Training a Transferrable Clinical Language Model from 75 million Notes

Madhumita Sushil, Dana Ludwig, Atul J. Butte, Vivek A. Rudrapatna

Bakar Computational Health Sciences Institute, University of California, San Francisco

madhumita.sushil@ucsf.edu

RESEARCH QUESTION

In-domain training

Publicly available models are trained on the MIMIC-III corpus.

How does a larger clinical corpus impact performance?

In-domain vocabulary

Most existing models do not use a clinical vocabulary and tokenizer.

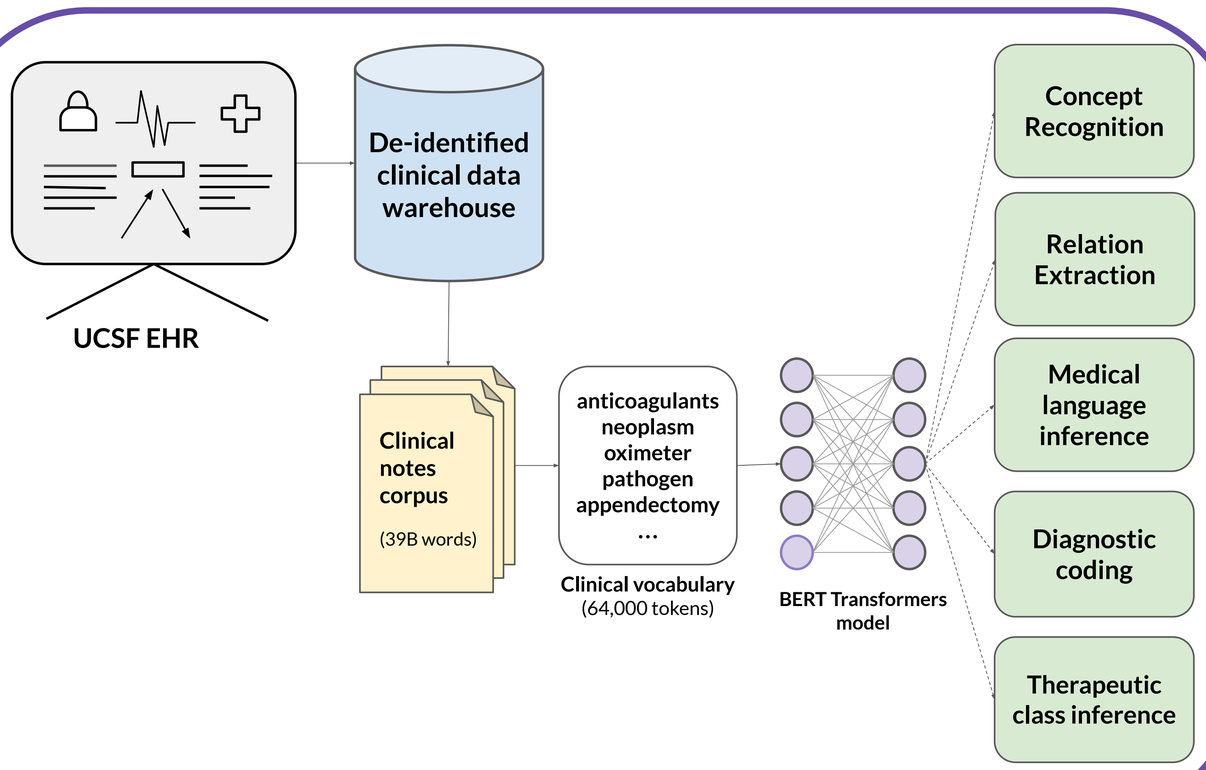
How does a clinical vocabulary impact model fine-tuning?

Where do these models still fail?

What can we learn from large clinical corpora?

What are the limitations of clinical language models?

UCSF-BERT TRAINING AND FINETUNING



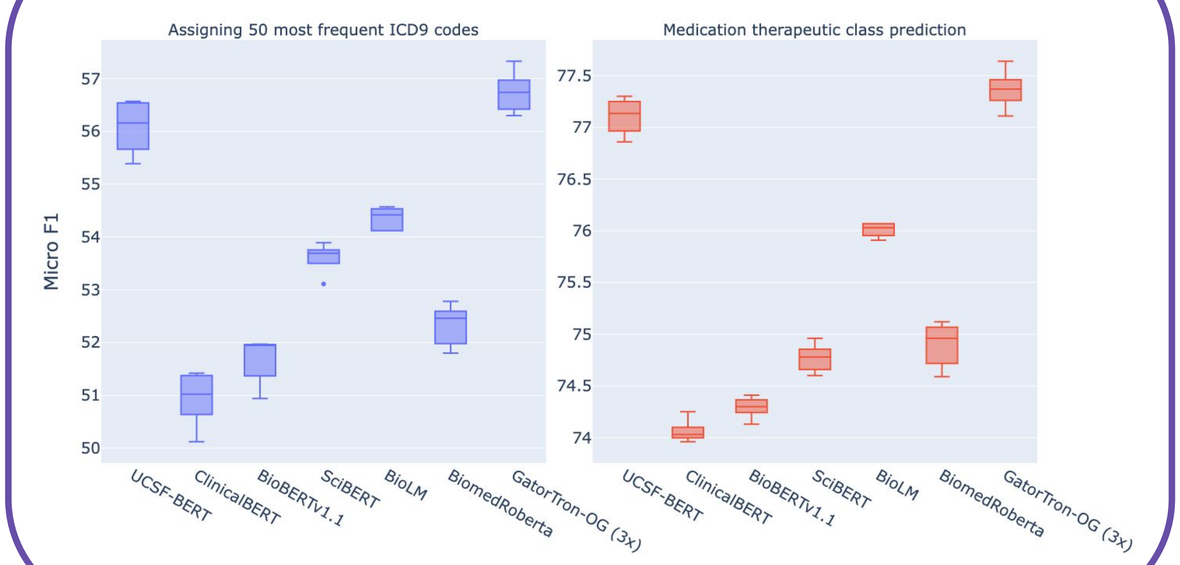
TRAINING CORPUS

- 28M encounters for 2.3M patients
- 2012 – early 2021
- 75 million deidentified clinical reports (39B words)
- Diverse note types: *Imaging reports, progress notes, telephonic encounter notes, consults, ED notes, Pathology and cytology reports, ECG reports, assessment and plan notes, procedure-related notes, discharge summaries, …*

Corpus	UCSF notes	PMC full-text articles	PubMed abstracts	MIMIC-III notes	Scientific papers	English Wikipedia	Books Corpus
Size	39.1	13.5	4.5	0.5	3.2	2.5	0.8

RESULTS

Model / Task	Metric	BioBERT	Clinical BERT	SciBERT	BioMed RoBERTa	BioLM RoBERTa	UCSF-BERT (512)	GatorTron OG
# parameters	N/A	110M	110M	110M	125M	125M	135M	345M
i2b2 2010 NER	%F1	86.0	86.3	86.3	85.0	88.1	**88.3**	**89.1**
i2b2 2012 NER	%F1	77.6	78.0	77.6	76.4	**79.5**	79.3	**80.2**
i2b2 2010 RE	%Micro F1	74.4	74.0	69.8	75.0	75.0	**75.7**	**77.4**
MedNLI	%Acc	82.5	81.8	79.7	85.1	**87.1**	86.8	**88.6**
ICD9-top50 coding	%Micro F1	52.0	51.0	53.7	52.5	54.4	**56.2**	**56.7**
Therapeutic class prediction	%Micro F1	74.3	74.0	74.8	75.0	76.0	**77.3**	**77.4**



CONCLUSIONS AND DISCUSSION

- Training language models on the same data source shows a significant benefit.
  - At-par with public models on public benchmarks
  - SOTA on within-system evaluation
- UCSF-specific clinical vocabulary supports processing of 20% longer sequences with the same model capacity.
- Great at contextual clinical language inference.
- Can potentially be improved by augmenting PubMed, MIMIC-III data

- UCSF-BERT limitations:
  - Domain-specific acronym and abbreviation resolution
  - Relative temporal ordering
  - Numeric inference of infrequent values
  - Implicit causality inference
  - Sequence length limited to 512 tokens

https://arxiv.org/abs/2210.06566