Developing a Clinical Language Model for Swedish: Continued Pretraining of Generic BERT with In-Domain Data

Anastasios Lamproudis anastasios@dsv.su.se
Aron Henriksson aronhen@dsv.su.se
Hercules Dalianis hercules@dsv.su.se
Department of Computer and Systems Sciences (DSV)
Stockholm University
Kista, Sweden

Abstract

The use of pretrained language models, fine-tuned to perform a specific downstream task, has become widespread in NLP. Using a generic language model in specialized domains may, however, be sub-optimal due to differences in language use and vocabulary. In this paper, it is investigated whether an existing, generic language model for Swedish can be improved for the clinical domain through continued pretraining with clinical text.

The generic and domain-specific language models are fine-tuned and evaluated on three representative clinical NLP tasks: (i) identifying protected health information, (ii) assigning ICD-10 diagnosis codes to discharge summaries, and (iii) sentence-level uncertainty prediction. The results show that continued pretraining on in-domain data leads to improved performance on all three downstream tasks, indicating that there is a potential added value of domain-specific language models for clinical NLP.

1 Introduction

Pretrained language models, trained on a variety of readily accessible and large-scale unlabeled corpora, and subsequently fine-tuned on downstream tasks using labeled datasets, have led to substantial performance gains across a whole host of NLP tasks. This has contributed to ameliorating a major bottleneck in the development of NLP systems, i.e. the need for access to very large amounts of labeled data for supervised learning.

In many cases, obtaining large, task-specific datasets in the form of human-annotated corpora is challenging and prohibitively expensive. As a result, the paradigm of pretraining and fine-tuning has become fundamental for contemporary NLP. In particular, with the introduction of models such as BERT (Devlin et al., 2018), which are based exclusively on self-attention, i.e. Transformers (Vaswani et al., 2017), and leverage transfer learning techniques, language models have become increasingly accessible; yet, pretraining language models from scratch requires substantial computational resources.

While generic language models, trained and released to the public by resource-rich organizations, can be utilized and fine-tuned to perform a particular downstream NLP task without a need for significant resources – neither computational nor in terms of data – it has been shown that their use in specialized domains may be sub-optimal as a result of differences in, for instance, language use and vocabulary (Lewis et al., 2020; Gururangan et al., 2020). This has motivated efforts to develop domain-specific language models, e.g. SciBERT (Beltagy et al., 2019) and BioBERT (Lee et al., 2020).

Specialized language models have been developed either by (i) pretraining a language model with in-domain data from scratch, possibly in combination with out-domain data, or by (ii) continuing to pretrain an existing, general language model with in-domain data (*domain-adaptive pretraining*), either by using large amounts of in-domain data, if available, or by only using task-related unlabeled data (*task-adaptive pretraining*).

The need for language models is particularly pronounced in low-resource settings – both in terms of languages and domains. While there is a publicly available generic language model for Swedish, KB-BERT (Malmsten et al., 2020), pretrained using text from the National Library of Sweden, there is no domain-specific variant for Swedish clinical text.

In this paper, we report on the development of a clinical language model for Swedish. The approach is based on continual pretraining of KB-BERT with in-domain data in the form of clinical text. The

Proceedings of Recent Advances in Natural Language Processing, pages 790–797
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_090

model is fine-tuned and evaluated on three downstream clinical NLP tasks: (i) detection of protected health information, i.e. a named entity recognition task, (ii) automatic assignment of ICD-10 codes to discharge summaries, i.e. a document-level multi-class, multi-label classification task, and (iii) uncertainty classification, i.e. a sentence-level multi-class, single-label classification task. The clinical KB-BERT is compared to the original KB-BERT and we report downstream performance of various checkpoints during the pretraining process. The domain-specific and generic BERT models are further evaluated on a generic NER task in order to understand if performance gains are best explained by the quantity or the domain-specificity of the additional pretraining data.

2 Related Work

There has been a substantial amount of effort in recent years dedicated to exploring and developing domain-specific and specialized language models by pretraining with in-domain data, particularly for the biomedical domain and for English.

An early and notable effort was the release of BioBERT (Lee et al., 2020), a BERT model pretrained on large-scale biomedical corpora (PubMed abstracts and PMC full-text articles) in addition to general-domain corpora (English Wikipedia and BooksCorpus). Rather than training the model from scratch, BioBERT was initialized with the general-purpose BERT model and also inherited its vocabulary, after which pretraining continued using biomedical data. It was shown that BioBERT significantly outperforms BERT on biomedical NLP tasks.

Subsequent efforts, e.g. BioMegatron (Shin et al., 2020), have shown that additional improvements can be gained by training larger models on even larger in-domain corpora and, in some cases, using a domain-specific vocabulary. In another study, experimental results indicated that training biomedical language models from scratch, as opposed to continued pretraining of a generic language model, may yield improved performance on downstream domain-specific tasks (Gu et al., 2020), although requiring substantial computational resources.

Domain-specific language models have also been developed for the clinical domain, albeit not for Swedish. Alsentzer et al. (2019) pretrained clinical BERT models on MIMIC-III (Johnson et al., 2016) using either (i) all types of clinical notes or (ii) discharge summaries only.

It was found that initializing the clinical BERT models with parameters from BioBERT, as opposed to parameters from BERT, led to better downstream performance, while the types of clinical notes used made little difference on most downstream tasks. However, the clinical language models yielded an increased performance on some – but not all – of the clinical NLP tasks compared to BERT and BioBERT.

There have been efforts to develop language models using a combination of biomedical and clinical data. In Lewis et al. (2020), the authors develop such models by applying recent advances in pretraining introduced by RoBERTa (Liu et al., 2019), while studying the impact of using different (combinations of) training corpora and model sizes along with a domain-specific vocabulary. Liu et al. (2019) compare their models with previously published language models on a number of downstream tasks in different domains. Their results suggest that using a larger, more powerful general-purpose language model may be better than using a smaller, less powerful domain-specific language model. However, it is also shown that using in-domain data does lead to improved performance: in particular, using clinical data for pretraining leads to large performance gains on clinical tasks but has little impact on biomedical tasks. Learning a domain-specific vocabulary yielded improvements on sequence labeling tasks, while the impact was less clear for classification tasks.

Another very relevant study was conducted by Gururangan et al. (2020), where they also explore the potential advantages in continuing to pretrain an existing BERT model with in-domain data. The authors explore a number of different settings, such as continuing pretraining on a collection of in-domain corpora for a limited amount of time, continuing the pretraining on the unlabeled training set of the intended downstream task, or continuing pretraining on available unlabeled data directly related to the future downstream task at hand. They explore the duration of each of the continued pretraining setups, and they show that this approach can be very beneficial, especially for the setup in which unlabeled data related to the task at hand is exploited for further pretraining. These results are promising and partly inspired the current work, as all the annotated corpora are very small subsets of

the pretraining data.

3 Methods and Data

In this paper, we report on the development of a clinical language model for Swedish. The hypothesis is that a clinical language model, in this case obtained by continuing to pretrain KB-BERT (Malmsten et al., 2020) – a generic language model for Swedish – using large amounts of in-domain clinical text, will yield improved performance over a generic language model on downstream clinical NLP tasks. The pretraining process of KB-BERT is continued using 17.8 GB of clinical text[1] from the research infrastructure Health Bank[2] – Swedish Health Record Research Bank at DSV/Stockholm University (Dalianis et al., 2015). This is, in fact, a similar amount of data that was used for pretraining KB-BERT.

The clinical BERT model is pretrained using a GeForce RTX 1080 GPU and 17.8 GB of uncompressed clinical text in the form of all available types of clinical notes over a period of seven years. The clinical BERT model is trained for one epoch, corresponding to a total duration of ten days. The clinical BERT model is fine-tuned and evaluated on three downstream clinical NLP tasks: (i) detection of protected health information, i.e. a named entity recognition task, (ii) automatic assignment of ICD-10 codes to discharge summaries, i.e. a document-level multi-class, multi-label classification task, and (iii) uncertainty classification, i.e. a sentence-level multi-class, single-label classification task. The clinical BERT model is compared to the original KB-BERT and we report downstream performance of various checkpoints during the pretraining process.

Furthermore, as the clinical BERT model is developed using more data compared to KB-BERT, it is important to investigate whether potential differences in downstream performance can be attributed to the amount of pretraining data rather than the domain-specificity of the data. To that end, the two language models are also evaluated on a NER task in the general domain. A similar improvement on this task could imply that the clinical BERT model is primarily benefiting from additional pretraining data, whereas degraded performance would indicate the value of pretraining specifically on in-domain data.

3.1 Data

Health Bank contains over 2 million patient records encompassing 500 clinical units from the years 2007-2014 from Karolinska University Hospital in Sweden. All clinical notes available in Health Bank – comprising 17.8 GB of uncompressed text – are used for continued pretraining of KB-BERT in order to develop a clinical BERT model for Swedish. In addition, the following four annotated datasets are used for fine-tuning and evaluation, corresponding to three clinical NLP tasks and one generic NLP task.

The Stockholm EPR PHI Corpus comprises 21,653 sample sentences, 380,000 tokens and contains 4,480 annotated entities corresponding to 9 PHI (Protected Health Information) classes: *First Name*, *Last Name*, *Age*, *Phone Number*, *Location*, *Health Care Unit*, *Organisation*, *Full Date*, and *Date Part*. Identifying PHI in clinical notes is a fundamental step in de-identification and is typically approached as a NER task. Details about the dataset can be found in (Dalianis and Velupillai, 2010a).

The Stockholm EPR Gastro ICD-10 Corpus, or **ICD-10 Corpus** for short, consists of 6,062 samples in the form of discharge summaries belonging to a number of ICD-10 diagnosis blocks. This is a document-level multi-class, multi-label classification task, where the ICD-10 codes are grouped into 10 groups with a more coarse granularity compared to the full ICD-10 codes. These groups are decided based on body parts and range from the ICD-10 code K00 to K99 and average 1.2 labels per sample. Details about the dataset can be found in (Remmer, 2021; Remmer et al., 2021).

The Stockholm EPR Sentence Uncertainty Corpus contains 5,515 samples in the form of sentences classified as *Certain*, *Uncertain*, and *Undefined*. The dataset is highly unbalanced with 88% of the samples belonging to the *Certain* class, 10% to the *Uncertain* class and the rest to the *Undefined* class. This is a sentence-level multi-class, single-label classification task. Details about the dataset can be found in (Dalianis and Velupillai, 2010b).

Swedish Web News Corpus comprises approximately 8,000 samples[3] in the form of sentences and contain the entities PER, LOC, ORG and MISC.

[1] This research has been approved by the Swedish Ethical Review Authority under permission no. 2019-05679.

[2] http://dsv.su.se/healthbank

[3] https://github.com/klintan/swedish-ner-corpus/

The dataset comes from Webbnyheter 2012[4], which was annotated semi-automatically. This dataset is used for the general-domain NER task.

3.2 Pretraining

The pretraining of the model with the Masked Language Modeling (MLM) task started from the released checkpoint of KB-BERT and lasted for 40,000 steps, approximately corresponding to one data epoch, or ten days in real time for our GPU. The pretraining procedure of BERT is closely followed with the notable difference that only sequence lengths of 512 are pretrained, acknowledging the significant evidence in the literature suggesting improved performance in later downstream tasks (Liu et al., 2019). Due to the high variability of the different note lengths in the pretraining data, and to construct the 512 chunks of text, the aforementioned work is carefully followed and each note is treated as a document, concatenating the different notes and separating them with extra [SEP] tokens to indicate the end of each document. Lastly, following the original BERT pretraining, a learning rate of $1 \cdot 10^{-4}$ with a linear schedule is used, a batch size of 256 utilizing gradient accumulation, and 10,000 warm up steps. Below, in Table 1, the hyper parameters of the pretraining session are presented.

hyper parameters	KB-BERT	Clinical KB-BERT
learning rate	10^{-4}	10^{-4}
batch size	256	256
Adam optimizer	✓	✓
β_1	0.9	0.9
β_2	0.999	0.999
L2 weight decay	0.01	0.01
warm up steps	10,000	10,000
dropout	0.1	0.1
linear learning rate decay	✓	✓
update steps	1,000,000	+40,000
training sequence length	128 and 512	only 512
MLM probability	15%	15%

Table 1: Pretraining hyper parameters comparison with original KB-BERT.

3.3 Fine-tuning

The primary way in which pretrained models can be evaluated is to fine-tune them to perform a number of tasks and evaluating their performance on these downstream tasks. Utilizing transfer learning, BERT allows for fine-tuning a model to any traditional NLP task with minimal changes. For each task, the core of the language model is kept intact, in this case the KB-BERT model or the subsequent checkpoints of clinical KB-BERT, and only the final classification layers are changed as appropriate depending on the task. The parameters of BERT are not held frozen but are updated for each task since this has been shown to yield an increased performance compared to only training the final layer (Devlin et al., 2018). As the main aim is to compare the further pretrained checkpoints of KB-BERT with the original KB-BERT, an extensive hyper-parameter search is avoided for the different downstream tasks; instead, the hyper parameters used are within the suggested ranges described by Devlin et al. (2018). As such, and as shown in Table 2, a batch size of 32 is used with a learning rate of $2 \cdot 10^{-5}$ for the multi-label classification task, while a batch size of 64 along with a learning rate of $3 \cdot 10^{-5}$ is used for the NER and multi-class classification tasks. In all of the cases, training proceeds until loss convergence and early stopping is utilized to stop the training process at that point.

hyper parameters	PHI	ICD-10	Uncertainty	Web news
learning rate	$3 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	$3 \cdot 10^{-5}$	$3 \cdot 10^{-5}$
batch size	64	32	64	64

Table 2: Fine-tuning hyper parameters.

However, it should be noted that when the goal is to reach the best possible performance, it is critical to perform a proper hyper-parameter search, as other parameter choices may yield a better result. Furthermore, learning rate schedulers, warm up steps, and gradient constraining approaches, such as gradient clipping, should also be explored as possible performance-enhancing changes. In this study, no extensive hyper-parameter search is conducted, nor are other optimization techniques applied, as the goal is to compare the relative performance of two models rather than obtaining state-of-the-art results on the downstream tasks.

All tasks were performed in a conventional setup where three subsets of each dataset are used: a training set that contains approximately 80% of the dataset, a validation set containing approximately 10%, and a test set containing approximately 10%. However, in the case where a train-test split is already provided, as in the case of the Web News Corpus, the test set is left unchanged and the training-validation split corresponds to 90-10% of the original training set.

[4]https://spraakbanken.gu.se/en/resources/webbnyheter2012

4 Results

After fine-tuning KB-BERT and clinical KB-BERT to each of the four tasks – three clinical and one generic – they are evaluated and the results are reported in Table 3 below.

dataset	model	*P*	*R*	F_1
PHI	KB-BERT	**90.53%**	90.34%	90.98%
(Clinical)	Clinical KB-BERT	90.51%	**94.04%**	**92.48%**
ICD-10	KB-BERT	85.53%	75.75%	80.35%
(Clinical)	Clinical KB-BERT	**86.83%**	**79.06%**	**82.76%**
Uncertainty	KB-BERT	92.89%	93.84%	93.05%
(Clinical)	Clinical KB-BERT	**94.70%**	**94.38%**	**93.69%**
Web news	KB-BERT	**89.81%**	**82.47%**	**84.14%**
(General)	Clinical KB-BERT	87.50%	78.38%	80.58%

Table 3: Comparison of the performance of KB-BERT with its clinical KB-BERT counterpart after the end of 1 epoch of further pretraining on 17.8 GB of clinical text. The first three tasks belong to the clinical domain, while the fourth task belongs to the general domain. The best scores are annotated in **bold**.

The results indicate that there is a clear benefit in continuing the pretraining process with in-domain data. The clinical KB-BERT outperforms its KB-BERT counterpart on all three clinical NLP tasks, in terms of F_1-score. On the PHI NER task, it performs close to two percentage points better in terms of F_1-score, with a large increase in recall and more or less the same precision. The improvement on the ICD-10 code assignment task is in the same range as the PHI NER task, but in this case yielding a further increase of both precision and recall. On the Uncertainty task, the performance improvement is not quite as large as for the other two clinical tasks. However, clinical KB-BERT still improves in all the metrics when compared to its KB-BERT counterpart.

However, on the general-domain NER task, the clinical KB-BERT underperforms compared to KB-BERT. It falls short by around 4 percentage points in terms of F_1-score and recall, and by more than 2 percentage points in terms of precision. This indicates that adding more pretraining data does not necessarily lead to better downstream performance, and also that the improved performance on the clinical NLP tasks can likely be attributed to including in-domain data specifically, and not simply more data in general.

Furthermore, a number of checkpoints during the pretraining process of clinical KB-BERT are evaluated on the downstream tasks, the results of which are reported in Figure 1. As can be seen in the figure, for the clinical NLP tasks, there is a positive trend in the performance as the pretraining session progresses. This indicates that, as more data is used, clinical KB-BERT becomes better at incorporating and encoding the differences in the distribution of the clinical text and, as a consequence, it becomes better at performing the downstream tasks.

However, in the case of the Uncertainty multi-class classification task, this trend is not quite as clear: although the vast majority of checkpoints of clinical KB-BERT seem to benefit from the continued pretraining with in-domain data, it experiences a low spike towards the end of the epoch, recovering right at the end. In contrast, the performance of clinical KB-BERT in the general-domain downstream task seems to follow a steadily degrading trend as the pretraining epoch progresses, and does not show any clear signs of recovering.

Finally, to illustrate the differences between a general-domain corpus and a clinical-domain corpus, the KB-BERT tokenizer is used to process the texts in the PHI Corpus and the Web News Corpus, respectively. This tokenizer is a word piece tokenizer, as described by Schuster and Nakajima (2012), and is responsible for constructing the vocabulary – gradually building it from the character level and upwards – by maximizing the likelihood of the training data with respect to the vocabulary. The goal is to investigate how the sentences in the general-domain and clinical-domain corpora are split into tokens, subtokens, and character-level tokens. This is done by calculating the average sentence length, in terms of number of tokens, in the respective corpora when applying the KB-BERT tokenizer versus tokenization based on whitespace and regular expressions. As demonstrated in Figure 2, the clinical-domain corpus, after being preprocessed with the KB-BERT wordpiece tokenizer, leads to a larger increase in average sentence length compared to the general-domain corpus.

5 Discussion

As demonstrated by the experimental results, there is potentially much to be gained from continuing the pretraining process of an existing generic language model with in-domain data, confirming the findings of previous work. Adapting a generic language model to a specific domain by exploiting the availability of unlabeled in-domain data helps BERT to better capture the semantics of the target domain as reflected by differences in the underly-

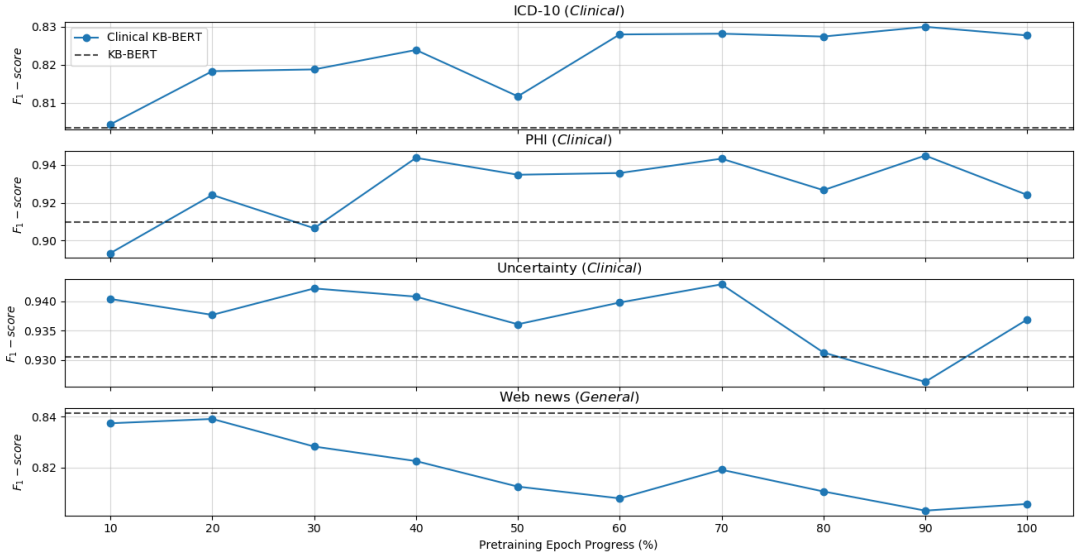

Figure 1: Downstream performance of various checkpoints of clinical KB-BERT during the pretraining process.

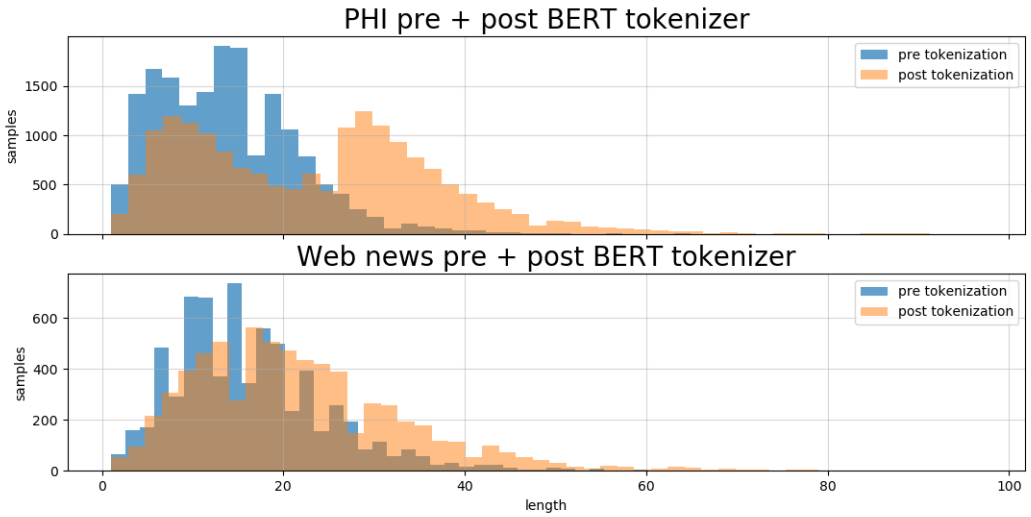

Figure 2: Sentence length distribution, in terms of number of tokens per sample, before and after applying the KB-BERT tokenizer in a clinical-domain corpus (PHI) vs. a general-domain corpus (Web news).

ing distribution. This is highlighted, not only by improved performance on the clinical tasks, but also by the decreased performance on the general-domain task. These results allow us to better understand the reasons behind the improved performance in the target domain and is an aspect that is often overlooked in similar studies. It indicates that improvements yielded by clinical KB-BERT are not solely due to being pretrained on more data – irrespective of domain – but that the domain of the data used for pretraining is indeed an important factor.

It is interesting to observe that these improvements were yielded by continuing to pretrain the language model for only one epoch, and it is possible that further improvements could be obtained by continuing to pretrain on the in-domain data for several more epochs. Moreover, in contrast to similar studies, we also evaluate numerous checkpoints during the pretraining of the the clinical language model. An important observation is that the clinical KB-BERT outperforms the original KB-BERT on all three clinical NLP tasks after using only 20% of the in-domain data. This indicates that it may be worthwhile to adapt general language models and

make them domain-specific even in the absence of enormous amounts of in-domain data.

The general domain and the clinical domain differ primarily in the use of different vocabularies. The vocabulary of a news paper article, or a work of literature, follows a different language distribution compared to a clinical note or a discharge summary. Clinical texts are typically written in a rather peculiar fashion and contain a large amount of technical terms, as well as (ad-hoc) acronyms and abbreviations, that are not as prevalent and may not even exist in the general domain. There may also be domain-conflicting homonyms, where a word has a completely different meaning in one domain compared to another. Due to these differences in vocabulary and frequency, the result of applying a generic language model's tokenizer – in this case that of KB-BERT – to a clinical corpus is that the words are likely to be split into subwords, even potentially reaching a character-level split. This was indeed confirmed by the analysis presented in Figure 2. This, in turn, entails that the BERT model will use more relevant word-level token representations and more common subword token combinations for the general-domain corpus compared to the clinical-domain corpora, where, instead, there is likely to be a high contribution of subword or even character-level token representations. This impact of the tokenizer in turn implies that the major workload and information encoding falls onto this subset of subword and character-level representations during continued pretraining on in-domain data. This not only helps to explain the increased performance on the clinical tasks, but also potentially the performance degradation on the general-domain task since there is a potential mismatch between the representations that are more frequent in the general domain versus the ones that are more frequent – and updated during continued pretraining – in the clinical domain.

In future work, this challenge regarding tokenization can be addressed by pretraining a clinical language model from scratch, which would create a tokenizer and vocabulary based on the in-domain clinical data. As shown by previous work, this may lead to further improvements in performance on the clinical tasks. Another approach is to manually add specific tokens to the vocabulary of a pretrained model, as explored by Tai et al. (2020). An informed set of tokens could potentially be extracted by a new tokenizer specifically trained on the in-domain data, and in a later step, incorporate the set difference to the original tokenizer's vocabulary.

Furthermore, we plan to continue pre-training the current clinical BERT model for more epochs in order to investigate whether this will lead to further improvements in performance, as well as training a new model with pseudonymized data with the aim to make this model publicly available.

Lastly, we also plan to explore and compare different transformer approaches, as well as different pretraining continuation setups, such as using specific parts of the dataset in the spirit of Gururangan et al. (2020). These could include more pretraining continuation setups, such as task-specific pretraining, where the unlabeled training set would be used during the pretraining for more epochs.

6 Conclusions

In this paper, we reported on the development of a clinical language model for Swedish – the first of its kind. The development of the domain-specific BERT model followed the common practice of continuing to train an existing generic language model, KB-BERT, with in-domain data. Compared to previous efforts to develop clinical language models for English, the model was trained using non-pseudonymized clinical data and, in contrast to previously reported results (Alsentzer et al., 2019), yielded improvements also on the de-identification sub-task of identifying protected health information in clinical text.

Furthermore, we carefully investigated the effect of further pretraining an existing language model with in-domain data and evaluated a number of checkpoints during the pretraining process on the downstream tasks. The results showed that continued pretraining with in-domain data yielded improvements on the in-domain tasks, but led to worse performance on a general-domain task, indicating that performance gains on the clinical NLP tasks can indeed be attributed to the domain-specificity rather than the sheer size of the additional pretraining data. Finally, these results further demonstrate the value of developing domain-specific and specialized language models.

Acknowledgments

This work was partially funded by the DataLEASH project and by Region Stockholm through the project *Improving Prediction Models for Diagnosis and Prognosis of COVID-19 and Sepsis with NLP*.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* .

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. pages 3606–3611.

Hercules Dalianis, Aron Henriksson, Maria Kvist, Sumithra Velupillai, and Rebecka Weegar. 2015. HEALTH BANK – A Workbench for Data Science Applications in Healthcare. *CEUR Workshop Proceedings Industry Track Workshop* pages 1–18. http://ceur-ws.org/Vol-1381/paper1.pdf.

Hercules Dalianis and Sumithra Velupillai. 2010a. De-identifying Swedish clinical text-refinement of a gold standard and experiments with Conditional random fields. *Journal of Biomedical Semantics* 1(1):1–10.

Hercules Dalianis and Sumithra Velupillai. 2010b. How Certain are Clinical Assessments? Annotating Swedish Clinical Text for (Un) certainties, Speculations and Negations. In *Proceedings of the of the Seventh International Conference on Language Resources and Evaluation, (LREC), Valletta, Malta, May 19-21*. pages 3442–3446.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779* .

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

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, 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.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4):1234–1240.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. pages 146–157.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* .

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden–Making a Swedish BERT. *arXiv preprint arXiv:2007.01658* .

Sonja Remmer. 2021. *Automatic Diagnosis Code Assignment with KB-BERT ICD Classification using Swedish Discharge Summaries*. Master's thesis, Stockholm University.

Sonja Remmer, Anastasios Lamproudis, and Hercules Dalianis. 2021. Multi-label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT. In *Proceedings of Recent Advances in Natural Language Processing, RANLP 2021, Varna, Bulgaria*.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 5149–5152.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. Biomegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 4700–4706.

Wen Tai, HT Kung, Xin Luna Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 1433–1439.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* .