformers library to produce universal sentence embeddings.

5 Experiment-Setup

We conducted the experiments using PyTorch Hugging Face's transformers (Wolf et al., 2019), and we used google-colab to import the transformers, cross-encoders, and sentence transformers. The latter showed that BERT outperforms other transformer-like networks when used as a bi-encoder. Baselines are just Bio-Bert output on a gold dataset. Baseline with gold has been measured only with bi-encoder

5.1 Datasets

Sentence pair scoring can be differentiated in regression and classification tasks. Regression tasks assign a score to indicate the similarity between the inputs.

BIOSSES: Several approaches have been proposed for semantic sentence similarity estimation for generic English. Biosses is a benchmark data set consisting of 100 sentence pairs from the biomedical literature that is manually annotated by five human experts and used for evaluating the proposed methods.

MedNLI: MedNLI is a dataset annotated by doctors, performing a natural language inference task (NLI) grounded in patients' medical history. The MedNLI dataset consists of the sentence pairs developed by Physicians from the Past Medical History section of MIMIC-III clinical notes annotated for Definitely True, Maybe True, and False.

5.1.1 Benchmarks

Both the datasets are present in BLUE Benchmark, for pairwise sentences for similarity and inference.

Models	Biosses	MedNLI
ELMO	60.2	71.4
BioBERT	82.7	80.5

Table 3: Accuracy of different language models for corpus

Table 3 [2] gives us the comparison of different models on different tasks of BLUE benchmark.

6 Results

The results section includes experimentation on the only gold dataset and the gold and silver dataset, using both methods, random sampling, and semantic similarity sampling, given in the Table 3.

Depicting the silver dataset helps improve the model, produced by using SBERT augmentation and mixed with gold labels. We have also compared it with the results of ELMO and BIOBERT results given in Table 3.

7 Conclusion

As language models get bigger, so do datasets. And although we have seen an explosion of data in the past decade, it is often not accessible, especially in niche domains like Bio-domain. And there are datasets, like Biosses, which has only 100 pairs of sentences. Thus finding a substantial annotated dataset becomes difficult.

Sentence-BERT augmentation can be used to improve pairwise sentence models and sentence similarity datasets. Like in our case, it has improved results by up to 23.9 percent in the case of the Biosses and 482 percent in case of MedNLI, as discussed in Table 2.

[2] data has been taken from https://arxiv.org/pdf/1906.05474v2.pdf

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

Özgür Gizem Soganciogl u, Öztürk H. 2017. Biosses: A semantic sentence similarity estimation system for the biomedical domain.

dhingra jin. 2019. probing biomedical embeddings from language models.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2020. Augmented SBERT: data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *CoRR*, abs/2010.08240.