Evaluating Pretraining Strategies for Clinical BERT Models

Anastasios Lamproudis, Aron Henriksson, Hercules Dalianis
{anastasios, aronhen, hercules}@dsv.su.se

Department of Computer and Systems Sciences

Stockholms universitet

Language models

- Models that contain information and semantics relations about a language or multiple languages
- Aim is to use that information for tasks such as Named Entity Recognition (NER), Categorization or Classification, and more.
- Lately machine learning with Transformer architectures are the most popular approach for language models

Use of big amounts of data in the development in a process called **pre-training** in machine learning. A number of different models can be used with BERT being the most popular.

Pre-training consists of one or more tasks. Popular tasks include **Masked language modelling** (MLM), **Next Sentence prediction** (NSP), and more.

Domain-adapted language models

Models specialized to a domain have been shown to perform better with tasks from this domain.

Many different approaches to adapt language models to a domain with some examples being
- Through pre-training with domain data
- Through changing the vocabulary

Research Question

How do the three different approaches of pre-training a new model, using domain-adaptive pre-training, and adapting the vocabulary in a domain, impact the performance of the final domain adapted model?

In this work

We evaluate three different pre-train approaches for domain adaptation. These include
- Continue the pre-training of an already trained language model with domain specific text.
- Change the language model's vocabulary to a domain specific vocabulary and then continue the pre-training with domain specific text.
- Train a new model with domain specific text.

We use **Clinical text** written in Swedish from the research infrastructure Health Bank that originates from Karolinska University Hospital. The text belongs to approximately 2 million patients over the years 2007 - 2014 collected from 500 clinical units, approximating 18GB.

Baseline

We use KB-BERT, trained with 17 GB of uncompressed text from the National Library of Sweden. The text originates from a great variety of sources such as government documents, Swedish Wikipedia and newspapers.

Models

Clinical KB-BERT v1
KB-BERT domain-adaptive pre-trained with clinical text.

Clinical KB-BERT v2
KB-BERT with changed vocabulary and domain-adaptive pre-training with clinical.

Pure Clinical BERT
New BERT model pre-trained with clinical text.

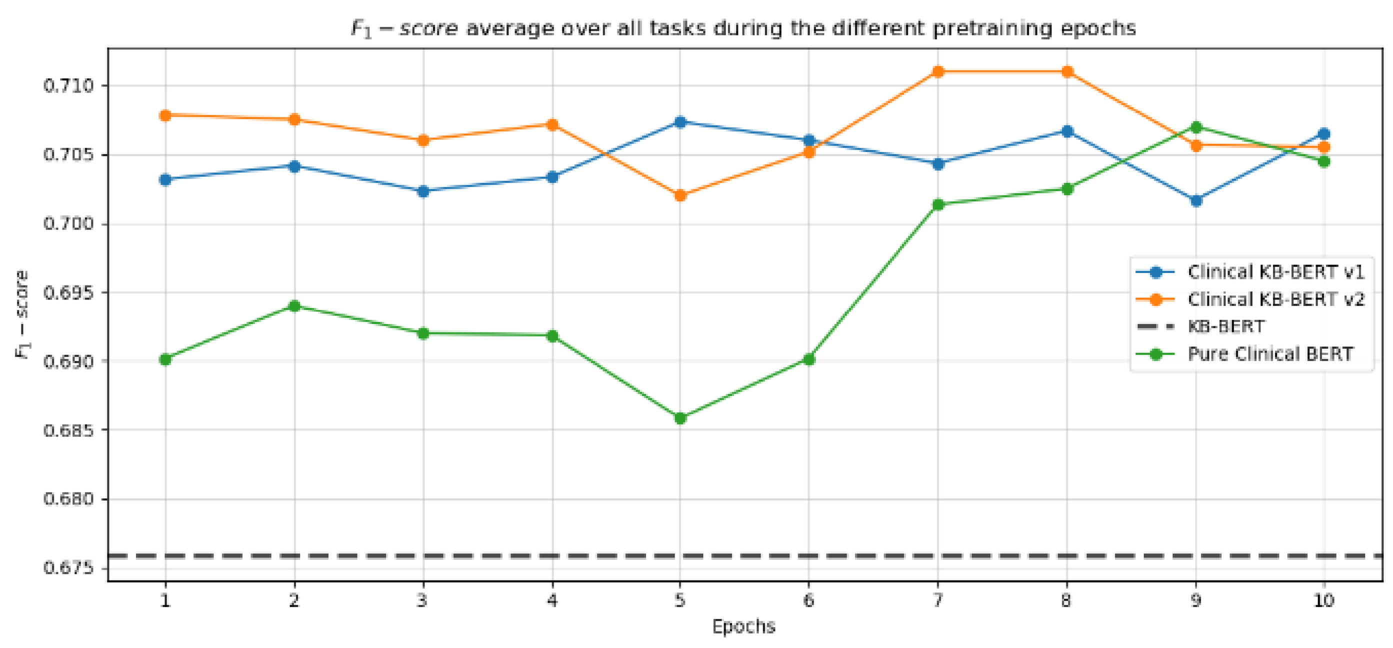
Downstream tasks

For all the tasks, fine-tuning takes place

ICD-10 code assignment, **PHI** entity recognition, **Clinical** entity recognition, **Adverse Drug Event** classification, **Factuality** classification, **Factuality** entity recognition.

Results

Model	ICD-10 Classification	PHI NER	Clinical Entity NER	ADE Classification	Factuality Classification	Factuality NER
KB-BERT	0.799	0.920	0.803	0.183	0.635	0.630
Clinical KB-BERT v1	0.841	**0.948**	**0.862**	**0.199**	0.732	0.690
Clinical KB-BERT v2	**0.848**	0.946	**0.862**	0.196	**0.734**	**0.696**
Pure Clinical BERT	0.844	0.939	0.857	0.193	0.726	0.694



Conclusions

All three approaches benefit the performance of the language model in later downstream tasks.

From the results, **domain-adaptive** pre-training yields the most improvement in performance with the new **vocabulary** adding slightly.

Pre-training a new model lacks behind slightly in performance, becoming competitive later on but not reaching exactly the performance of the other two approaches.