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

Yuping Wu[1*], **Lifeng Han**[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
valerio.antonini3@mail.dcu.ie
lifeng.han, yuping.wu, g.nenadic@manchester.ac.uk

Abstract

Pre-trained language models (PLMs) have been deployed in many natural language processing (NLP) tasks and in various domains. Language model pre-training from general or mixed domain rich data plus fine-tuning using small amounts of available data in a low resource domain demonstrated beneficial results by researchers. In this work, we question this statement and verify if BERT-based PLMs from the biomedical domain can perform well in clinical text mining tasks via fine-tuning. We test the state-of-the-art models, i.e. Bioformer which is pre-trained on a large amount of biomedical data from PubMed corpus. We use a historical n2c2 clinical NLP challenge dataset for fine-tuning its task-adapted version (BioformerApt), and show that their performances are actually very low. We also present our own end-to-end model, TransformerCRF, which is developed using Transformer and conditional random fields (CRFs) as encoder and decoder. We further create a new variation model by adding a CRF layer on top of PLM Bioformer (BioformerCRF). We investigate the performances of TransformerCRF on clinical text mining tasks by training from scratch using a limited amount of data, as well as the model BioformerCRF. Experimental evaluation shows that, in a *constrained setting*, all tested models are *far from ideal* regarding extreme low-frequency special token recognition, even though they can achieve relatively higher accuracy on overall text tagging. Our models including source codes will be hosted at https://github.com/poethan/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, recent work on BioBERT [1] (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. Furthermore, very recently, Bioformer[2] developed by WGLab[3] claimed even better performance in comparison to BioBERT (Lee et al., 2019) and PubMedBERT (Gu et al., 2021) using the same training data.

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. However, mixed domain knowledge can also have negative effect on model learning for domain specific tasks, especially by introducing ambiguity. For instance, the word "dude" can be normally understood and learned as a person in a general domain corpus, but in the veterinary/medical domain it is an acronym for "defecating, urinating, drinking and eating" [4].

[1] https://github.com/dmis-lab/biobert
[2] available at HuggingFace https://huggingface.co/bioformers
[3] https://github.com/WGLab and https://wglab.org
[4] *ref.* https://www.allacronyms.com/DUDE/

One of the key questions here is, to make the deployed task to be benefited from mixed domain pre-training, what size of data is low-resource enough scenario? No research has reported such statistics. In this paper, focusing on clinical domain text mining, we investigate into 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.

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 corpora and privacy concerns. In this work, we take PLMs from the *biomedical domain* and examine how well they perform on (kind of closely) related clinical data from international shared tasks via adaptation and fine tuning.

We report evaluation scores from the experiments. The findings have surprised us because of the huge performance drops when we use the alternative development and test datasets in such a related domain. Furthermore, we introduce our own end-to-end models using a Transformer and conditional random fields (CRFs) structure, and using the same constrained very small amount of data, our models demonstrate higher Accuracy evaluation scores in comparison to the currently openly available state of the art models even without pre-training. Finally, we design a new variation model by adding a CRF layer on top of PLM Bioformer and demonstrate how interestingly this new model works. We also report that all of theses evaluated models are still far from our expectations in practice on CTM tasks, especially on the recognition of special tokens.

This paper is organised as follows: Section 2 further explains more details of related work, Section 3 introduces the designed models for investigation including the selection of the current state of the art models and our own models, Section 4 presents the experimental evaluations on designed models and their comparison, and Section 5 summarises this paper with conclusion and a 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. 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 LM in biomedical domain performs on n2c2-shared task EHR data using adaptation and domain fine-tuning.

To facilitate this research, we adopt off-the-shelf open-source Bioformer [5] from HuggingFace (Wolf et al., 2019) which is one of the latest PLMs in biomedical domain and reported comparable and even better scores than BioBERT and PubMed-

medical

[5] https://huggingface.co/bioformers/

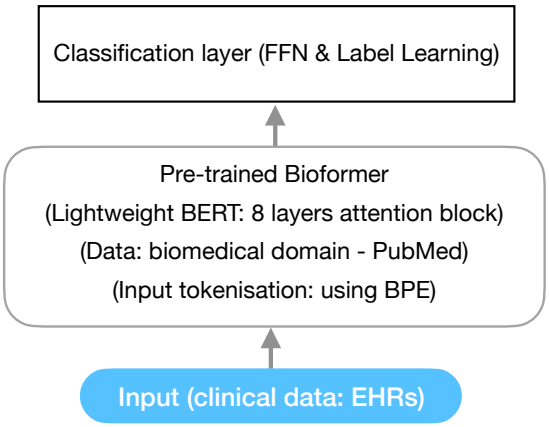

Figure 1: BioformerApt structure: Bioformer adapted to text mining task.

BERT, as well as 3 times faster than BERT base forming a lightweight BERT model for biomedical NLP.

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) which applied Transformer and CRF model for spoken language understanding and Chinese NER tasks. However, none of above mentioned work has ever applied TransformerCRF model into clinical domain text mining which topic is the focus of this paper.

3 Cross-Domain Model Selection and Designs

We explain more details of Bioformer that we use as baseline PLM and its adaptation (BioformerApt) to clinical text mining task, then present our own model TransformerCRF and a variation BioformerCRF.

3.1 Adapting Bioformer: BioformerApt

We adopt Bioformer cased version 1.0 [6] which is trained using 33 million PubMed abstract and 1 million PMC full-text. PubMed is a search engine maintained by the US National Library Medicine, which offers access to bibliographical database of scientific articles and abstracts on life sciences and biomedical domains, since 1996 [7].

The parameters reported from Bioformer include 8 layers of Transformer (Vaswani et al., 2017) blocks, 8 self-attention heads, and total number of parameters 42,820,610. Bioformer was pre-trained for 2 million steps, using batch size 256, maximum input sequence length 512, with vocabulary size 32,768 (2^{15}).

However since Bioformer was pre-trained for masked language modelling (MLM) and next sentence prediction (NSP) tasks, we need to build another classification layer on top of it to be able to apply it for text mining task under investigation, i.e. "drug event and medication extraction". The adapted Bioformer structure (**BioformerApt**) is displayed in Figure 1. Pre-trained Bioformer is applied to encode and generate representation vectors for each token from our input sentences. Pre-trained Bioformer as many other BERT based models used Byte Pair Encoding (BPE) (Sennrich et al., 2016) to address low frequency token issues. To fit the vocabulary of Bioformer, BPE was applied to input sequences before feeding to Bioformer encoders, and the first component from BPE of a token is taken as the token representation for label prediction, if the token under-study is split into multiple components by BPE. Using token representation vectors as input, we add an extra classification layer 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.

3.2 Our Model: TransformerCRF

To re-examine the advantage of training domain-specific data from scratch to avoid cross-domain knowledge influences, as well as to combine current popular neural structure, i.e. the Transformers (Vaswani et al., 2017), with conventional state of the art statistical model, i.e. graph-based CRF (Lafferty et al., 2001), we introduce our own open-source developed model structure called **TransformerCRF** as in Figure 2 which was presented in HealTAC 2022 in a set of research projects (Han et al., 2022; Madrid et al., 2022). In Han et al. (2022), we presented both character and word level

[6] https://huggingface.co/bioformers/bioformer-cased-v1.0

[7] official website https://pubmed.ncbi.nlm.nih.gov

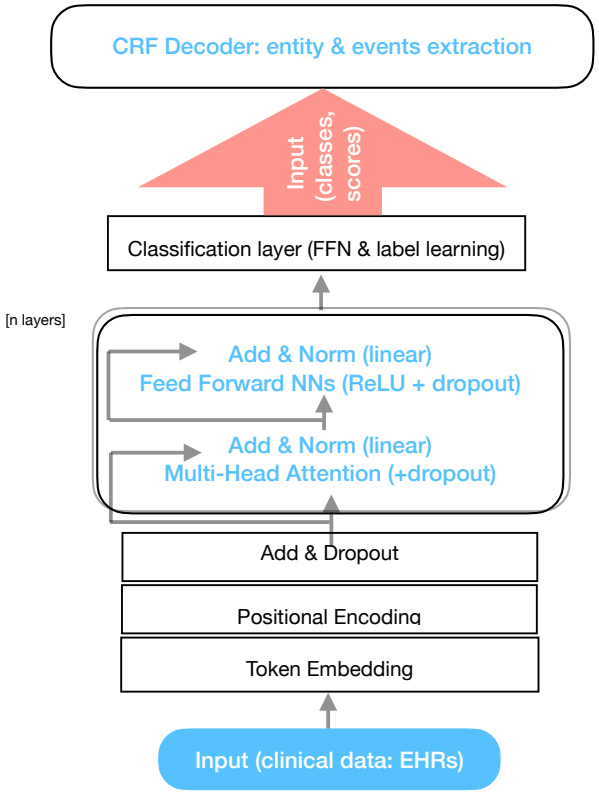

Figure 2: TransformerCRF: Transformer (Vaswani et al., 2017) encoder with CRF (Lafferty et al., 2001) decoder structure.

combined TransformerCRF structure. However, in this work, we only use word / token based Transformer as encoder and CRF as a continuous learning layer (kind of decoder) for CTM. In this model, we do not use any pre-trained knowledge, instead we will train the model from fully randomised parameter values (Epoch 1) using this structure, i.e. in a data constrained setting.

3.3 BioformerCRF: New Variation

To further testify if pre-trained LM using Bioformer can be improved by continuously learning via graph-based CRF models, we design a new model variation **BioformerCRF**.

BioformerCRF has the same structure of TransformerCRF, instead just replacing the input of pre-CRF block, i.e. the "input (classes, score)" in Figure 2) with BioformerApt output which is learned from the already pre-trained embeddings. Different from BioformerApt, instead of predicting labels of tokens in a sequence independently, the extra CRF learning layer takes the neighbouring tokens and their corresponding labels into account when predicting potential labels for a current token understudy.

4 Experimental Investigations

We introduce the n2c2-2018 corpus we deploy for model evaluations and the experimental findings.

4.1 Corpus and Model Setting

Regarding 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) [8]. The World Health Organisation (WHO) [9] 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" [10]. 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 purpose.

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

There are overall 36 (9x4) special labels plus O as indicator of normal text, which covers 9 categories of different events and medications including ADE, Dosage, Drug, Duration, Form, Fre-

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

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

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

quency, Reason, Route, and Strength. It uses BIOES labelling format from CoNLL2003 shared task (Tjong Kim Sang and De Meulder, 2003) with B for beginning, I for inner, O for normal text, E for ending, and S for single token event/medication.

The parameter settings for the designed model are listed as below: batch size 4, dropout rate 0.1, patience 20, maximum number of epochs 100, optimiser Adam learning rate 0.005, Transformer number of heads 8 and number of layers 6[11], Transformer d-model 512, dff 2048, Max sequence 128, token embedding dimension 600, tokenizer Spacy, gpu 1, number of cpu threads 8. Dependency requirements for installation include Tensorflow (Abadi et al., 2015) version 2.9.1, Tensorflow-addons version 0.17.0, Transformers (Vaswani et al., 2017), and Stanford-corenlp-4.0.0 (Manning et al., 2014).

4.2 Evaluation Results

We report the experimental results of overall tagging accuracy (Acc), precision (Pre), recall (Rec), and F1 score at event and medication level. We also report the number of correctly recognised event and medication using the parameter Corr. These overall scores are located in the heading part of the evaluation tables.

Each event and medication can have multiple tokens, for example, 20 (B-strength) mg (I-strength) per (I-strength) day (E-strength) having four tokens. Event and medication level 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. Furthermore, we report these scores on each of the nine category, as well as the number of found (tagged) events and medications in each category including wrongly tagged numbers. These detailed scores from each of the nine categories are in the lower part of the evaluation tables.

To further understand the model better, following these, we will also show their token level evaluation in different figure illustrations, both using the evaluation scripts from NeuroNER (Dernoncourt et al., 2017) in line with the CoNLL2003 format (Tjong Kim Sang and De Meulder, 2003).

For BioformerApt, we report the 1st epoch score since the later epochs did not catch up with the 1st epoch's performance. The 1st epoch of BioformerApt represents the original parameters learned from Bioformer pre-trained model with our adapted classification layer.

Regarding TransformerCRF and BioformerCRF, since the later epochs are performing much better than the starting one, we report the later stabled higher performance epoch scores from Epoch 8 for TransformerCRF and Epoch 10 for BioformerCRF.

These evaluation results are reported in Table 1, 2 and 3 separately from BioformerApt, TransformerCRF and BioformerCRF at event and medication level. The learning curves of F1 score vs epoch using TransformerCRF and BioformerCRF are presented in Figure 3 and 4.

The corresponding token level scores in each detailed event and medication categories are displayed in Figure 5 for BioformerApt and 6 for TransformerCRF and BioformerCRF (almost same scores). In addition, the total number of trainable parameters from each of these models is displayed in Table 4.

From the overall evaluation scores in Table 1, 2 and 3, and the parameter numbers in Table 4, it demonstrates that in comparison to BioformerApt model, TransformerCRF has a large margin 17% parameter decreases while achieving much higher overall accuracy score 75.01% versus original 18.67%. By adding CRF layer to BioformerApt, BioformerCRF almost remains the same amount of trainable parameters but produces much higher overall accuracy score 75.01% versus 18.67%.

Even though BioformerApt recognised more events and medications correctly in Corr (correct) criterion (1615 *vs* 717), i.e. how many correctly recognised special tokens/terms, its overall tagging accuracy is too low 18.67%, in comparison to other two models. TransformerCRF has very similar evaluation scores to BioformerCRF but using much less trainable parameters, as well as without using large PLMs and external data set, i.e. the 33 million PubMed abstracts and 1 million PMC full-text. Instead, TransformerCRF was learned from scratch using only 273 EHR letters for training set and 30 EHRs for validation set.

In comparison to event and medication level (let's call it "term level") evaluation (*ref.* Table 1, 2 and 3), the token level evaluation in Figure 5 and 6 demonstrate higher scores. For instances, the recall values on ADE and Dosage from BioformerApt are 6.62% and 3.13% respectively on term level

[11] to be consistent among models, this is also set for BioformerApt, instead of 8 layers used by Bioformer

Acc	Pre	Rec	F1	Corr
18.67%	0.37%	4.61%	0.69%	1615
Catergory	Pre	Rec	F1	found
ADE	0.07%	6.62%	0.14%	61714
Dosage	0.16%	3.13%	0.30%	56548
Drug	1.54%	7.44%	2.55%	52783
Duration	0.01%	0.73%	0.01%	50488
Form	0.37%	3.73%	0.67%	46304
Frequency	0.31%	2.74%	0.55%	44625
Reason	0.29%	4.48%	0.54%	42380
Route	0.49%	5.51%	0.90%	39608
Strength	0.11%	0.95%	0.20%	36737

Table 1: BioformerApt evaluation results using PLM Bioformer.

Acc	Pre	Rec	F1	Corr
75.01%	0.81%	2.05%	1.16%	717
Catergory	Pre	Rec	F1	found
ADE	0	0	0	0
Dosage	0	0	0	0
Drug	2.59%	6.45%	3.69%	27258
Duration	0	0	0	30614
Form	0	0	0	0
Frequency	0	0	0	0
Reason	0	0	0	0
Route	0	0	0	0
Strength	0.04%	0.28%	0.07%	30250

Table 2: TransformerCRF evaluation results using limited constrained training data learned from scratch.

(Table 1), but they are 12.46% and 15.95% respectively on token level performances (Figure 5). This verifies that similar to other domain applications, in clinical domain, the term level recognition on events and medications is much more challenging than single token level among most of the tested categories available in this challenge corpus. In token level evaluation, BioformerApt wins TransformerCRF/BioformerCRF models on F1 scores in most of the special token categories.

Looking back to the overall evaluation scores in Table 1, 2 and 3, we realise that the overall higher accuracy of TransformerCRF and BioformerCRF most owes to the correctness on plain context labelling instead of the event and medication terms, especially, six out of the nine categories with "found" value as 0. They are ADE (adverse

Acc	Pre	Rec	F1	Corr
75.01%	0.81%	2.05%	1.16%	717
Catergory	Pre	Rec	F1	found
ADE	0	0	0	0
Dosage	0	0	0	0
Drug	2.59%	6.45%	3.69%	27257
Duration	0	0	0	30614
Form	0	0	0	0
Frequency	0	0	0	0
Reason	0	0	0	0
Route	0	0	0	0
Strength	0.04%	0.28%	0.07%	30249

Table 3: BioformerCRF evaluation results using PLM Bioformer plus additional CRF learning layer.

Biof.Apt	Transf.CRF	Biof.CRF
42,523,136	35,273,750	42,543,638
–	0.8295(↓17%)	1.0005(↑0.05%)

Table 4: Number of trainable parameters of each model and comparison ratios from TransformerCRF and BioformerCRF to BioformerApt.

Figure 3: TransformerCRF F1 score with each epoch for all classes.

drug event), Dosage, Form, Frequency, Reason, and Route.

To reflect the performances on event and medication focused text, we list such evaluation scores in Figure 7 for BioformerApt and 8 for TransformerCRF and BioformerCRF (they have almost the same scores). These statistics show that even though BioformerApt has much lower overall accuracy, i.e. 18.67%, its term recognition is higher

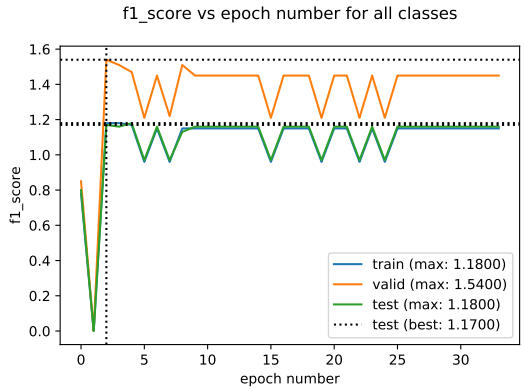

Figure 4: BioformerCRF F1 score with each epoch for all classes.

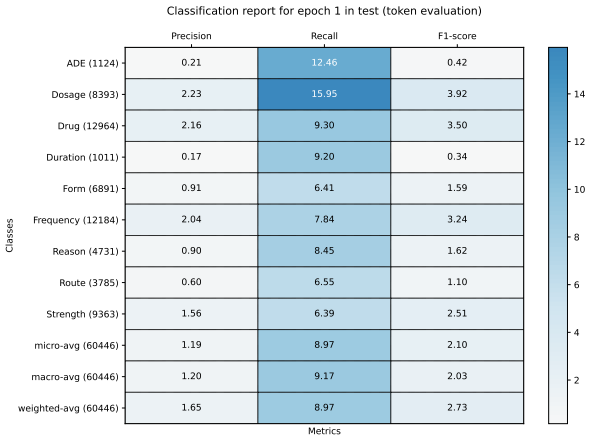

Figure 5: BioformerApt token level performances.

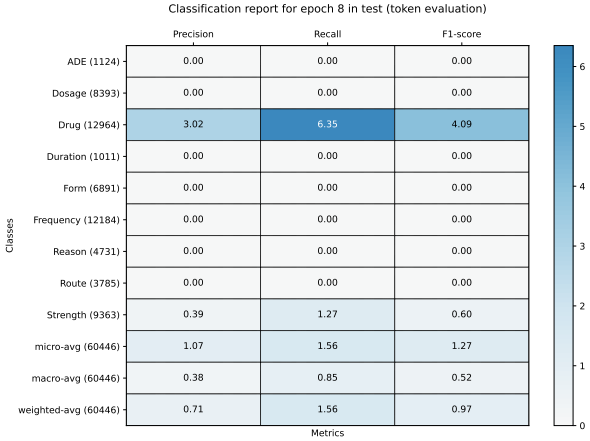

Figure 6: TransformerCRF and BioformerCRF token level performances.

than other two models, especially on the recall value, 87.16% versus 11.12%. This is due to that the tagging output of special labels from BioformerApt is more evenly spread across nine categories, while the other two models shifted to the focus on

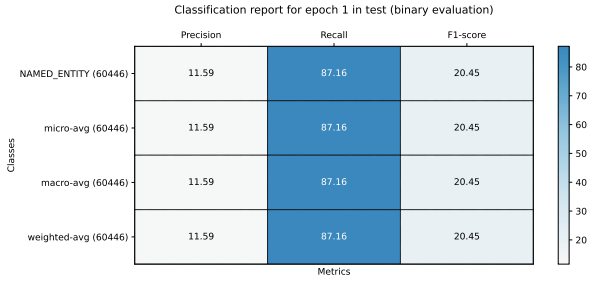

Figure 7: BioformerApt binary evaluation scores.

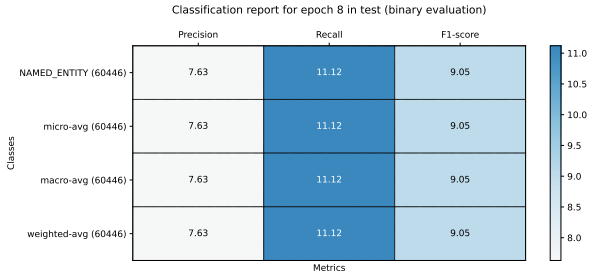

Figure 8: TransformerCRF and BioformerCRF binary evaluation scores.

only three special terms i.e. Drug, Duration, and Strength. We will discuss further on the model behaviours and analysis in the next section.

4.3 Discussion

There are many interesting findings from our experimental investigations that have never been reported by other researchers.

Firstly, the TransformerCRF model which is learned from scratch using only 303 EHR letters in a *data constrained setting* converges to a very similar performance with the BioformerCRF stabled model which is pre-trained using Bioformer from 33 millions of PubMed abstract and 1 million of full text then continuously learned using CRF layer. To investigate if these two models already had the same performances before their stabled models, we look into their pre-stable model output performances, i.e. the earlier epochs, in Table 5 and 6 respectively for epoch 7 from TransformerCRF (stable from epoch 8 to 20) and epoch 9 from BioformerCRF (stable from epoch 10 to 14). They are actually very different before the stabled models, e.g. the Acc (overall Accuracy) score from Epoch 7 of TransformerCRF is 63.63% and the "found" value on Duration is 54871, while these two values are 69.57% and 24785 respectively from Epoch 9 of BioformerCRF model. These are very interesting results, which indicate that the differences

Acc	Pre	Rec	F1	Corr
63.63%	0.64%	2.05%	0.97%	717
Catergory	Pre	Rec	F1	found
ADE	0	0	0	0
Dosage	0	0	0	0
Drug	2.59%	6.45%	3.69%	27258
Duration	0	0	0	54871
Form	0	0	0	0
Frequency	0	0	0	0
Reason	0	0	0	0
Route	0	0	0	0
Strength	0.04%	0.28%	0.07%	30250

Table 5: Pre-stable TransformerCRF evaluation results using Epoch 7.

Acc	Pre	Rec	F1	Corr
69.57%	0.77%	2.17%	1.13%	759
Catergory	Pre	Rec	F1	found
ADE	0	0	0	0
Dosage	0	0	0	0
Drug	2.59%	6.45%	3.69%	27257
Duration	0	0	0	24785
Form	0	0	0	0
Frequency	0.39%	0.84%	0.53%	10702
Reason	0	0	0	0
Route	0	0	0	5829
Strength	0.04%	0.28%	0.07%	30249

Table 6: Pre-stable BioformerCRF evaluation results using Epoch 9.

made by earlier pre-trained Bioformer model can be covered by later CRF learning layer and the model will converge to an almost same stabled version as TransformerCRF which is learned only from scratch in our model settings.

Secondly, using the off-the-shelf biomedical domain PLM Bioformer with adapted classification layer, BioformerApt gives more evenly distributed attention to each of the 9 label categories; however, both TransformerCRF and BioformerCRF converge to stabled models that shift the attention to only three special labels, i.e. Drug, Duration, and Strength. To further investigate into this phenomenon, we carry out statistical analysis on the n2c2-2018 corpus on label distribution. The label distribution of 9 different categories is actually very consistent regarding their percentage out of

label	train	valid	test	train_pcnt	valid_pcnt	test_pcnt
B-ADE	374	26	251	0.05%	0.03%	0.04%
B-Dosage	2857	328	2039	0.37%	0.39%	0.36%
B-Drug	1958	244	1345	0.25%	0.29%	0.24%
B-Duration	506	47	363	0.07%	0.06%	0.06%
B-Form	1124	131	803	0.14%	0.16%	0.14%
B-Frequency	3697	375	2474	0.47%	0.45%	0.44%
B-Reason	1468	175	1091	0.19%	0.21%	0.19%
B-Route	332	21	222	0.04%	0.03%	0.04%
B-Strength	5620	621	3988	0.72%	0.74%	0.71%
E-ADE	374	26	251	0.05%	0.03%	0.04%
E-Dosage	2857	328	2039	0.37%	0.39%	0.36%
E-Drug	1958	244	1345	0.25%	0.29%	0.24%
E-Duration	506	47	363	0.07%	0.06%	0.06%
E-Form	1124	131	803	0.14%	0.16%	0.14%
E-Frequency	3697	375	2474	0.47%	0.45%	0.44%
E-Reason	1468	175	1091	0.19%	0.21%	0.19%
E-Route	332	21	222	0.04%	0.03%	0.04%
E-Strength	5620	621	3988	0.72%	0.74%	0.71%
I-ADE	303	27	208	0.04%	0.03%	0.04%
I-Dosage	4790	491	3482	0.62%	0.59%	0.62%
I-Drug	1059	114	686	0.14%	0.14%	0.12%
I-Duration	400	28	238	0.05%	0.03%	0.04%
I-Form	2022	192	1525	0.26%	0.23%	0.27%
I-Frequency	6768	747	4707	0.87%	0.89%	0.83%
I-Reason	1201	99	914	0.15%	0.12%	0.16%
I-Route	25	0	26	0.00%	0.00%	0.00%
I-Strength	1375	225	1040	0.18%	0.27%	0.18%
O	693886	74155	504301	89.14%	88.71%	89.30%
S-ADE	571	51	414	0.07%	0.06%	0.07%
S-Dosage	1164	157	833	0.15%	0.19%	0.15%
S-Drug	13131	1515	9588	1.69%	1.81%	1.70%
S-Duration	81	12	47	0.01%	0.01%	0.01%
S-Form	5160	546	3760	0.66%	0.65%	0.67%
S-Frequency	3338	388	2529	0.43%	0.46%	0.45%
S-Reason	2149	288	1635	0.28%	0.34%	0.29%
S-Route	4590	549	3315	0.59%	0.66%	0.59%
S-Strength	513	74	347	0.07%	0.09%	0.06%

Figure 9: Label distribution among Train/Dev/Test data.

all special labels across training, development, and test set, as in figure 9. The coverage on all kind of special labels is very low in percentage from 0.01% on S-Duration to 1.70% on S-Drug, even with 0 on I-Route label in the test set. The *relatively* larger ratio on Drug label might explain why these two models shifted attention to it.

Thirdly, the fine-tuning of Bioformer pre-trained model on n2c2-2018 data does not make improvement on model performances, i.e. the later epochs of BioformerApt do not improve evaluation score

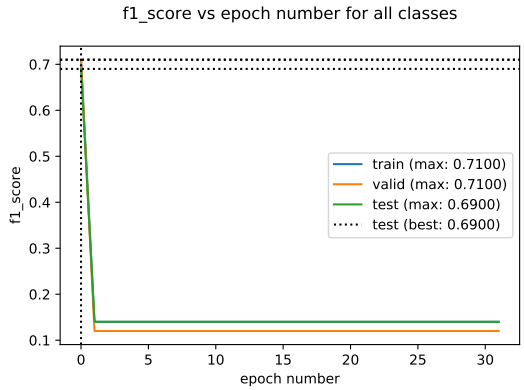

Figure 10: BioformerApt F1 score with each epoch for all classes.

in comparison to Epoch 1, as in Figure 10. Either it is due to the vocabulary used in pre-trained Bioformer is very different to clinical domain data from EHRs used by n2c2 challenge task, or the *constrained fine-tuning* resource is too low to make an influence, i.e. only 303 manually labelled EHRs available from this shared task. This indicates that clinical domain fine tuning using PLM from related or general domain data is not a trivial task, and it can not be taken for granted that the fine-tuned model will surely perform better.

4.4 Revisiting n2c2-2018 Official Results

Looking back to the official results from n2c2-2018 challenge, the high F1 scores achieved by the tops systems were owing to the external knowledge based features they 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). In addition, most of submitted models in n2c2-2018 applied BiLSTM-CRF models, instead, we use the latest deep learning structure Transformer in concatenation with CRF layers. The deployed TransformerCRF model is kind of the only *constrained models* that only use the official training data of 303 annotated letters. 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, held by different shared task organisers, such as data constrained machine translation (Han et al., 2021), following this route.

5 Conclusion and Future Work

In this work, we firstly question the assumption that has always been taken for granted which is *PLMs from mixed/cross domain data can benefit the low resource domain NLP tasks via fine tuning and domain adaptation*. Our experimental results show that using pre-trained Bioformer from large amount of biomedical data from PubMed abstract and PMC full-text corpus, the adaptation and fine-tuning (BioformerApt) using limited amount of available human-annotated clinical EHR data (303 letters, around 45K sentences in our constrained setting) did not improve the model performances in clinical text mining, even though it is related domain, e.g. the shared terminology on drugs.

Instead, the alternative model TransformerCRF taking advantages of both the state of the art Transformer neural attention-based model and the conventional graph-based statistical CRF model outperformed PLM BioformerApt with large margins on overall model Accuracy, even though it only trained from scratch using the same amount of limited data in a *constrained manner*. However, TransformerCRF loses competition to BioformerApt on special token recognition. Furthermore, we demonstrated that by adding a CRF layer on top of PLM Bioformer for continuous learning, BioformerCRF can achieve almost the same performances as TransformerCRF model regarding overall accuracy.

Both TransformerCRF and BioformerCRF are mostly performing correctly on plain context instead of special labels. This indicates that clinical domain text mining remains a very challenging task in data constrained setting. This challenge is partially due to the complexity and diversity of clinical domain data. For instance, n2c2-2018 shared task annotation has nine special labels in comparison to three or four commonly used categories by conventional named entity recognition (NER) tasks e.g. person, location, and organisation (Han et al., 2013).

In the future work, we will examine the *unconstrained* TransformerCRF model by using pre-trained embeddings from entire MIMIC-III clinical data e.g. for semantic reranking of model predictions (Maldonado et al., 2017; Moreau et al., 2018); we will also carry out labelled data augmentation for model training which includes extending manually created golden data and automatic data-augmentation via graph-based model such as graph construction and label propagation to cre-

ate silver data (Han et al., 2022, 2015). We will investigate the influence of character embedding and BPE encoding on our model optimisation for low frequency terms. Finally, we will optimise our current model learning to shift the attention to all special labels instead of only on the higher ratio labels.

Acknowledgements

We thank Google's Colab service at our model developing stage. We thank Hao Li for helpful discussion on our experiments. The authors thank the project support from HIPS (R126035 task A05) and from JigSaw (R124782 task A07) at The University of Manchester. We thank the CSF3 server support from The University of Manchester. We thank the anonymous reviewers for valuable comments on the earlier version of our manuscript. We acknowledge the usage of the following packages: NeuroNER, Transformers, conll2003 eval, Stanford coreNLP, Bioformer and ClinicalBERT.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

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 L. F. Han, Derek F. Wong, and Lidia S. Chao. 2013. Chinese named entity recognition with conditional random fields in the light of chinese characteristics. In *Language Processing and Intelligent Information Systems*, pages 57–68, Berlin, Heidelberg. Springer Berlin Heidelberg.

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, Valerio Antonini, Ghada Alfattni, Alfredo Madrid, Warren Del-Pinto, Judith Andrew, Ana María Aldana Gelves, Robyn Hamilton, Karim Webb, Meghna Jani, William G Dixonmand, and Goran Nenadic. 2022. A transformer-based machine learning framework using conditional random fields as decoder for clinical text mining. In *Poster Presentations in HealTAC 2022: the 5th Healthcare Text Analytics Conference*, London Hub, UCL, London, UK.

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.

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.

Alfredo Madrid, Caitlin Bullin, Lifeng Han, Judith Andrew, Warren Del-Pinto, Ghada Alfattni, Oswaldo Solarte, Ernestina M Ruiz, Luis Rodríguez, Ana María Aldana Gelves, Robyn Hamilton, Karim Webb, Meghna Jani, William G Dixonmand, and Goran Nenadic. 2022. Diagnosis certainty and progression: A natural language processing approach to enable characterisation of the evolution of diagnoses in clinical notes. In *Poster Presentations in HealTAC 2022: the 5th Healthcare Text Analytics Conference*, London Hub, UCL, London, UK.

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.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Erwan Moreau, Ashjan Alsulaimani, Alfredo Maldonado, Lifeng Han, Carl Vogel, and Koel Dutta Chowdhury. 2018. Semantic reranking of CRF label sequences for verbal multiword expression identification. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 177 – 207. Language Science Press.

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.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

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.

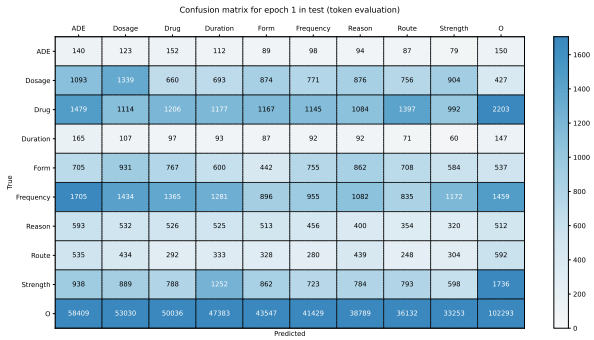

Figure 11: BioformerApt Confusion matrix at token level evaluation from Epoch 1.

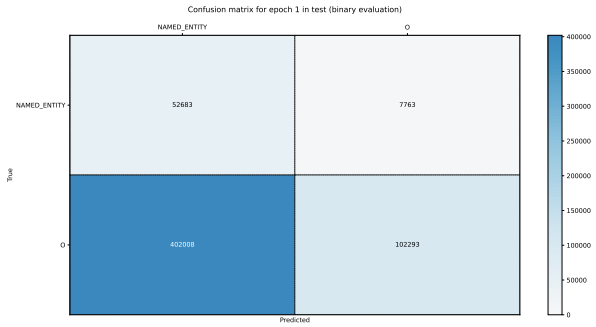

Figure 12: BioformerApt Confusion matrix for binary evaluation at token level from Epoch 1.

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.

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 confusion matrices from each of the model evaluation, detailed BIOES level evaluation scores, as well as the scores from different batch size settings.

Appendix-A: Confusion Matrix

We report the confusion matrices from three different models on how they mis-label each of the nine classes and on overall binary evaluation in Figure 11 and 12 for BioformerApt, Figure 13 and 14 for TransformerCRF (BioformerCRF having very similar score to TransformerCRF).

Appendix-B: BIOES Level Scores

The BIOES level of classification report from BioformerApt is shown in Figure 15. The BIOES level classification report of TransformerCRF and BioformerCRF mostly only fused on S-Drug and S-strength two sub-classes.

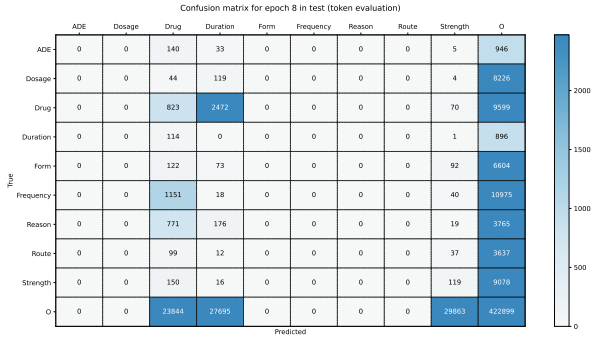

Figure 13: TransformerCRF Confusion matrix at token level evaluation from Epoch 8.

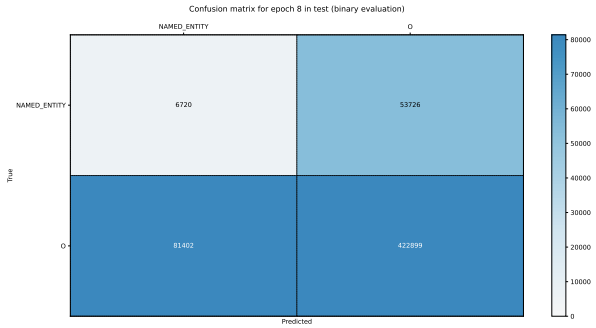

Figure 14: TransformerCRF Confusion matrix for binary evaluation at token level from Epoch 8.

Appendix-C: Batch Size Matters

We report more experimental results from different batch size settings and the findings show that batch size difference does make an influence on model performances. This can be related to our model settings or the very limited of amount of constrained training data from n2c2 shared task.

Figure 16 and 17 show the F1 score learning curves from BioformerApt and BioformerCRF at batch size 10. Figure 18 and 19 show the F1 score learning curves from TransformerCRF and Bio-

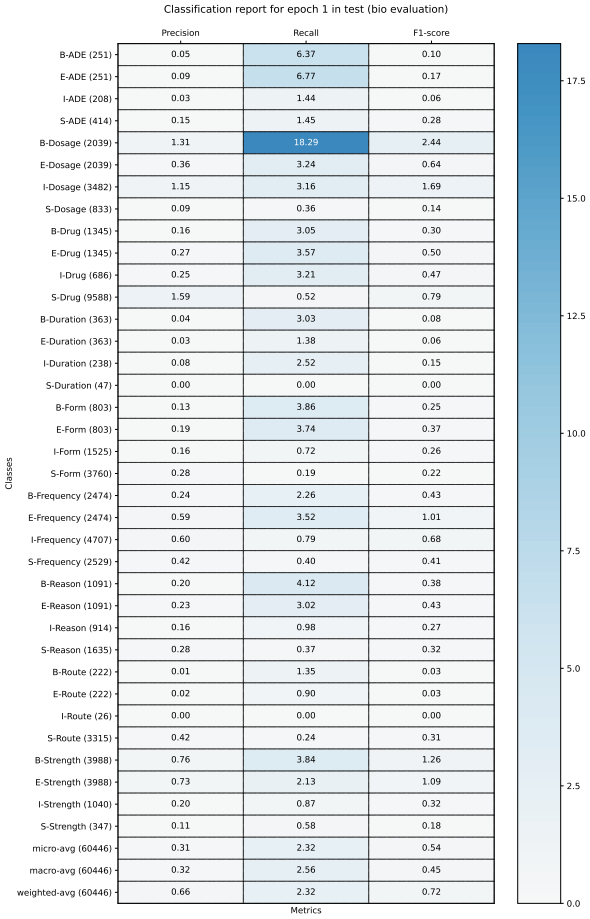

Figure 15: BioformerApt classification report for all BIOES sub-classes at Epoch 1.

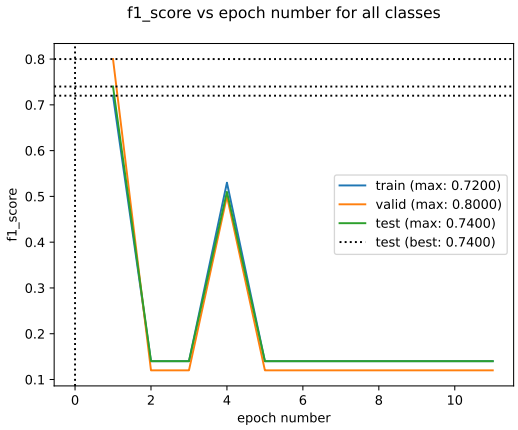

Figure 16: BioformerApt F1 score with each epoch for all classes with batch size as 10.

formerCRF at batch size 1.

Appendix-D: ClinicalBERT-CRF

We also deployed Clinical-BERT (ClinicalBERT) (Alsentzer et al., 2019) which was fine-tuned from

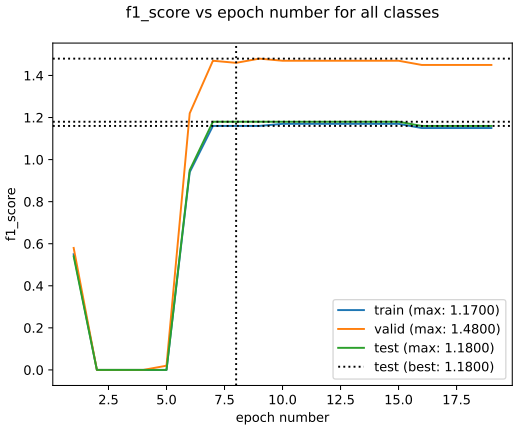

Figure 17: BioformerCRF F1 score with each epoch for all classes with batch size as 10.

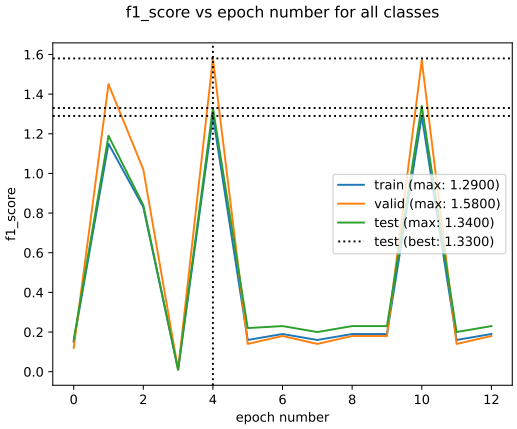

Figure 18: TransformerCRF F1 score with each epoch for all classes with batch size as 1.

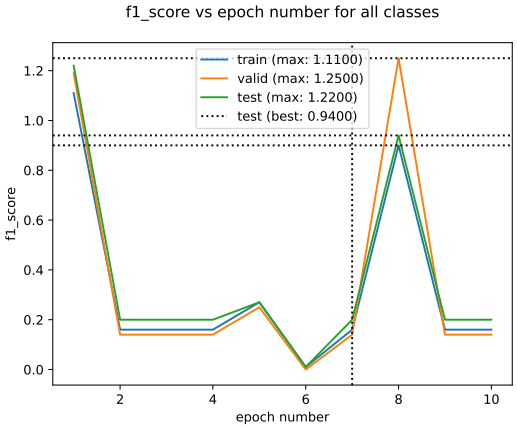

Figure 19: BioformerCRF F1 score with each epoch for all classes with batch size as 1.

BioBERT using MIMIC clinical data such as dis-

charged summaries [12]. We applied CRF layer on top of ClinicalBERT for our experiments. However, the evaluation results do not show much difference/improvement from BioformerCRF (pre-trained plus fine-tuning) and TransformerCRF (learned from 300 letters only). We believe this is related to difficulty of this challenge task itself, e.g. the nine categories of special tokens and their very low frequencies in the context.

Appendix-E: The Open Source Project

Currently, our open-source project https://github.com/poethan/TransformerCRF hosts the following models which are available.

- TransformerCRF: word level model training from scratch
- BioformerCRF: PLM Bioformer with CRF fine-tuning
- BioformerApt: Using pre-trained parameters from Bioformer with an adaptation layer for text mining task
- ClinicalBERT-CRF: using pre-trained ClinicalBERT with CRF layer for fine-tuning

In the future, we plan to carry out the follow-up models using clinical domain pre-trained word embeddings together with sub-word / character-level learning.

The model dependencies/requirements include:

- Python 3.7.13
- Tensorflow 1.14.0 / 2.9.1 (using 2.9.1)
- tensorflow-addons 0.17.0 (if use tf2)
- transformers 4.20.1
- matplotlib
- sklearn
- spacy
- pytorch

[12] model available at https://github.com/EmilyAlsentzer/clinicalBERT