BIOptimus: Pre-training an Optimal Biomedical Language Model with Curriculum Learning for Named Entity Recognition

Vera Pavlova
rttl.ai
Dubai, UAE
v@rttl.ai

Mohammed Makhlouf
rttl.ai
Dubai, UAE
mm@rttl.ai

Abstract

Using language models (LMs) pre-trained in a self-supervised setting on large corpora and then fine-tuning for a downstream task has helped to deal with the problem of limited label data for supervised learning tasks such as Named Entity Recognition (NER). Recent research in biomedical language processing has offered a number of biomedical LMs pre-trained using different methods and techniques that advance results on many BioNLP tasks, including NER. However, there is still a lack of a comprehensive comparison of pre-training approaches that would work more optimally in the biomedical domain. This paper aims to investigate different pre-training methods, such as pre-training the biomedical LM from scratch and pre-training it in a continued fashion. We compare existing methods with our proposed pre-training method of initializing weights for new tokens by distilling existing weights from the BERT model inside the context where the tokens were found. The method helps to speed up the pre-training stage and improve performance on NER. In addition, we compare how masking rate, corruption strategy, and masking strategies impact the performance of the biomedical LM. Finally, using the insights from our experiments, we introduce a new biomedical LM (BIOptimus), which is pre-trained using Curriculum Learning (CL) and contextualized weight distillation method. Our model sets new states of the art on several biomedical Named Entity Recognition (NER) tasks. We release our code and all pre-trained models.[1]

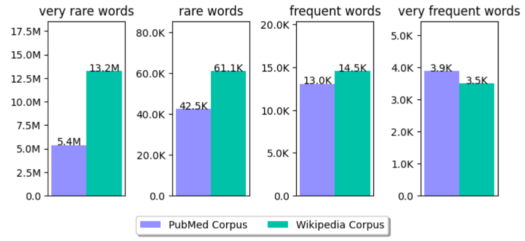


Figure 1: Comparison of word frequency (raw count) for Wikipedia Corpus and PubMed corpus. Frequencies are divided into four categories. Each bar chart represents how many words of that frequency category are found in both corpora.

1 Introduction

Since the introduction of transformer architecture (Vaswani et al., 2017), transfer learning has gained immense popularity in Natural Language Processing (NLP). Pre-training LMs on a large corpus (Devlin et al., 2019; Liu et al., 2019) allows a large-scale knowledge extraction. In the current pipeline of knowledge transfer in NLP, the pre-training stage can be viewed as preparing the model with broad general knowledge (Raffel et al., 2019) that can later be fine-tuned to new tasks or adapted to different domains. However, applying these models in a new domain directly without additional pre-training steps cannot achieve high performance because of a word distribution shift from a general domain vocabulary to a new domain vocabulary. The importance of a pre-trained model for the biomedical domain is even more significant due to the difficulties of constructing supervised training data because it requires expert knowledge for quality annotations.

The biomedical domain has greatly benefitted from domain-specified LMs. To date, various methods have been developed to pre-train biomedical language models: pre-training from scratch (Gu et al., 2021; Beltagy et al., 2019), continual pre-training (Lee et al., 2019; Huang et al., 2019), and hybrid approaches (Tai et al., 2020; Poerner et al., 2020; Sachidananda et al., 2021). Nonetheless, more research is needed to compare the optimal ways to pre-train the biomedical LMs. Comparison of different pre-training approaches and techniques primarily targets general domain LMs (Wettig et al.,

[1] The source code and the models are released on our Github https://github.com/rttl-ai/BIOptimus, including the Rust code for qik-find.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 337–349
July 13, 2023 ©2023 Association for Computational Linguistics

2023; Dai et al., 2022; Yamaguchi et al., 2021; Gu et al., 2020). What still needs to be clarified is the effects of these methods on the pre-training biomedical LM.

Typically, general domain LMs are pre-trained on Wikipedia, Book Corpora, and Common-Crawl (Devlin et al., 2019; Lan et al., 2020; Radford et al., 2019; Raffel et al., 2019; Clark et al., 2020). These corpora are considerably different from the biomedical corpora in terms of domain specifics (use of biomedical terminology) and the style of scientific text (most of the models are trained on PubMed). Thus, the methods and techniques that work well for the general domain are not guaranteed to work the same for the biomedical domain. Figure 1 illustrates the difference in the raw counts of word frequencies between Wikipedia and PubMed, which contains only abstracts (PubMed Corpus). The corpora are of similar sizes (3.1 and 3.6 billion tokens, correspondingly). We can observe that, for example, the count of rare words noticeably differs between two corpora and twice less for the PubMed Corpus. Yu et al. (2022) and Ethayarajh (2019) showed how the presence of rare words in the pre-training corpus might greatly impact the performance of the LM. This work aims to compare different techniques and methods for creating biomedical LMs. We compare such approaches as pre-training from scratch, continued pre-training with no specialized vocabulary, and continued pre-training with biomedical vocabulary. Our evaluations show that a biomedical LM pre-trained in a continued setting speeds up the pre-training. To account for the lack of specialized vocabulary that inevitably accompanies models pre-trained following a continued approach, we propose a new method to initialize weights for new tokens by distilling existing weights from the BERT model inside the context where they were found. Furthermore, we compare how masking rate, corruption strategy, and masking strategies impact the performance of the biomedical LM. We observe that the corruption strategy of "80-10-10" (Devlin et al., 2019) may play a role in the representation degeneration problem of LM by increasing the degree of anisotropy in the contextualized word representations. With relation to the word representation, anisotropy characterizes an embedding space that is restricted into a narrow cone-like shape (Ethayarajh, 2019). This phenomenon leads to a decrease in word embeddings' expressiveness and is referred to as the representation degeneration problem (Gao et al., 2019).

Finally, based on the results of our experiments, we introduce a model which pre-trained with a CL schedule. CL is an easy-to-hard training strategy (Soviany et al., 2022). Though very successful and widely adopted in NLP, it is still under-explored in pre-training LMs. To the best of our knowledge, CL was not used before to pre-train the biomedical LM. We propose the CL method based on the task complexity of predicting masked labels. Our model (BIOptimus), pre-trained with the contextualized weight distillation and following our CL method, sets new states of the art on several biomedical Named Entity Recognition (NER) tasks. In summary, our contributions are:

- Firstly, we propose a new approach to initialize weights to pre-train a biomedical LM that leverages the efficiency of pre-training in a continued fashion with specialized biomedical vocabulary that improves the model's performance on NER tasks.

- Secondly, we comprehensively compare our approach with other pre-training approaches. In addition, we experiment with masking ratio, corruption strategy, and masking strategies, testing what works more optimally to prepare a language model for the biomedical domain.

- Following experimental insights, we propose a biomedical model (BIOptimus) pre-trained with the CL method and contextualized weight distillation.

2 Prior Literature

Recent studies indicate that adapting existing LMs pre-trained on general corpora for a new domain or pre-training a new LM from scratch for a specific domain gives a substantial gain in terms of performance for a downstream task. The authors of BioBERT (Lee et al., 2019) performed domain adaptation by continuing pre-training BERT (Devlin et al., 2019) on a large corpus of biomedical data. The model showed notable achievement on downstream tasks, outperforming BERT and the state-of-the-art models. Huang et al. (2019) further pre-train BioBERT using clinical corpora to address clinical text's challenges and released ClinicalBERT. Gururangan et al. (2020) showed the benefits of continued domain pre-training followed by task-adaptive pre-training (TAPT) of RoBERTa on four domains, including biomedical. One of the downsides of continued pre-training is the lack of domain-specific vocabulary. The authors of the SciBERT model (Beltagy et al., 2019) constructed a new Scivocab using the SentencePiece library, introducing 58% of new tokens, which shows how much scientific vocabulary differs from a general vocabulary of BERT-base. The pre-training corpus comprises 82% of the biomedical domain; the rest is the computer science domain. Training from scratch solely on the biomedical domain showed further gain and was presented by the authors of PubMedBert (Gu et al., 2021). They construct a new domain-specific vocabulary, proving its effectiveness when fine-tuning for downstream tasks. Another approach to pre-train the biomedical LM from scratch using citation links is the LinkBERT model, which achieved new state-of-the-art on several BioNLP tasks (Yasunaga et al., 2022). To mitigate the high costs of pre-training from scratch, Tai et al. (2020) introduced extension vocabulary that is not found in the vocabulary of BERT. The weights of the embedding layer of the extension vocabulary are randomly initialized, then further pre-trained on the biomedical corpus. To avoid randomly initializing weights for new tokens and speed up pre-training, Poerner et al. (2020) used word2vec. Word2vec vectors are trained on the biomedical domain and then aligned with wordpiece vectors from BERT. Sachidananda et al. (2021) experimented with subword-based initialization using the mean of RoBERTa fixed subword embeddings.

Curriculum Learning Curriculum learning (Elman, 1993; Bengio et al., 2009) has been successfully explored in NLP to solve different tasks. It was widely adopted in machine translation (Guo et al., 2020; Liu et al., 2020; Wang et al., 2020, Zhan et al., 2021). Some works have been done in the area of answer generation (Liu et al., 2018), relation extraction (Huang and Du, 2019), reading comprehension (Tay et al., 2019), and NER (Jafarpour et al., 2021). Xu et al. (2020) used CL during the fine-tuning stage of BERT. Nagatsuka et al. (2021) applied CL to pre-train RoBERTa by gradually increasing the block size of the text. Lee et al. (2022) proposed a concept-based curriculum masking.

3 Method

In continued biomedical domain adaptation, pre-training using the checkpoint of the model pre-trained on the general domain can significantly speed up training. Moreover, the benefits of continued domain adaptation are highly aligned with the present trends of NLP communities (Bommasani et al., 2021) of reducing the environmental impact and making knowledge readily available in limited resources settings. One of the problems of continued pre-training of LMs for domain adaptation is, loosely speaking, the absence of domain-specific vocabulary (Guo and Yu, 2022). The issue of out-of-vocabulary (OOV) words is more elegantly solved with subword tokenization such as Byte-Pair Encoding (BPE) (Sennrich et al., 2016), WordPiece (Song et al., 2021), and SentencePiece (Kudo and Richardson, 2018). Nevertheless, the presence of a whole word or more meaningful subwords in domain-specific vocabulary greatly enhances performance on downstream tasks (Beltagy et al., 2019; Gu et al., 2021; Guo and Yu, 2022). There are ways to address this issue by incorporating domain-specific tokens and assigning new token weights without initializing them from scratch (Sachidananda et al., 2021; Poerner et al., 2020). Nevertheless, none of the current approaches leverage the advantage of contextualized embeddings of transformer models. To boost knowledge transfer, we introduce the new vocabulary by distilling existing weights from the BERT model inside the context of the domain where they were found. Our approach uses WordPiece Tokenization (Song et al., 2021) to tokenize PubMed Corpus [2] (BioMedTokenizer) with a vocabulary size of 30522. We contextualize tokens in the following way (see Figure 2):

- First, we perform tokenization with BioMedTokenizer. For example, suppose BioMedTokenizer has a token "bronchoconstriction" absent from the original bert-base-uncased vocabulary. In that case, the token will be broken into six pieces: [’bro’,’##nch’,’##oco’,’##nst’,’##ric’, ’##tion’] (see Figure 2). We use the mean operation of constituting tokens to compute a single representation for a new token (distilled representations):

$$t_{distilled} = f(t_1, ..., t_k) \\ f \in \{mean\} \tag{1}$$

The variable k represents the number of the

[2] https://pubmed.ncbi.nlm.nih.gov. Abstracts published before Jan 2023.

subtokens that make up the token of interest. We average token weights of only the last layer of the BERT model as it is more expressive of context and domain information (Peters et al., 2018).

- In the next step, we compute aggregated representation by sampling sentences that contain tokens of interest from the same PubMed Corpus and compute aggregated weight across several sentences (contextualized representations):

$$t_{context} = g(t_{distilled}, ..., t_m) \quad g \in \{mean\} \tag{2}$$

  The size of sentences sampled per token *m* may vary. We sample uniformly at random, setting the upper bound equal to 20 and the lower bound to 1. In case the token is not found in the corpus (such tokens represent less than 1/10 of the whole vocabulary), we assign distilled representation to this token. We use mean operation, which was found to be the most efficient operation for aggregating across several contexts (Bommasani et al., 2020).

The averaging of the weights of subtokens constitutes more meaning when placed in context. Moreover, contextualization of tokens' embeddings leverages the ability of the BERT model to create domain-specific word embeddings that align with the corpus where they were found. (Aharoni and Goldberg, 2020). We call this approach **contextualized weight distillation**.

4 Models

To perform a more rigorous comparison and find an optimal approach to the pre-train biomedical LM, we pre-train four models (see Table 1):

- The first model is pre-trained from scratch, further referred to **as the biomedical model from scratch (BM from scratch)**. We train WordPiece tokenizer on PubMed corpus and construct new biomedical vocabulary. The model is trained on the same corpus with all the weights initialized randomly.
- The second model is pre-trained from the BERT-base model checkpoint[3]. We call it **the biomedical model continued (BM continued)**.
- The third model is pre-trained in a hybrid approach called **the biomedical model averaged token weights (BM averaged)**. We use BioMedTokenizer to construct the biomedical vocabulary and find an intersection between BERT-base vocabulary and biomedical vocabulary. The weights of the tokens, which are common for both vocabularies, are copied from BERT-base vocabulary to new model weights by directly mapping weights to the corresponding tokens or subtokens. The weights of new domain-specific tokens absent in the BERT-base model are synthesized by averaging the corresponding subtokens representations and assigning this averaged representation to the token of interest in the embedding matrix of the new model. This method corresponds to the step "distilled representations" described in Section 3.
- The fourth model is pre-trained following the proposed method in Section 3. We refer to this approach **biomedical model contextualized weights (BM contextualized)** or **BIOptimus 0.1**. We assign tokens' weights that are common for BERT-base vocabulary and newly constructed biomedical vocabulary in the same manner as described above in the **biomedical model averaged token weights**. Contextualization of tokens is performed only for the tokens that are not found in the BERT-base vocabulary.

	Pre-training method	**Weight initialization**
BM from scratch	from scratch	randomly
BM continued	continued	from the existing checkpoint
BM averaged	continued	from the existing checkpoint + averaging subtokens
BM contextualized	continued	from the existing checkpoint + contextualized weight distillation

Table 1: Four pre-trained models with different pre-training approaches and initialization methods.

5 Data

The primary data used for pre-training is the same PubMed Corpus used for training BioMedTokenizer. To speed up experiments, we choose a random subset of this corpus of 1.8 billion words for pre-training models, representing approximately half of the data generally used for pre-training

[3] https://huggingface.co/bert-base-uncased

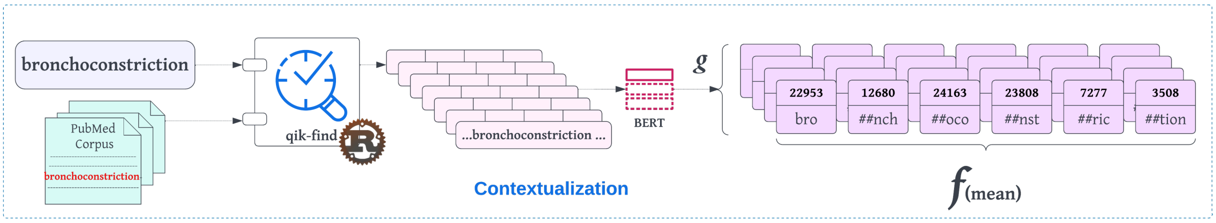


Figure 2: Contextualization of tokens' embeddings with our resource-efficient and performant qik-find tool written in the Rust programming language. Qik-find is a purpose-built tool to find tokens of interest and extract corresponding sentences from a large corpus exploiting the native capability of the Rust programming language for efficient multiprocessing.

models for biomedical domain adaptation. To test our models on the downstream task, we use NER tasks. NER is one of the most fundamental biomedical tasks; it is an essential first step in processing literature for biomedical text mining. Moreover, the NER is an excellent test for evaluating a domain-specific model's success to recognize different types of biomedical terminology. The recent initiative of Microsoft Research (Gu et al., 2021) unified and released a new BLURB benchmark (Biomedical Language Understanding Reasoning Benchmark). The Benchmark includes five NER datasets: BC5-chem (Li et al., 2016), BC5-disease (Li et al., 2016), NCBI-disease (Dogan et al., 2014), BC2GM (Smith et al., 2008), JNLPBA (Kim et al., 2004). Label distribution (Crichton et al., 2017) is presented in Appendix A.

6 Experimental Set-up

We pre-train all models with the same hyperparameters (details can be found in Appendix A). We further fine-tune each model for five NER datasets and report results in Table 2. Fine-tuning details can be found in Appendix A. To compare the effectiveness of pre-training approaches more closely, we show the effects of each pre-training epoch on the performance of our models on three datasets: JNLPBA, BC5-chem, and NCBI-disease (Figure 3). We use the same evaluation metric and fine-tuning approach for the NER task described in PubMedBERT (Gu et al., 2021).

Moreover, we experiment with other pre-training techniques. The above-described models are pre-trained by masking random tokens. Gu et al. (2021), and Cui et al. (2021) showed that pre-training with whole-word masking (WWM) can be more beneficial. Wettig et al. (2023) experimented with different percentages of masking rate and found out that for the BERT-base, 0.2 percent may be more optimal to prompt the model to learn better. Thus, we pre-train two more models using our method but with different techniques, one model using WWM and the second using WWM and 0.2% masking rate. Results are presented in Table 3.

	BM	BM	BM	BM
	from scratch	continued	averaged	contextualized
BC5-chem	92.51	**93.7**	92.93	93.52
BC5-disease	80.66	**83.92**	81.83	83.41
NCBI-disease	86.34	86.67	87.38	**88.44**
BC2GM	81.4	**83.78**	82.82	83.53
JNLPBA	76.14	77.59	75.89	**78.71**

Table 2: Performance (F1) of test models on five NER tasks.

7 Analysis

As shown in Table 2, **BM continued** shows superior performance on three datasets. The worst-performing model is **BM from scratch**. Such performance is expected due to the experimental set-up and pre-training for a shorter period. The performance of PubMedBERT (Table 6) suggests that pre-training from scratch with a domain-specific vocabulary can outperform continued pre-training but would require more extended training to reach this point. **BM averaged** performs better than the model pre-trained from scratch, demonstrating that

	contextualized	**contextualized**	**contextualized**
	+ MLM	**+ WWM**	**+ WWM**
	+ 0.15 mr	**+ 0.15 mr**	**+ 0.2 mr**
BC5-chem	93.52	93.77	**93.96**
BC5-disease	83.41	84.02	**84.28**
NCBI-disease	88.44	87.41	**88.85**
BC2GM	83.53	83.76	**83.89**
JNLPBA	78.71	78.35	**78.8**

Table 3: Performance (F1) of models pre-trained with contextualized embeddings and different pre-training techniques. The scores of the best-performing models are in bold, and underlined are the scores of the second-best-performing models.

averaging weights of subtokens to initialize weights for new tokens gives an initial boost for the model to learn. However, the second best-performing model is **BM contextualized** or **BIOptimus 0.1**. It outperforms **BM averaged** and **BM from scratch** on all datasets and **BM continued** on two datasets.

Figure 3 shows the models' performance on the dev set after each epoch. We can see that **BM continued** learns quickly and performs better on JNLPBA and BC5-chem datasets. However, the model shows the lowest learning curve on the NCBI-disease dataset, which the absence of specialized vocabulary can explain. Figure 3 also illustrates the importance of domain-specific vocabulary, and we can see that all the models with specialized vocabulary learn quickly and perform relatively well. Overall, introducing domain-specific vocabulary is beneficial. However, it is less efficient to pre-train a new model entirely from scratch, initializing all weights randomly. In the case of BIOptimus 0.1, adding new domain-specific tokens alongside the corresponding weights initialized in a way that leverages the contextualizing ability of transformers gives noticeable gains. Additionally, it is important to compare how biomedical LMs respond to pre-training on WWM against pre-training with masking only tokens/subtokens. Table 3 shows that it is not precisely the case that the WWM pre-training performs better across the board (Dai et al., 2022; Gu et al., 2021). Pre-training with masking only tokens performs better than WWM on two datasets out of five with a masking rate of 0.15 (underlined scores). One of the explanations is that individual subtokens often coincide with morphological components of the words, and learning the meanings of these components separately may be beneficial for the biomedical LM (Hofmann et al., 2020). Moreover, Table 3 demonstrates that increasing the masking rate contributes to a tapered performance enhancement.

8 Curriculum Learning

8.1 Motivation and Method

Analysis of our experiments motivates us to hypothesize that different training techniques like masking rate and masking strategies might help broaden the model's experience (Mitchell, 1997) and gain more diversified knowledge of textual input. In addition, during the experimental stage, we've observed that pre-training with specific techniques like WWM and increased masking rate slow down the training process if introduced right from the beginning due to increased task complexity. That brings us to the idea that introducing more complex tasks gradually, using CL's easy-to-hard strategy, assists in guiding the model's learning process more smoothly (Bengio et al., 2009). Using the results of our experiments, we implement a CL method for pre-training a biomedical LM. In Masked Language Modeling (MLM), the objective is to predict the masked token based on the surrounding context:

$$\begin{aligned} L_{MLM} &= -\sum_i \log P_\Theta(\omega_i|\tilde{\omega_i}) \\ &= -\sum_i \log \frac{\exp\left(E(\omega_i)^\top \tilde{h_i}\right)}{\sum_{j=1}^{|V|} \exp\left(E(\omega_i)^\top \tilde{h_i}\right)} \end{aligned} \tag{3}$$

It remains a challenging task to measure the difficulty of a task or a training sample in CL. It is primarily an issue in the case of pre-training masked language models. There are a few newly proposed approaches to tackle this challenge. Nagatsuka et al. (2021) pre-trains RoBERTa by increasing the block size of input text. Lee et al. (2022) proposed a curriculum based on masking easy-to-predict tokens first. We formulate our curriculum strategy from the perspective of the complexity of the prediction task. Predicting whole words is more complex than predicting just tokens which may be a part of the word, giving the LM more hints from the surrounding context (Gu et al., 2021; Cui et al., 2021). Increasing the masking rate makes prediction more challenging (Wettig et al., 2023) since less context is available for prediction.

We use a pre-trained model of the same architecture and number of parameters to measure the prediction task difficulty and evaluate its performance as MLM (Dudy and Bedrick, 2020; Liu et al., 2019). We use a corpus from a different domain to account for domain shift. In our case, we evaluate the performance of the BERT-base-uncased[4] on the RealNews dataset (Raffel et al., 2019). The evaluation results are presented in Figure 4. Based on this evaluation, we divide our curriculum into four phases (see Table 4).

We start with pre-training vanilla MLM (masking random tokens with a masking rate of 0.15). As we can see from Figure 4, this task is the easiest to handle for MLM. This task also ensures the model learns the subtokens that constitute complex biomedical terminology. At the next stage, we increase the complexity of the prediction task and

[4] https://huggingface.co/bert-base-uncased

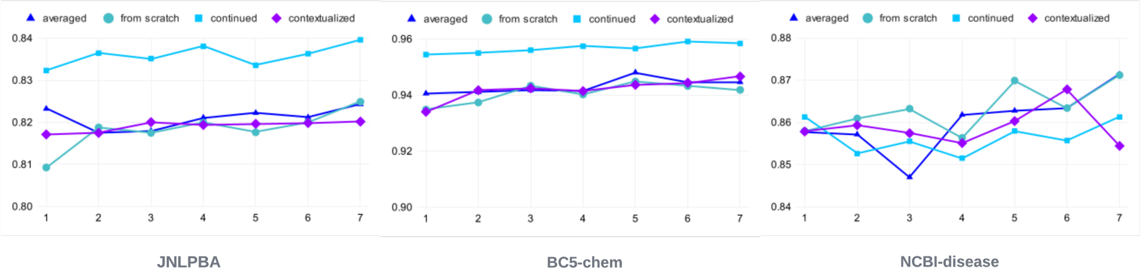


Figure 3: Effects of pre-training epochs on performance (F1) of four test models on the development set of JNLPBA, BC5-chem, and NCBI-disease datasets.

pre-train predicting the whole words to teach the model to combine separate subtokens into whole words. In the third phase, we raise the task complexity to one more level by increasing the masking rate to 0.2 in predicting the whole words. Wettig et al. (2023) observed that the corruption rule "80-10-10" hurts the performance for some downstream tasks and suggested using only [MASK] without corruption strategy for MLM pre-training. In our training experiments, we observe that prediction without corruption of tokens makes the prediction task even more complex and slows down the learning process; thus, we add it as an additional curriculum phase at the end.

We track and visualize contextualized word representations with different frequencies to observe how pre-training phases evolve over time and their impact on the performance and quality of pre-trained embeddings. Figure 5 shows that low-frequency words form a separate cluster from high-frequency words, which is more evident after the first stage. During pre-training, the gap decreases; however, only after the fourth phase, when the corruption strategy is removed, do the clusters join closer together. It is plausible that tokens' prediction with a corruption strategy plays a role in degenerating word embeddings, and it may explain why it hurts the performance in some downstream tasks (Wettig et al., 2023). We leave further experiments on this subject for future research.

8.2 Experimental Setting

Data. We use the same PubMed Corpus and NER datasets described in Section 5.

Implementation. The model's weights are initialized using our contextualized weight distillation approach, which helps speed up pre-training. We

v.	**Phases**	**Masking strategy**	**Masking rate**	**Corruption strategy**
0.1	phase 1	tokens	0.15	with corruption
0.2	phase 2	**WWM**	0.15	with corruption
0.3	phase 3	**WWM**	**0.2**	with corruption
0.4	phase 4	**WWM**	**0.2**	**no corruption**

Table 4: Stages of the CL method with increasing task complexity.

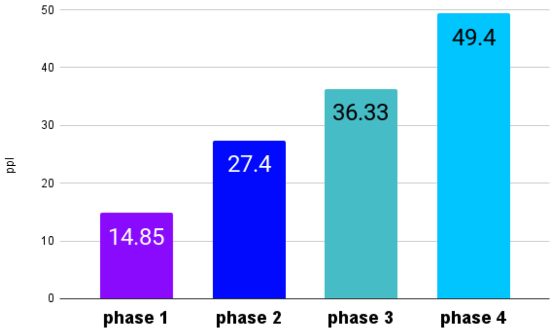


Figure 4: Evaluation of the performance of the BERT-base-uncased on the RealNews corpus (perplexity). We average the scores across text files randomly sampled from the corpus. We take ten samples, each of size of around 12M tokens.

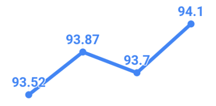
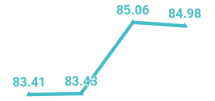
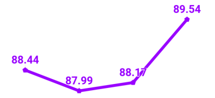
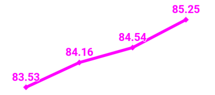
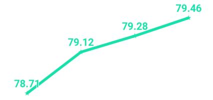
v.	BC5-chem	BC5-disease	NCBI-disease	BC2GM	JNLPBA
0.1	93.52	83.41	88.44	83.53	78.71
0.2	93.87	83.43	87.99	84.16	79.12
0.3	93.7	85.06	88.17	84.54	79.28
0.4	94.1	84.98	89.54	85.25	79.46
					

Table 5: Evaluation of each CL phase on NER datasets.

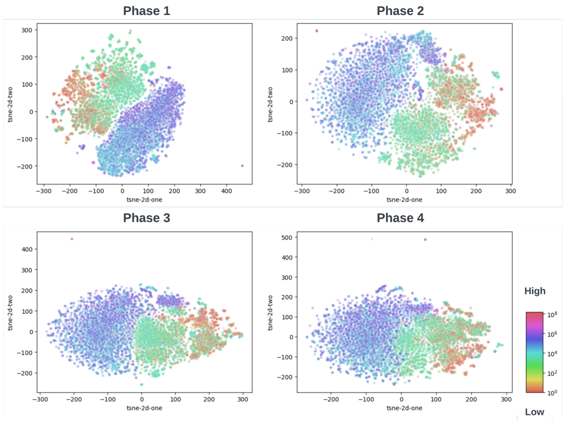


Figure 5: A t-SNE visualization of word embeddings during four pre-training phases with the CL method. "High" on the color bar corresponds to high frequencies, and "Low" to low frequencies.

pre-train with the same hyperparameters presented in Appendix A and increase the number of optimization steps for one epoch for each subsequent phase to account for the complexity of the prediction task. The initial learning rate is set at 1e-4, and when restarting the training at the next phase, we decrease the learning rate following the learning rate scheduler.

Models and Baselines. We track the progress of the CL method by evaluating models after each phase (see Table 5). Each phase's resulting model is considered an independent model and brings competitive results right from the first phase. Furthermore, we compare our models with BioBERT (Lee et al., 2019), PubMedBERT (Gu et al., 2021), and BioLinkBERT-base (Yasunaga et al., 2022) (Table 6).

8.3 Results and Discussion

Masking strategy (from phase 1 to phase 2) (see Table 5) gives noticeable gains in performance for BC5-chem, BC2GM, and JNLPBA datasets, there is only a slight improvement on the BC5-disease dataset, and the NCBI-disease dataset responded quite poorly to this transition. **Increasing the masking rate (from phase 2 to phase 3)** helps to advance performance on all datasets except BC5-chem, while BC5-disease shows a considerable gain from increasing the masking rate. **Removing the corruption strategy "80-10-10" (from phase 3 to phase 4)** is generally beneficial for all datasets, with considerable gains for BC5-chem and NCBI-disease and a slight drop for BC5-disease. BC2GM and JNLPBA datasets respond with stable improvement in all pre-training phases of the CL method. BC5-chem, BC5-disease, and NCBI-disease datasets exhibit more diverse responses to changes in curriculum phases.

	Bio-BERT	**PubMed-BERT**	**BioLink-BERT**	**BIOpti-mus 0.4**
BC5-chem	92.85	93.33	93.75	**94.1**
BC5-disease	84.7	85.62	**86.1**	84.98
NCBI-disease	89.13	87.82	88.18	**89.54**
BC2GM	83.82	84.52	84.9	**85.25**
JNLPBA	78.55	**80.06**	79.03	79.46

Table 6: Comparison of pre-trained language models on five NER datasets. The scores of the best-performing models are in bold, and underlined are the scores of the second-best-performing models. BIOptimus 0.4 sets a new state-of-the-art on BC5-chem, NCBI-disease, and BC2GM datasets.

8.4 Ablation Study

In this section, we conduct an ablation study to assess the effect of the proposed CL method. To measure the impact of pre-training with our CL method, we pre-train a model in a continued setting with the contextualized weight distillation method, using vanilla MLM (with a masking rate of 0.15, applying a corruption rule "80-10-10") but without CL. Essentially the model is BIOptimus 0.1 pre-trained for the same number of optimization steps as BIOptimus 0.4. The performance of this model is presented in Table 7 (“No CL”). Removing the CL method hurts downstream performance. The drop occurs with all NER datasets and is more apparent with BC5-disease, NCBI-disease and BC2GM datasets. This suggests that pre-training with the CL method helps boost biomedical LM’s performance on NER task.

	BC5-chem	BC5-disease	NCBI-disease	BC2GM	JNLPBA
BIOptimus 0.4	**94.1**	**84.98**	**89.54**	**85.25**	**79.46**
No CL	93.86	84.29	88.92	84.53	79.07

Table 7: Ablation study on the CL method.

9 Conclusion

This paper presented a new method to initialize tokens’ weight for new biomedical vocabulary when pre-training from the existing checkpoint (continued approach). We also compared this method of token weight initialization with other pre-training methods (see Table 2 and Table 3). This method showed considerable gains in speeding up the pre-training phase and improving performance on NER. Comparing pre-training techniques showed that WWM is not the best-performing approach for all NER tasks, and masking only tokens/subtokens shows competitive performance. Increasing the masking rate and removing the corruption strategy are generally beneficial techniques for pre-training biomedical LM. Finally, we introduced the CL method based on the task complexity to pre-train LMs. The “easy-to-hard” CL method introduces the biomedical LM to a broader scope of language experience, speeds up pre-training, and enhances performance on downstream tasks like NER. It is important to highlight that our model BIOptimus 0.4 achieves high performance with the pre-training time reduced by at least half, proving the pre-training approach’s efficiency.

Acknowledgment

We thank Sebastian Hurubaru and our anonymous reviewers for their valuable feedback. This work was supported by tnsr capital.

Limitations

(1) To be able to present a rigorous comparison and analysis of pre-training an optimal biomedical LM, we focused on running extensive evaluation on five NER tasks from the BLURB benchmark[5]. We do not evaluate the performance of our models on other downstream tasks in the framework of this paper and leave it for future work. (2) While we performed additional experiments to explain the reason for the performance drop on some tasks when implementing the corruption strategy "80-10-10," one plausible explanation is that it might increase the degree of anisotropy in the contextualized word representations. Separate work is needed to search for how the corruption strategy might cause a drop in performance on some downstream tasks. (3) Due to the expensive nature of pre-training experiments, we were not able to experiment with all possible combinations of pre-training methods and techniques. (4) Our models were pre-trained on the English language only.

Ethics Statement

While our research does not directly introduce any social or ethical bias nor amplify the bias in the data, we inherit a considerable amount of the underlying limitations of LMs. LM pre-training is computationally expensive and causes environmental damage. Our research focuses on overall computing resource efficiency and thus has a marginal environmental footprint.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML ’09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–

[5] https://microsoft.github.io/BLURB/

4781, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics*, 18.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Yong Dai, Linyang Li, Cong Zhou, Zhangyin Feng, Enbo Zhao, Xipeng Qiu, Piji Li, and Duyu Tang. 2022. “is whole word masking always better for Chinese BERT?”: Probing on Chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1–8, Dublin, Ireland. 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.

Rezarta Dogan, Robert Leaman, and Zhiyong lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47.

Shiran Dudy and Steven Bedrick. 2020. Are some words worth more than others? In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 131–142, Online. Association for Computational Linguistics.

Jeffrey L. Elman. 1993. Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71–99.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. *CoRR*, abs/1907.12009.

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*, 3(1):1–23.

Yuxian Gu, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. 2020. Train no evil: Selective masking for task-guided pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6966–6974, Online. Association for Computational Linguistics.

Junliang Guo, Xu Tan, Linli Xu, Tao Qin, Enhong Chen, and Tie-Yan Liu. 2020. Fine-tuning by curriculum learning for non-autoregressive neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7839–7846.

Xu Guo and Han Yu. 2022. On the domain adaptation and generalization of pretrained language models: A survey.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2020. DagoBERT: Generating derivational morphology with a pretrained language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3848–3861, Online. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission.

Yuyun Huang and Jinhua Du. 2019. Self-attention enhanced CNNs and collaborative curriculum learning for distantly supervised relation extraction. 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 389–398, Hong Kong, China. Association for Computational Linguistics.

Borna Jafarpour, Dawn Sepehr, and Nick Pogrebnyakov. 2021. Active curriculum learning. In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pages 40–45, Online. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, JNLPBA '04, page 70–75, USA. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Mingyu Lee, Jun-Hyung Park, Junho Kim, Kang-Min Kim, and SangKeun Lee. 2022. Efficient pre-training of masked language model via concept-based curriculum masking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7417–7427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn Mattingly, Thomas Wiegers, and Zhiyong lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Cao Liu, Shizhu He, Kang Liu, and Jun Zhao. 2018. Curriculum learning for natural answer generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4223–4229. International Joint Conferences on Artificial Intelligence Organization.

Jinglin Liu, Yi Ren, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Task-level curriculum learning for non-autoregressive neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3861–3867. International Joint Conferences on Artificial Intelligence Organization. Main track.

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.

Tom M Mitchell. 1997. *Machine learning*, volume 1. McGraw-hill New York.

Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. 2021. Pre-training a BERT with curriculum learning by increasing block-size of input text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 989–996, Held Online. INCOMA Ltd.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. Inexpensive domain adaptation of pretrained language models: Case studies on biomedical NER and covid-19 QA. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1482–1490, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Vin Sachidananda, Jason Kessler, and Yi-An Lai. 2021. Efficient domain adaptation of language models via adaptive tokenization. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 155–165, Virtual. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Larry Smith, Lorraine Tanabe, Rie Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig Struble, Richard Povinelli, Andreas Vlachos, William Baumgartner Jr, Lawrence Hunter, Bob Carpenter, and W. Wilbur. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9 Suppl 2:S2.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. Association for Computational Linguistics.

Yi Tay, Shuohang Wang, Luu Anh Tuan, Jie Fu, Minh C. Phan, Xingdi Yuan, Jinfeng Rao, Siu Cheung Hui, and Aston Zhang. 2019. Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2020. Learning a multi-domain curriculum for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7711–7723, Online. Association for Computational Linguistics.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling?

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. Association for Computational Linguistics.

Atsuki Yamaguchi, George Chrysostomou, Katerina Margatina, and Nikolaos Aletras. 2021. Frustratingly simple pretraining alternatives to masked language modeling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3116–3125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Sangwon Yu, Jongyoon Song, Heeseung Kim, Seongmin Lee, Woo-Jong Ryu, and Sungroh Yoon. 2022. Rare tokens degenerate all tokens: Improving neural text generation via adaptive gradient gating for rare token embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29–45, Dublin, Ireland. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021. Meta-curriculum learning for domain adaptation in neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14310–14318.

A Appendix

A.1 Datasets

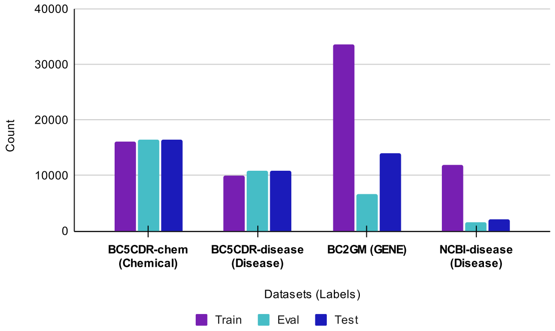


Figure 6: Label distribution of BC5CDR-chem, BC5CDR-disease, BC2GM, NCBI-disease datasets.

A.2 Computing Infrastructure

Computing Infrastructure	8 x A100 GPUs
Hyperparameter	**Assignment**
number of epochs	7-34
batch size	256
maximum learning rate	0.0005
learning rate optimizer	Adam
learning rate scheduler	None or Warmup linear
Weight decay	0.01
Warmup proportion	0.06
learning rate decay	linear

Table 8: Hyperparameters for pre-training biomedical LMs.

Computing Infrastructure	2 x NVIDIA RTX 3090 GPU
Hyperparameter	**Assignment**
number of epochs	5-11
batch size	4, 8, 16
learning rate	1e-5, 2e-5
dropout	0.1

Table 9: Hyperparameters for fine-tuning on NER datasets.