On Cross-Domain Pre-Trained Language Models for Clinical Text Mining: How Do They Perform on Data-Constrained Fine-Tuning?

Samuel Belkadi[1*], **Lifeng Han**[1*], **Yuping Wu**[1], **Valerio Antonini**[2], and **Goran Nenadic**[1]
[1] Department of Computer Science, The University of Manchester, UK
[2] INSIGHT Centre, Dublin City University, Ireland
* *co-first authors*
samuel.belkadi@student.manchester.ac.uk
lifeng.han, yuping.wu, g.nenadic@manchester.ac.uk

Abstract

Fine-tuning Large Language Models (LLMs) pre-trained from general or related domain data to a specific domain and task using a limited amount of resources available in the new task has been a popular practice in NLP fields. In this work, we re-visit this assumption, and carry out investigation in clinical NLP, specifically named-entity recognition on Drugs and their related Attributes. We compare Transformer models that are learned from scratch to fine-tuning BERT-based LLMs including BERT-base, BioBERT, and ClinicalBERT. We also investigate the comparison of such models and their extended models with a CRF layer for continuous learning. We use n2c2-2018 shared task data for model development and evaluations. The experimental outcomes show that 1) the CRF layer makes a difference for all neural models; 2) on BIO-strict span level evaluation using macro-average F1, while the fine-tuned LLMs achieved scores 0.83+, the TransformerCRF model learned from scratch achieved 0.78+ demonstrating comparable performances but using much less cost, e.g. 39.80% less training parameters; 3) on BIO-strict span level evaluation using weighted-average F1, the score gaps are even smaller (97.59%, 97.44%, 96.84%) for models (ClinicalBERT-CRF, BERT-CRF, TransformerCRF). 3) efficient training using down-sampling for better data-distribution (SamBD) further reduced the data for model learning but producing similar outcomes around 0.02 points lower than the full set model training.

Our models including source codes will be hosted at https://github.com/HECTA-UoM/TransformerCRF.

1 Introduction

Pre-trained language models (PLMs) with fine tuning have been one of the dominant methods in current natural language processing (NLP) tasks, including text mining (Zhang et al., 2021), named entity recognition (Dernoncourt et al., 2017), reading comprehension (Sun et al., 2020), machine translation (Vaswani et al., 2017; Devlin et al., 2019), and summarisation (Gokhan et al., 2021; Wu et al., 2022), etc. They also demonstrated strong performances in comparison to conventional methods. Domain applications of PLMs have spanned in a much wider variety including financial, legal, and biomedical texts, in addition to traditional news and social media domains. For instance, experimental work on BioBERT (Lee et al., 2019) and BioMedBERT (Chakraborty et al., 2020) using BERT-based (Devlin et al., 2019) learning structure trained on biomedical data has demonstrated high evaluation scores. Fine-tuned SciFive, BioGPT, and BART models produced reasonable experimental outputs on biomedical abstract simplification tasks (Li et al., 2023).

There is an ongoing investigation in this field regarding whether PLMs with fine-tuning would perform better than training from scratch on certain tasks (Gu et al., 2021). Researchers tend to take it for granted that this assumption can receive a positive answer when the task under study is having a very low amount of resource and the PLM can help with extra knowledge learned from large amounts of available out-of-domain or related-domain data which is usually from general or mixed domain. One of the key questions here is, to make the deployed task benefit from mixed domain pre-training, what size of data is a low-resource enough scenario? No research has reported such statistics. In this paper, focusing on clinical domain text mining, we investigate the above-mentioned assumption, i.e. whether PLMs can perform better than models learned from scratch using small amounts of available data in a constrained setting and to what degrees.

In comparison to other domains, clinical text mining (CTM) is still a relatively new task for PLM applications, as CTM is well known for data-scarce issue due to small amount of human-annotated cor-

pora and privacy concerns. In this work, we take PLMs from the *general* domain BERT, *biomedical* domain BioBERT, and *clinical* domain ClinicalBERT, examining how well they perform on clinical information extraction task on drugs and drug-related attributes using n2c2-2018 shared task data via adaptation and fine-tuning. We compare these models with light-weighted models learned from scratch. We also investigate if a CRF layer makes a difference across the deployed models.

This paper is organised as follows: Section 2 further explains more details of related work, Section 3 introduces the methodologies for our investigation including experimental designs, Section 4 includes the data-preprocessing and experimental setups, Section 5 details the evaluation outputs, Section 6 gives discussions on ablation evaluations with more insight, and finally, Section 7 concludes this paper with future work plan. We also give more detailed experimental analysis and relevant findings in Appendix.

2 Related Work

Most related work to ours is the development on pre-trained language model applications into biomedical and clinical domains in recent years. These include BioBERT (Lee et al., 2019) which was the first work discussing the benefit of pre-training BERT based model using biomedical data from "scratch". In comparison to traditional methods that use pre-trained BERT on general domain data (e.g. Wikipidia or BooksCorpus (Zhu et al., 2015)) with fine-tuning as a second step using domain specific corpus, BioBERT shows that training BERT model from scratch using PubMed abstract and PubMed Central (PMC) full-text articles for 23 days on 8 NVIDIA V100 GPUs produced higher score on NER and RE tasks in biomedical domain.

However, *in fact*, BioBERT is pre-trained using mixed domain data in a way of continuous learning based on pre-training, instead of purely using biomedical corpus for model learning. This is because it was initialised with standard BERT model then *continuously-trained* using biomedical data. To re-examine the advantages of purely trained model from biomedical in-domain data, PubMedBERT (Gu et al., 2021) used the same data from PubMed as in BioBERT to avoid mix-domain influences. The rational behind this is that the word distributions from different domains are represented differently in the corresponding vocabularies. Furthermore, PubMedBERT created a new benchmark data set BLURB which covers more tasks than BioBERT. The biomedical terms covered in PubMedBERT include disease, drug, gene, organ, and cell.

PubMedBERT and BioBERT both focused on biomedical specific domain knowledge, leaving other closely related domains such as clinical domain into their future exploration. Subsequently, Alsentzer et al. (2019) demonstrated that clinicalBERT trained from generic clinical text and discharge summaries can perform better on medical language inference task (i2b2-2010 and 2012) and de-identification task (i2b2-2006 and 2014). Similarly, Huang et al. (2019) investigated that clinicalBERT trained on clinical notes and do better performance on predicting hospital readmission after fine-tuning on this task.

In our work, focusing on clinical domain, we will use the n2c2-2018 shared task corpus which is from electric health record (EHR) letters that are often semi-structured with the heading part specifying drug names, patient names, doses, relations, etc., and the rest of the letter with free text describing the diagnoses and treatment procedure and effects. We aim at examine how the pre-trained LLMs in biomedical and clinical domain perform on n2c2-2018 task EHR data using domain adaptation and task-specific fine-tuning.

Regarding Transformer based models for text mining, Wu et al. (2021) applied Transformer with adaptation layer for information and communication technology (ICT) domain patent entity extraction. Al-Qurishi and Souissi (2021) applied CRF layer on top of BERT model to carry out Arabic NER on mixed domain data, such as news and magazines. Yan et al. (2019) demonstrated that Transformer encoder has a superior performance on traditional NER tasks in comparison to BiLSTMs. Other related work also includes (Zhang and Wang, 2019; Gan et al., 2021; Zheng et al., 2021; Wang and Su, 2022) which applied Transformer and CRF combined models for spoken language understanding, Chinese NER, and power meter NER, and forest disease text.

3 Methodology and Experimental Designs

Figure 1 displays the design of our investigation, where it includes the pre-trained LLMs BERT (Devlin et al., 2019), BioBERT (Lee et al., 2019), and ClinicalBERT (Alsentzer et al., 2019), in addition

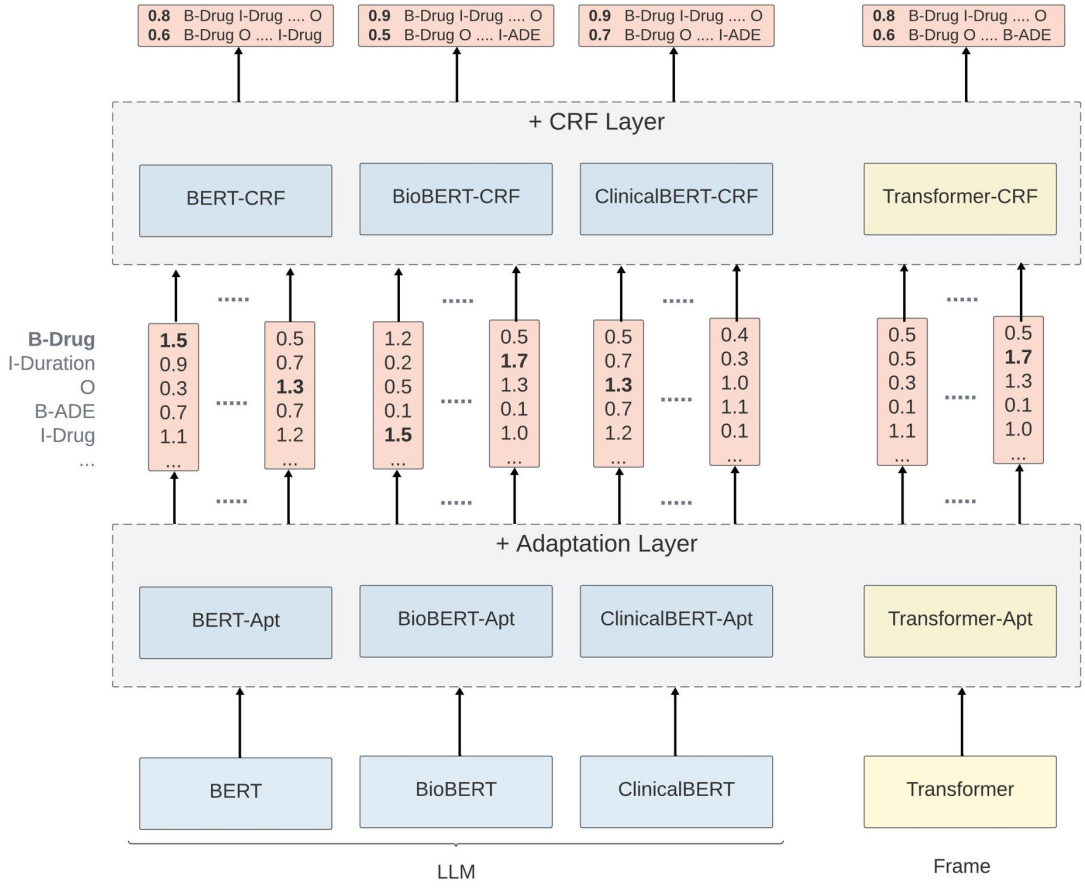

Figure 1: Model Designs for Investigation

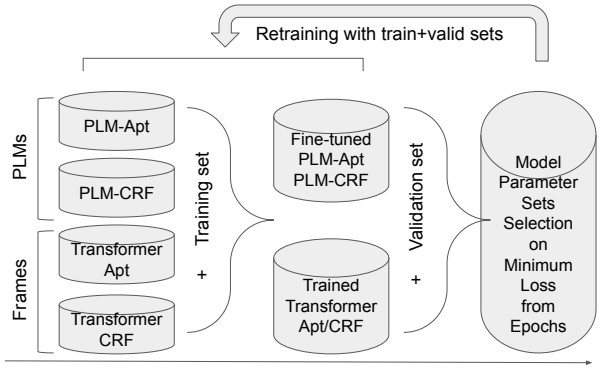

Figure 2: Model Optimisation and Training

to scratch-learned Transformer (encoder only) from (Vaswani et al., 2017). For the encoder-only Transformer, we applied "distilbert-base-cased" structure without using the learned word embeddings.

The first step is to adapt these models for clinical named entity recognition (NER) task on Drug names and their related attributes, which results in BERT-Apt, BioBERT-Apt, ClinicalBERT-Apt, and Transformer-Apt. These adaptation models can carry out our designed NER tasks themselves. The adaptation layer is to predict probability distribution over all labels for each token. The label with the highest predicted probability is chosen as the predicted label for the token.

Then, the second step is to add a Conditional Random Field (CRF) (Lafferty et al., 2001) layer on top of the adaptation models forming BERT-CRF, BioBERT-CRF, ClinicalBERT-CRF, and Transformer-CRF models. These CRF layers are expected to continue model learning for label classification and optimisation. Different from the adaptation models, instead of predicting labels of taken in a sequence independently, the extra CRF layer takes the neighbouring tokens and their corresponding labels into account when predicting potential labels for the current token under-study. The design of Transformer-Apt and Transformer-

CRF is to examine the performance of the learned models trained from scratch using the limited data, in comparisons to LLMs.

4 Data Pre-processing and Experimental Setups

We introduce the n2c2-2018 corpus we deploy for model development and evaluations, the model optimisation strategies, efficient training, and evaluation metrics here.

4.1 Corpus and Model Setting

Regarding the data set, we use standard n2c2-2018 shared task data from Track-2 (Henry et al., 2020): Adverse Drug Events and Medication Extraction in Electric Health Records (EHRs) [1]. The World Health Organisation (WHO) [2] defines ADE as "an injury resulting from medical intervention related to a drug", and the Patient Safety Network (PSNet) uses another definition "harm experienced by a patient as a result of exposure to a medication" [3]. The aim of this task is to investigate if "NLP systems can automatically discover drug-to-adverse event relations in clinical narratives." There are three sub-tasks under this track including Concepts, Relations, and End-to-End. Among these, the first task is to identify drug names, dosages, duration, and other entities, the second task is to identify relations of drugs with adverse drug events (ADEs) and other entities given gold standard entities, and the third task is the same as the second one but on entities that are predicted by systems. In total, this track offered 505 annotated files on discharge summaries from the Medical Information Mart for Intensive Care III (MIMIC-III) clinical care database (Johnson et al., 2016). The annotation work was carried out by four physician assistant students and three nurses. The presence of drugs and ADEs is reflected by entity tags and the corresponding attributes. The 505 files were divided into 303 for model training and 202 for model testing purposes.

We split 10 percent of original training data, i.e. 30 files into validation data, and the rest 273 files as the training (on scratch model) or fine-tuning (on PLMs) data for our designed investigation. The testing data is the original n2c2-2018 track-2 test set of 202 files. *Regarding sentence counts, there are about 41,497 sentences in the training set, 4,536 in the validation set, and 30,614 in the test set.*

There are 18 (9x2) special labels plus O as an indicator of normal text, which covers 9 categories of different events and medications including ADE, Dosage, Drug, Duration, Form, Frequency, Reason, Route, and Strength. We use the BIO labelling format from CoNLL2003 shared task (Tjong Kim Sang and De Meulder, 2003) with B for the beginning and I for an inner token of events/medications, and O for normal context.

4.2 Model Optimisation and Training

Figure 2 shows the designed model training procedures. Firstly, we use the sub-training set for continuous learning on LLMs and for the training of TransformerApt and TransformerCRF. The learned model will be fed with our validation set for parameter optimisation and selection based on their F1 scores and the minimum loss values from epochs. Finally, the selected parameter sets for each model will be used for their re-training using the original n2c2-2018 training data consisting of the sub-training and validation sets. For the foundation models, we used "bert-base-cased", "biobert-base-cased-v1.2", "ClinicalBERT" for LLMs, and "distilbert-base-cased" for Transformer frame.

The number of trainable model parameters and their comparison ratios are listed in Table 1. It shows that TransformerCRF reduces the trainable parameters from BERT-CRF by 39.80 percent, ClinicalBERT-CRF has 24.39 percent more parameters than BERT-CRF, and BioBERT-CRF has the same parameter size to BERT-CRF. These ratios kept the same for Adaptation (Apt) models.

4.3 Efficient Training

Because there are around 90 percent O labels that are not special tokens for the model, we design an efficient training step for this task by removing a certain amount of plain samples (down-sampling) for better data distribution, which we name as **SamBD**. Specifically, we down-sample the text with only the label "O", to approach a similar distribution with the CONLL2003 (Tjong Kim Sang and De Meulder, 2003) data, which has around 20 percent non-special label sentences. This also reduces the computational cost for model training and validation. The label distributions on original divided train-valid sets and the optimised sets using SamBD are displayed in Figure 3. We keep the test

[1] https://portal.dbmi.hms.harvard.edu/projects/n2c2-2018-t2/

[2] https://www.who.int/

[3] https://psnet.ahrq.gov/primer/medication-errors-and-adverse-drug-events

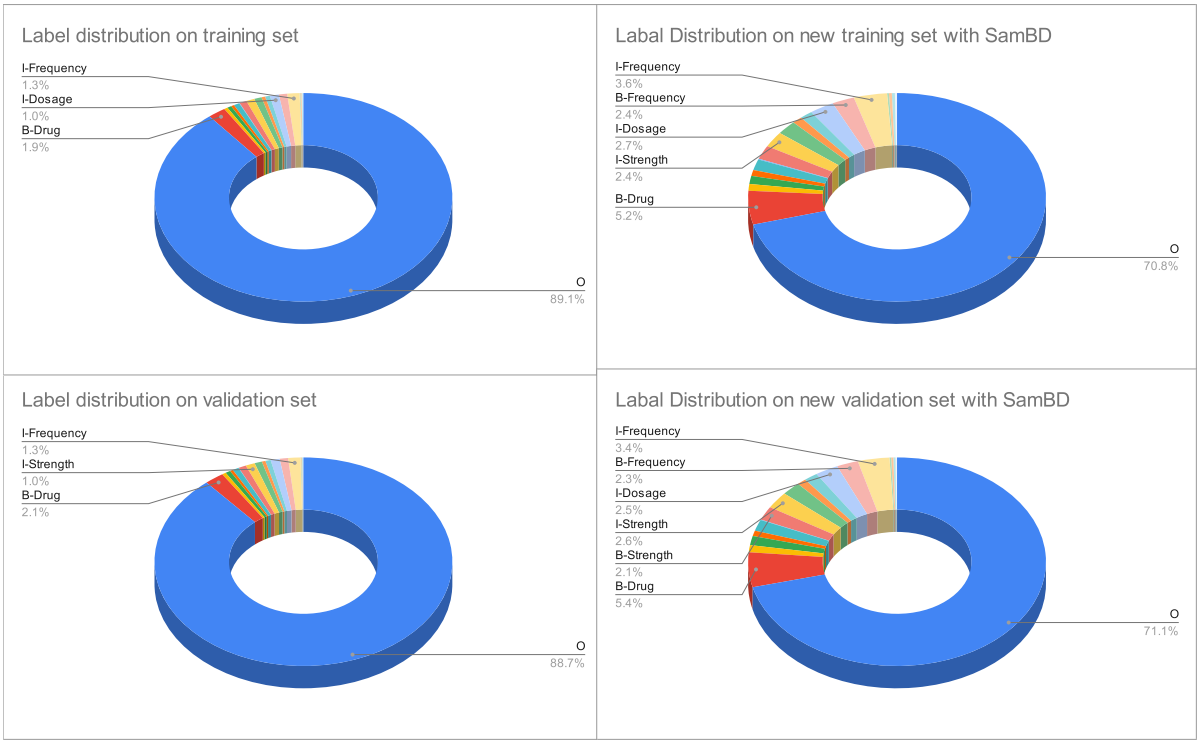

Figure 3: Label distribution on training and validation set before (left) and after (right) SamBD.

CRF Model	Parameters	%to Bert-crf
BERT-crf	108,325,282	1
BioBERT-crf	108,325,282	1
ClinicalBERT-crf	134,749,090	24.39%↑
TransformerCRF	65,205,922	39.80%↓
Apt Model	**Parameters**	**%to BERTApt**
BERT-Apt	108,324,883	1
BioBERT-Apt	108,324,883	1
ClinicalBERT-Apt	134,748,691	24.39%↑
Transformer-Apt	65,205,523	39.80%↓

Table 1: Number of trainable parameters of each model and comparison ratios

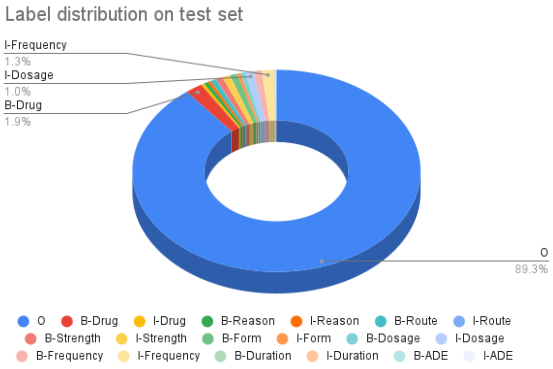

Figure 4: Label distribution on test set.

set as it is and the label distributions are shown in Figure 4.

In the original training set from our setting, most special labels have less than 1% appearance, except for I-Frequency (1.3%), I-Dosage (1.0%), and B-Drug (1.9%); however, in the efficient training set with SamBD, the special labels have improved frequencies. e.g. I-Strength label has 2.4% statistically in Figure 3. Similar stories apply to the original validation set on our setting and the efficient validation set with SamBD. With a full data set, the overall evaluation score might be high but it can be a biased model toward all O labels; with efficient training by reducing O labels, the model is expected to learn balanced knowledge and increase the learning on special NER-labelled tokens. We will report the model learning and evaluation with the full data set, i.e. with all "O" labels kept, and the efficient training setting.

4.4 Evaluation Metrics

We calculate the macro- and micro-averaged precision, recall, and F1 scores, as well as the weighted scores. For multi-class labelling tasks, macro-averaging assigns the same weight to each label

category, however, micro-averaging assigns the same weight to each sample token. In this situation, the precision, recall, and F1 scores are the same for the micro-average which is also often called as the model accuracy score. For weighted-average scores, the contribution of each class label is weighted by its size among all samples. In situations when the labels are very unevenly distributed, these scores tell different stories.

Because each event and medication can have multiple tokens, for example, 20 (B-strength) mg (I-strength) per (I-strength) day (I-strength) having four tokens. BIO-level event and medication calculation only assigns a correct score if the full span of the multiple tokens within this event/medication are recognised correctly, while token-level calculation takes each single correctly labelled token into account. We will report the evaluation scores on both of these two categories.

5 Experimental Outcomes

We present both strict BIO-level (span-level) evaluations on the models and the confusion matrix to look into detailed error analytics, and then we present the token-level (attribute-level) evaluation, and the outcomes from efficient training.

5.1 BIO-Strict Evaluations using Full Set

The model evaluation scores using macro average, weighted average, and accuracy (micro-F1) on n2c2-2018 official testing set when they are trained on the full training data are displayed in Figure 5 and Figure 7 including the adaptation models and CRF models. We also draw a bar chart for macro-averaged scores for better illustration in Figure 6.

For macro average evaluations, Clincial-BERT-Apt has the highest Precision score 85.30% (BioBERT-CRF the 2nd 84.26%), BERT-CRF has the highest Recall score 83.43% (BioBERT-CRF the 2nd 83.30%) and the highest F1 score 83.72% (ClinicalBERT-CRF the 2nd 83.71%). In general ClinicalBERT-CRF and BERT-CRF have very close F1 scores. While TransformerApt and TransformerCRF have Precision scores above 0.80, i.e. 81.99% and 82.51% each, their Recall scores fall under 0.80. However, their performances are still comparable to the fine-tuning of pre-trained LLMs, producing 78.42% and 78.74% F1 scores only using the 303 training letters learned from scratch.

For weighted and micro average evaluation shown in Figure 7, ClinicalBERT-CRF wins the

Macro Average Evaluation on Test Set

Models	P	R	F1
BERT-Apt	0.835418197	0.830983584	0.832294977
BERT-crf	0.841795668	**0.834369933**	**0.837159335**
BioBERT-Apt	0.045421434	0.052631578	0.048761414
BioBERT-crf	0.842558512	0.833045927	0.835699285
clinicalBERT-Apt	**0.852971384**	0.809422989	0.825485325
clinicalBERT-crf	0.849750868	0.829104303	0.837110897
TransformerApt	0.819853660	0.760761994	0.784214658
TransformerCRF	0.825062735	0.761722835	0.787392525

Figure 5: BIO-strict macro-models-test. bold case is the highest score and underline is the second highest in the same column.

highest scores on all columns including weighted precision, recall, F1 and accuracy. ClinicalBERT-Apt has the 2nd highest weighted Recall and model accuracy, while BioBERT-CRF has the 2nd highest weighted precision and F1 scores. Overall, the weighted and micro evaluation scores do not tell much difference between each model, except for BioBERT-Apt. Most of these models produced score ranges between (96.60%, 97.55%) for weighted F1, and between (96.75%, 97.59%) for accuracy which include the scratch-learned TransformerApt and TransformerCRF. Combining the results from both Figure 5 and Figure 7, we understand that the BioBERT-Apt model mostly predicted the correct "O" label but failed the special labels on Drugs and Drug-related attributes.

Looking back at these overall evaluation scores, there are interesting findings: 1) Micro/Weighted Evaluations make less sense. In situations when special labels have really low ratios, the challenge for the model prediction is to label these special labels correctly. However, if the model just labels all tokens into the majority label, e.g. "O", it will still get really high micro- and weighted-average evaluation scores since this evaluation is based on the accuracy of each sample and the sample weight of the labels, without carefully considering the model bias and label distribution. For instance, BioBERT-Apt only has Macro P/R/F: 0.0454, 0.0526, and 0.0488 but the Micro (0.8630) and Weighted P/R/F1 (0.7448, 0.8630, 0.7995) are really high, resulting in a false high-accuracy model on this task. 2) Adding CRF layers has an impact. While in most cases, adding the CRF layer just improves a little bit of evaluation score, for the BioBERT model, it made a dramatic change with the CRF layer vs without. We are not sure if this is caused by the BioBERT structure setting or what

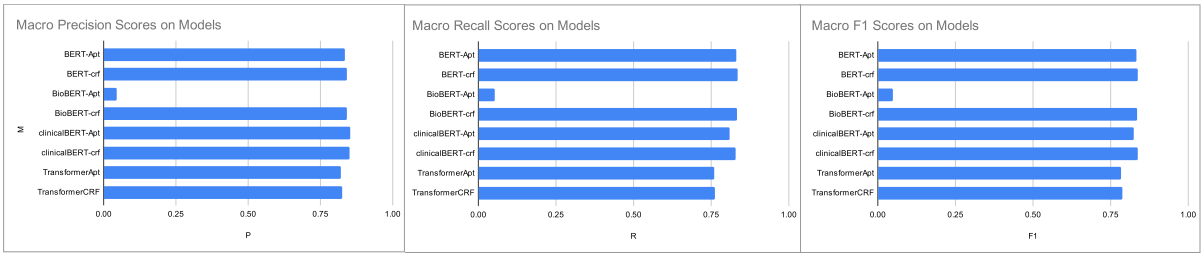

Figure 6: Macro macro-models-eval-bar

	Weighted Average Evaluation on Test Set			Acc/Micro-F1
Models	P	R	F1	
BERT-Apt	0.974074404	0.974150656	0.973976257	0.974150656
BERT-crf	0.974120811	0.974364681	0.974126955	0.974364681
BioBERT-Apt	0.744781531	0.863007260	0.799547642	0.863007260
BioBERT-crf	0.974881087	0.974948057	0.974677520	0.974948057
clinicalBERT-Apt	0.974355176	0.975095221	0.974305239	0.975095221
clinicalBERT-crf	**0.975518239**	**0.975910036**	**0.975497544**	**0.975910036**
TransformerApt	0.965489360	0.967515877	0.966092929	0.967515877
TransformerCRF	0.966459706	0.968439243	0.966982874	0.968439243

Figure 7: BIO-strict weighted and micro test. bold case is the highest score and underline is the second highest in the same column

other reasons. This is to be further investigated.

5.2 BIO Confusion Matrix of PLM-CRFs

To understand better the model performances regarding different labels, we list the detailed evaluation scores on different labels and the confusion matrix using the BIO-strict evaluation setting. Because across all models the CRF layers win the adaptation models, we present here the results from ClinicalBERT-CRF representing the best in Figure 8 and Figure 9, BERT-CRF representing the base model of LLMs in Figure 10 and Figure 11, and TransformerCRF representing the scratch-learned variants in Figure 12 and Figure 13.

Looking into the strict evaluation of BIO categories on all labels from the two models BERT-CRF and ClinicalBERT-CRF in Figure 10 and Figure 8, we get the following observations: 1). Both these two models have lower performances on ADE and Reason labels, with the F1 scores below 0.70. 2). In comparison to the BERT-CRF model, the ClinicalBERT-CRF model wins two F1 scores on B-Reason (0.63 vs 0.62) and I-Reason (0.61, 0.58), one tie one B-ADE (both 0.49), and lost on I-ADE (0.51, 0.52). Looking at Figure 12, TransformerCRF has the same phenomena. In addition, the B-Duration label also has an under 0.70 F1 score (0.682).

Looking at the confusion matrix in Figure 9, 11,

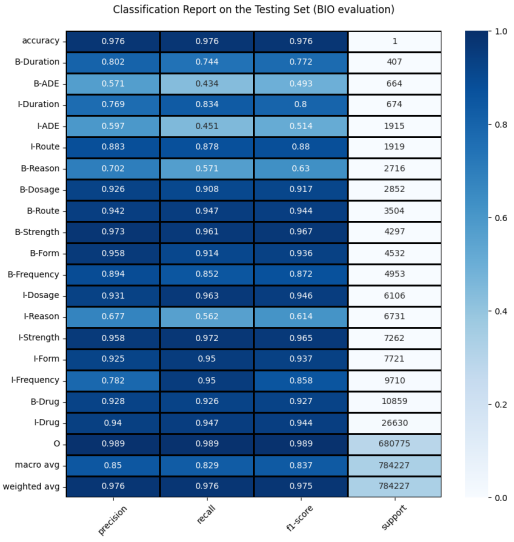

Figure 8: BIO-strict clinical-BERT-crf classification report on all labels

and 13, we can also observe some similar patterns.

- Most of the wrong labels predicted by models out of context token "O" are I-Drug, I-Reason, and I-Frequency. These numbers are (1157, 1450, 1931) by Clinical-BERT-CRF, (1484, 1931, 1886) by BERT-CRF, and (1245, 1393, 1967) by TransformerCRF on the top row of the confusion matrix.

- Between different special labels, the most common errors are I-Drug to B-Drug and B-Drug to I-Drug. They are (208, 334) by Clinical-BERT-CRF, (163, 395) by BERT-CRF, and (282, 284) by TransformerCRF.

- Between different attributes, the most common errors are I-ADE to I-Reason, I-Reason to I-ADE. They are (178, 95) by Clinical-BERT-CRF, (150, 147) by BERT-CRF, and (178, 164) by TransformerCRF. In this cate-

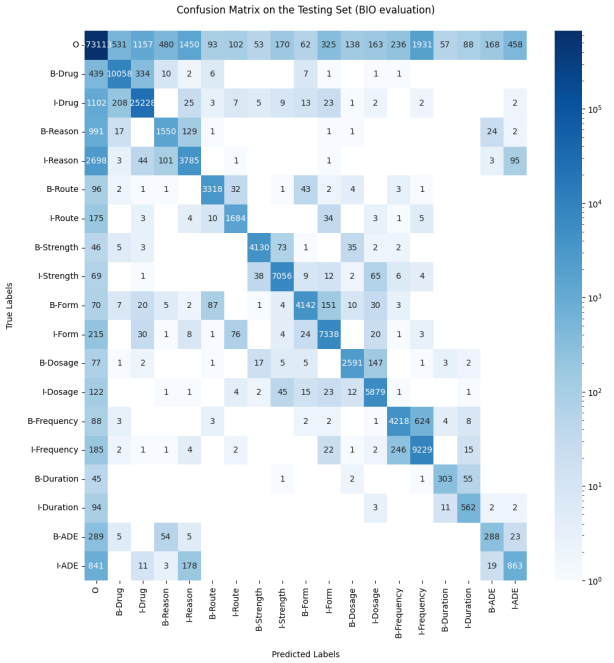

Figure 9: BIO-strict clinical-BERTcrf confusion matrix report on all labels

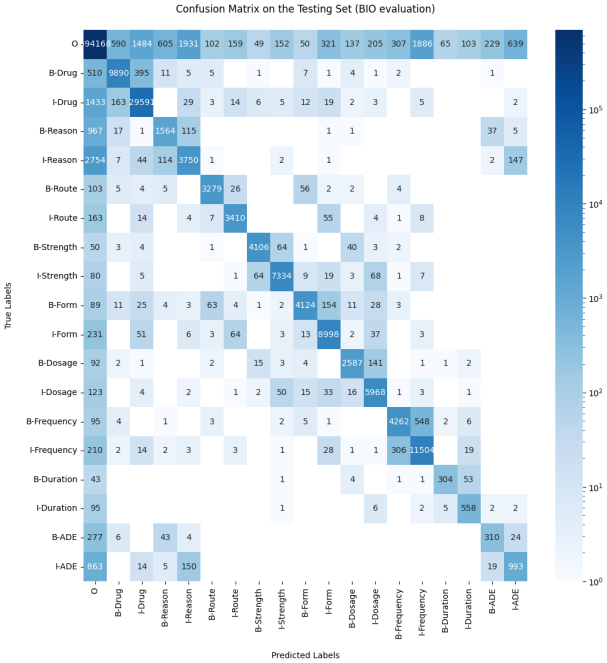

Figure 11: BIO-strict BERTcrf confusion matrix report on all labels

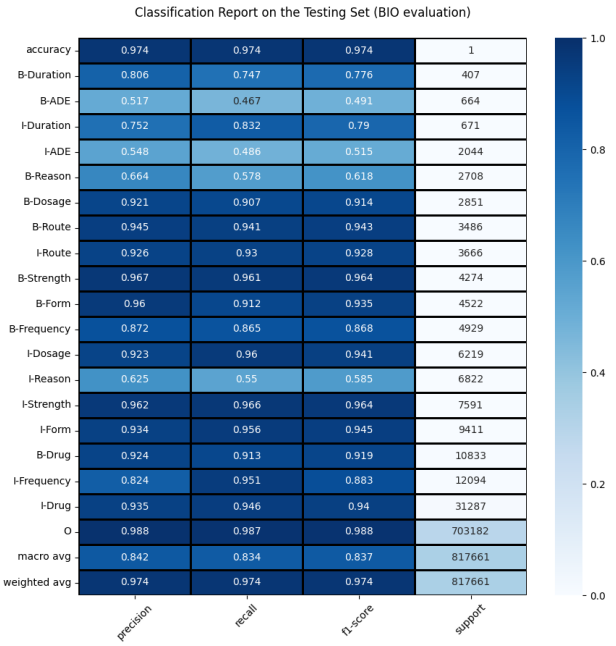

Figure 10: BIO-strict BERT-crf classification report on all labels

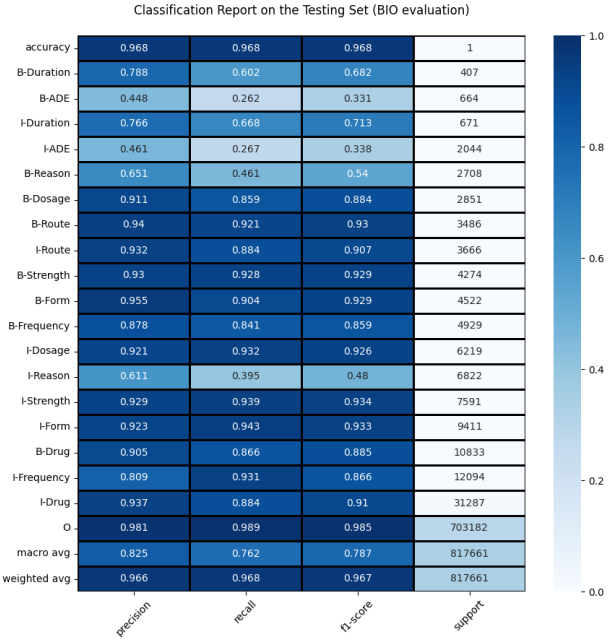

Figure 12: BIO-strict TransformerCRF classification report on all labels

gory, the following up pairs are B-Form to B-Route, I-Form to I-Route, I-Strength to I-Dosage, B-ADE to B-Reason, I-Dosage to I-Strength for Clinical-BERT-CRF. These kinds of errors are understandable, for instance, *ADE* and *Reason* are connected, and *Dosage* and *Strength* are also very *related* to each other.

5.3 Token-Level Evaluation using Full set

As we mentioned in the evaluation setting, if we do not care about whether the token is the beginning (B) or inner part of the Drugs and related events, i.e. the detailed granularity, we can calculate the token-level (or Attribute-level) performances of

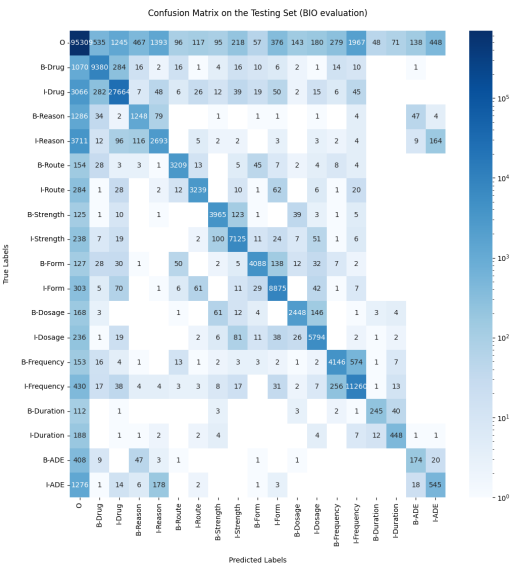

Figure 13: BIO-strict TransformerCRF confusion matrix report on all labels

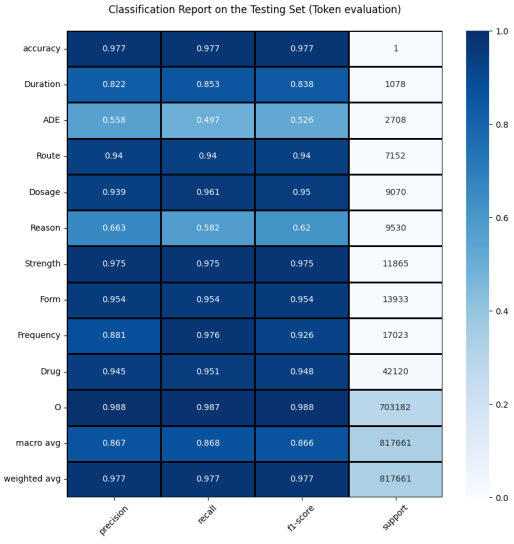

Figure 14: Token BERT-crf classification report on all labels

these models. In many practical situations, this is a good enough setting. Compared to the token-level, the BIO-level evaluation can be treated as a strict "span" level prediction. These two kinds of categories have also been popular in the Multiword Expression (MWE) prediction tasks organised by the MWE section of the ACL-SIGLEX group (Maldonado et al., 2017). The evaluation scores of these three models at the token-level are displayed in Figure 15, 14, and 16. In comparison to BIO-span level evaluation scores shown in Figure 5, we can see that the token-level evaluation has higher scores, increasing by (83.7 to 86.9), (83.72 to 86.6), and (78.74 to 81.9) for the three models Clinical-BERT-CRF, BERT-CRF, and Transformer-CRF on macro-averaged F1.

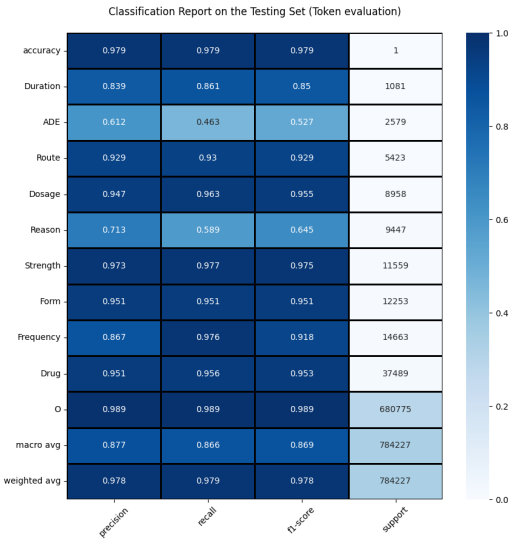

Figure 15: Token clinical-BERT-crf classification report on all labels

5.4 Token-Level Eval with Efficient Training

To investigate the downsampling experiments we carried out for better label distribution on the training and validation data set and to reduce the computational cost, we present the evaluation outcomes of these three models using SamBD efficient training.

Figure 17, 18, and 19 show the token-level (attribute-level) evaluations of efficient training outputs from models BERT-CRF, ClinicalBERT-CRF, and TransformerCRF. In comparison to the full data set training outcomes in Figure 14, 15, and 16, we can see that the efficient training with SamBD has very comparable macro average evaluation scores (0.829 to 0.866), (0.842 to 0.869), and (0.79 to 0.819) but using less computational cost. The weighted average F1 scores are even closer (0.966 to 0.977), (0.971 to 0.978), and (0.959 to 97) respectively on these three models, i.e. the fine-tuned PLMs BERT-CRF and Clinical-BERT-CRF, and scratch-learned TransformerCRF. In experiments with much larger data sets, efficient training will demonstrate a bigger potential for saving more model training and selection costs.

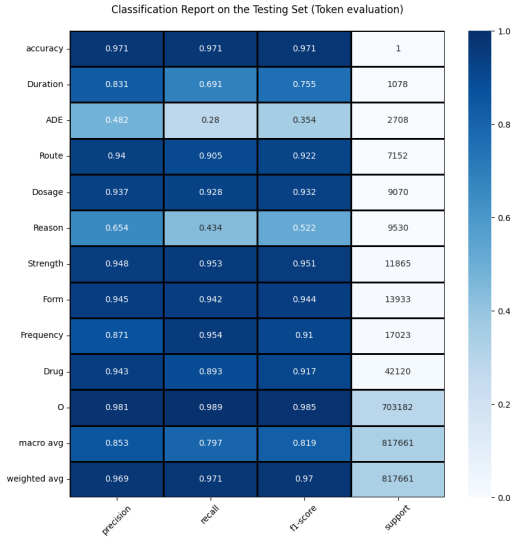

Figure 16: Token TransformerCRF classification report on all labels

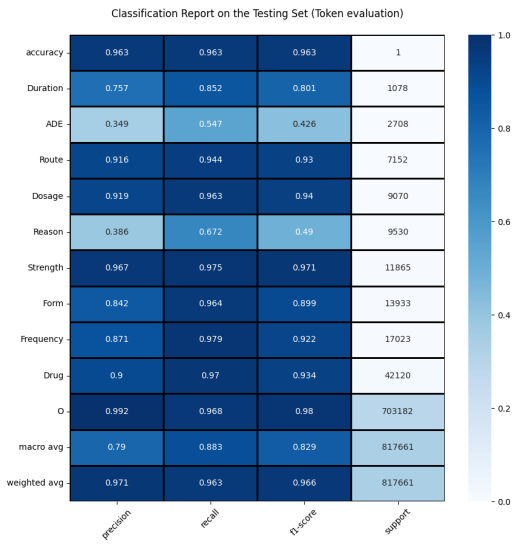

Figure 17: Token BERT-crf with SamBD classification report on all labels

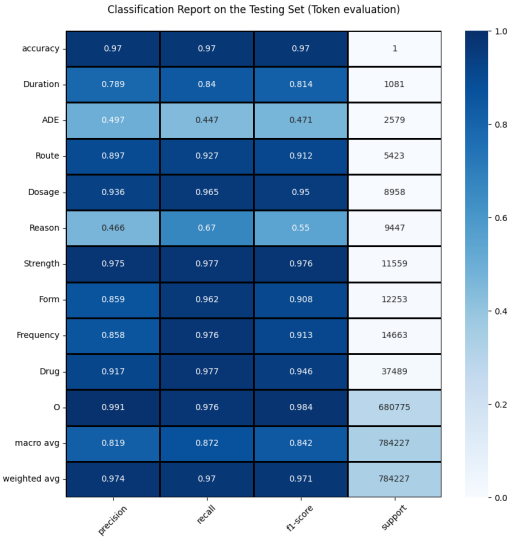

Figure 18: Token Clinical-BERT-crf with SamBD classification report on all labels

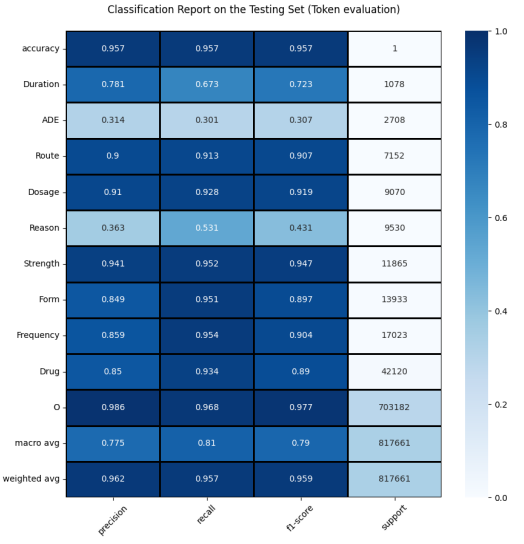

Figure 19: Token TransformerCRF with SamBD classification report on all labels

5.5 Revisiting n2c2-2018 Official Submissions

Looking back to the official results from n2c2-2018 challenge, the F1 scores achieved by the tops systems were all owing to the external knowledge-based features the teams used, such as pre-trained word embeddings upon the entire MIMIC III data and part-of-speech (POS) tags among others (Henry et al., 2020). Such external resources either make the model computationally costly to run or make it not easy to reproduce by others. In addition, most of the submitted models in n2c2-2018 applied BiLSTM-CRF models, instead, we examined the latest BERT-like deep learning structures with CRF layers using both pre-trained LLMs and scratch-learned TransformerCRF model. The deployed TransformerCRF model is kind of the only *constrained models* that only uses the official training data of 303 annotated letters for training. We call for attention from researchers on constrained

clinical text mining on how to improve model performances when the available resource is low or restricted. There have been many constrained NLP tasks in general domains to tackle scientific challenges, held by different shared task organisers, such as data-constrained machine translation e.g. by Han et al. (2021), following this route. In addition, the scratch-learned TransformerCRF model and its effecient-traning model using SamBD have both produced higher weighted average F1 scores on the token-level (Lenient F1) 0.97 and 0.959 in Figure 16 and Figure 19 than the highest F1 0.9418 from the official shared task submissions (Henry et al., 2020).

6 Ablation Evaluations on Token-level

To further look into the system outputs with more details, we carry out some ablation analysis on the evaluation settings. We report these three ablation evaluations on the token-level settings in Table 2.

- Firstly, since there are over-populated "O" labels in the data set and these labels are not the key concerns of the model, we report the evaluation scores removing the "O" label.
- Secondly, the "ADE" and "Reason" labels are two outliers across all models with much lower performances than other classes of special labels. To have an intuitive view of how the models perform without these two labels, we report the scores after removing these two labels.
- To further look at how the models perform on the exact 7 key labels without "ADE" and "Reason", we remove these two labels together with "O".
- Thirdly, because "ADE" and "Reason" are related labels semantically, we want to know if the models can predict correctly for the tokens that belong to any of the two if we merge these two labels, e.g. treating ADE label as Reason label. In this case, we keep all other labels as they are including the "O" label.

From this table, we can see that 1) by removing the context "O" label, the macro averaged F1 and weighted F1 scores are (0.853, 0.856, 0.801) and (0.91, 0.91, 0.877) from these three models. 2) if we remove the "Reason" and "ADE" outliers, the models have much higher macro- and weighted F1 scores (0.94, 0.94, 0.915) and (0.983, 0.984, 0.977) respectively. 3) by removing three labels "O", "ADE", and "Reason", the model performances on the other 7 key labels are around four points (0.04) lower than the last category on weighted average but very similar on macro average. 4) by merging the two labels "ADE" and "Reason", the weighted average scores are very close the the second category when we removed these two labels. 5) overall, the scratch-learned TransformerCRF model is about five points (0.05) below the other two models on the 9 key labels, but only three points (0.03) below on the other 7 key labels without (or merging) "ADE" and "Reason" using the macro average F1.

7 Conclusion and Future Work

In this work, we investigated the question "*On Cross-Domain Pre-Trained Language Models for Clinical Text Mining: How Do They Perform on Data-Constrained Fine-Tuning?*" We used the limited amount of available human-annotated clinical EHR data 303 letters, around 45K sentences from n2c2 shared task. We applied BERT, BioBERT, ClinicalBERT as examples of pre-trained LLMs from general, biomedical, and clinical domains. We compared their fine-tuning with adaptation layers and CRF layers versus scratch-learned TransformerApt and TransformerCRF models. On the full set training, using macro-averaged F1, ClinicalBERT-CRF and BERT-CRF produced similar highest scores (0.8371, 0.8372), while BioBERT-CRF has 0.8357, and scratch-learned TransformerCRF has comparable score 0.7839. However, TransformerApt and TransformerCRF have 39.80% fewer parameters in comparison to BERT-Apt and BERT-CRF for the model training, meanwhile, ClinicalBERT-Apt/CRF have 24.39% more parameters than BERT-Apt/CRF.

Using the efficient training SamBD mechanism, we further reduced the training cost by removing context plain labels and increasing the key label distributions over the training set, without much sacrifice of performance loss.

From the ablation evaluations, we further understand that while scratch-learned TransformerCRF performs around 0.05 points lower on Macro-F1 score than the other two fine-tuning models using LLMs on 9 key labels, the performance gaps narrowed to 0.03 on the 7 key labels without "ADE" and "Reason". This is very promising since the

Metrics	BERT-CRF			ClinicalBERT-CRF			TransformerCRF (scratch)		
	P	R	F	P	R	F	P	R	F
Removing label 'O': only using 9 key labels									
micro avg	0.908	**0.915**	0. 912	**0.913**	0.913	**0.913**	0.906	0.863	0.884
macro avg	0.853	**0.854**	0.853	**0.865**	0.852	**0.856**	0.839	0.776	0.801
weighted avg	0.905	**0.915**	**0.91**	**0.909**	0.913	**0.91**	0.896	0.863	0.877
Removing labels 'Reason' and 'ADE': keeping 'O' and other 7 key labels									
micro avg	0.982	0.983	0.982	0.982	0.985	0.984	0.975	0.98	0.977
macro avg	0.931	0.95	0.94	0.931	0.95	0.94	0.925	0.907	0.915
weighted avg	0.982	0.983	0.983	0.983	0.985	0.984	0.975	0.98	0.977
Removing labels 'O', 'Reason' and 'ADE': only keeping other 7 key labels									
micro avg	0.936	0.957	0.947	0.936	0.959	0.947	0.929	0. 919	0.924
macro avg	0.922	0.944	0.933	0.922	0.945	0.933	0.916	0.895	0.904
weighted avg	0.937	0.957	0.947	0.937	0.959	0.948	0.93	0.919	0.924
Merging two labels 'ADE' as 'Reason': keeping all others									
macro avg	0.902	0.91	0.906	0.908	0.911	0.908	0.897	0.855	0.872
weighted avg	0.977	0.978	0.977	0.979	0.979	0.979	0.97	0.972	0.971

Table 2: Ablation-eval on o-ade-reason: Token-level with full set learning

scratch-learned model costs much less from a computational perspective. This can be further improved by data augmentation on low-frequency labels.

From this findings, we ask a new question: "***On fine-tuning LLMs for Clinical NLP - do we really need to do this way***"? Or we shall consider training the domain and task-specific clinical text mining models from scratch in certain situations in a cost-performance balanced view?

In future work, we are planning to carry out data augmentation using synthetic data generation models, especially on sentences with key labels as well as applying the SamBD on this. Another way for data augmentation is the graph-based semi-supervised learning method by graph-construction on both labelled and unlabelled data followed with label propagation as used by (Han et al., 2015).

Author Contributions

S developed the adaptation (Apt) and CRF layer on top of investigated LLMs and Scratch-learned Transformer, developed SamBD, and carried out the experimental work. L designed the project and drafted the manuscript. Y contributed to the earlier version of TransformerCRF project with code-development and experimental work (https://arxiv.org/abs/2210.12770v2). V attended the co-development of the earlier version of this project. G co-supervised the work.

Acknowledgements

We thank Viktor Schlegel for helping debugging and Hao Li for useful discussion on our earlier experiments. We thank the anonymous reviewers for their valuable comments on the earlier version of our manuscript. The following projects have partially supported this research: UKRI/EPSRC grant EP/V047949/1 "Integrating hospital outpatient letters into the healthcare data space". We thank Google's Colab service at our model developing stage and the CSF3 server support from The University of Manchester. We acknowledge the usage of the following packages: SKlearn, BERT, BioBERT, and ClinicalBERT.

References

Muhammad Saleh Al-Qurishi and Riad Souissi. 2021. Arabic named entity recognition using transformer-based-crf model. In *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 262–271.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco

Mosconi. 2020. BioMedBERT: A pre-trained biomedical language model for QA and IR. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 669–679, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

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.

Yong Gan, Ruisen Yang, Chenfang Zhang, and Dongwei Jia. 2021. Chinese named entity recognition based on bert-transformer-bilstm-crf model. In *2021 7th International Symposium on System and Software Reliability (ISSSR)*, pages 109–118.

Tuba Gokhan, Phillip Smith, and Mark Lee. 2021. Extractive financial narrative summarisation using SentenceBERT based clustering. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 94–98, Lancaster, United Kingdom. Association for Computational Linguistics.

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

Aaron Li-Feng Han, Xiaodong Zeng, Derek F. Wong, and Lidia S. Chao. 2015. Chinese named entity recognition with graph-based semi-supervised learning model. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 15–20, Beijing, China. Association for Computational Linguistics.

Lifeng Han, Gareth Jones, Alan Smeaton, and Paolo Bolzoni. 2021. Chinese character decomposition for neural MT with multi-word expressions. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 336–344, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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*, 36(4):1234–1240.

Zihao Li, Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Matthew Shardlow, and Goran Nenadic. 2023. Large language models and control mechanisms improve text readability of biomedical abstracts.

Alfredo Maldonado, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel, and Qun Liu. 2017. Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 114–120, Valencia, Spain. Association for Computational Linguistics.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging Chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Conference on Neural Information Processing System*, pages 6000–6010.

Qi Wang and Xiyou Su. 2022. Research on named entity recognition methods in chinese forest disease texts. *Applied Sciences*, 12(8):3885.

Hengqin Wu, Geoffrey Qiping Shen, Xue Lin, Minglei Li, and Clyde Zhengdao Li. 2021. A transformer-based deep learning model for recognizing communication-oriented entities from patents

of ict in construction. *Automation in Construction*, 125:103608.

Yuping Wu, Ching-Hsun Tseng, Jiayu Shang, Shengzhong Mao, Goran Nenadic, and Xiao-Jun Zeng. 2022. Edu-level extractive summarization with varying summary lengths.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2019. TENER: adapting transformer encoder for named entity recognition. *CoRR*, abs/1911.04474.

Linhao Zhang and Houfeng Wang. 2019. Using bidirectional transformer-crf for spoken language understanding. In *Natural Language Processing and Chinese Computing*, pages 130–141, Cham. Springer International Publishing.

Taolin Zhang, Zerui Cai, Chengyu Wang, Minghui Qiu, Bite Yang, and Xiaofeng He. 2021. SMedBERT: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining. 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 5882–5893, Online. Association for Computational Linguistics.

Kaihong Zheng, Jingfeng Yang, Lukun Zeng, Qihang Gong, Sheng Li, and Shangli Zhou. 2021. Constructing bi-order-transformer-crf with neural cosine similarity function for power metering entity recognition. *IEEE Access*, 9:133491–133499.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

Appendices

We list the BIO-strict span-level evaluation reports and confusion matrices from BERT-Apt, BioBERT-CRF, ClinicalBERT-Apt, and TransformerApt.

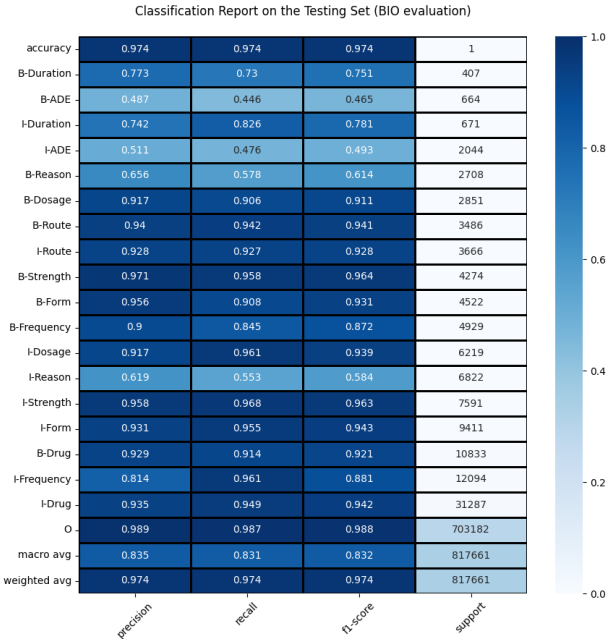

Figure 20: BIO-strict BERTapt classification report on all labels

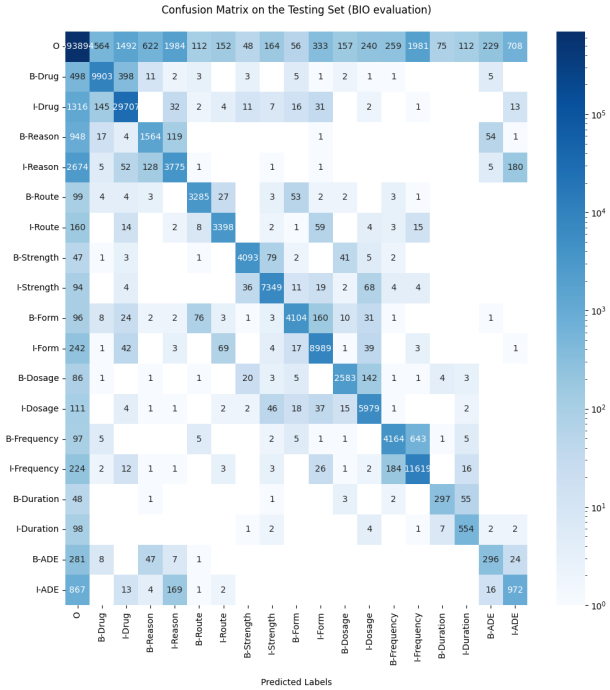

Figure 21: BIO-strict BERTapt confusion matrix report on all labels

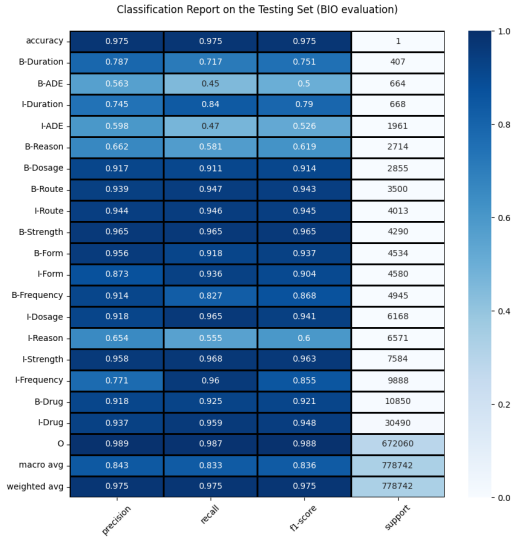

Figure 22: BIO-strict BioBERT-CRF classification report on all labels

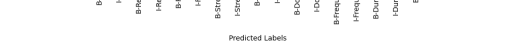

Figure 23: BIO-strict BioBERT-CRF confusion matrix report on all labels

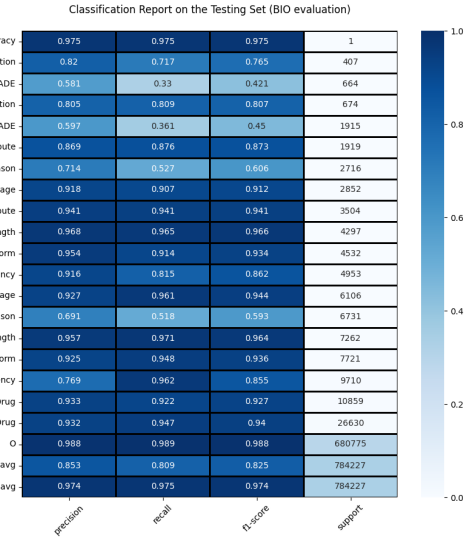

Figure 24: BIO-strict ClinicalBERT-Apt classification report on all labels

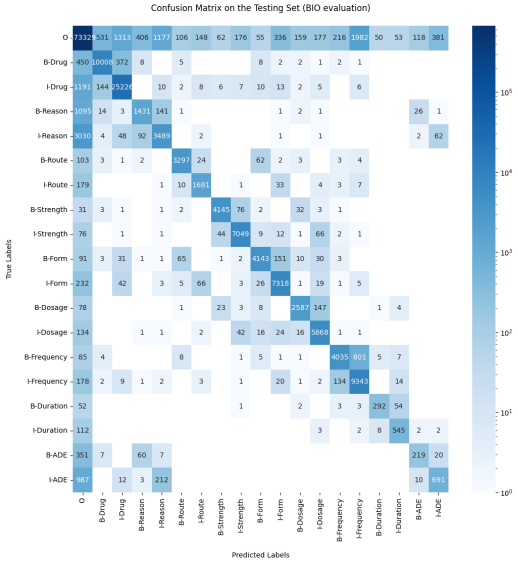

Figure 25: BIO-strict ClinicalBERT-Apt confusion matrix report on all labels

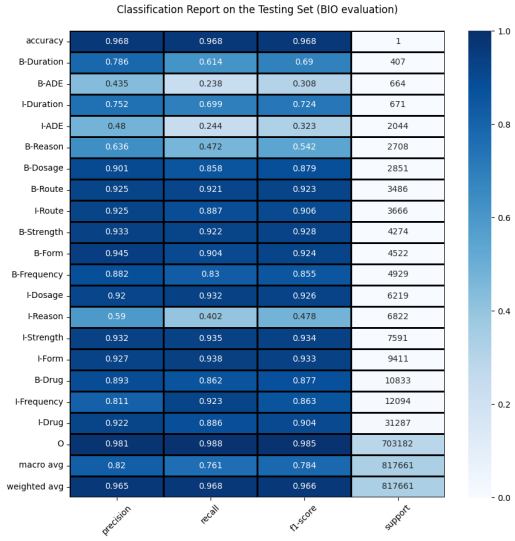

Figure 26: BIO-strict Transformer-Apt classification report on all labels

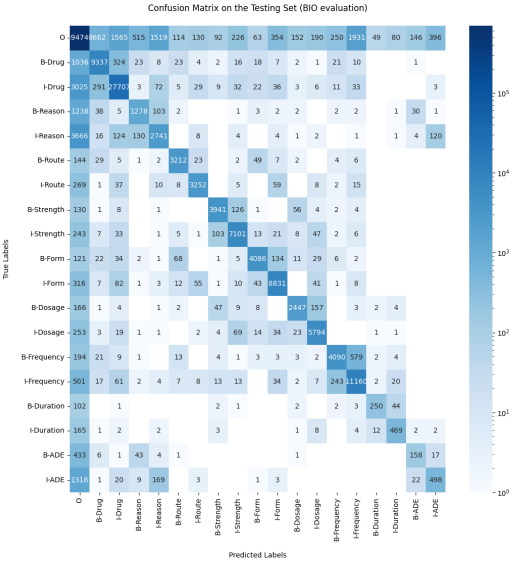

Figure 27: BIO-strict Transformer-Apt confusion matrix report on all labels