Predicting Clinical Events Based On Raw Text: From Bag of Words to Attention-Based Transformers

Dmitri Roussinov, Andrew Patterson

University of Strathclyde

Andrew Conkie

Red Star Consulting

Christopher Sainsbury

NHS Greater Glasgow & Clyde

Presented at HEALTHCARE TEXT ANALYTICS CONFERENCE 2021 JUNE 17-18

Funded by EPSRC HealTex Feasibility, Dec 2019 – Feb 2020

Synthetic Example of a Clinical Note

Pt placed on a spont breathing trial @ 13:00, pt resp one time within 10 sec - unfortunatly his SBP droppd from 100 to 70 rapidly and therefore the trail was d/c'ed.
Cardiac: BP stable 120-130/60. Pt is on Amiodarone via NGT TID.
Tolerating this well. HR 80–95 most of the shift. Has rare to occ.
Swan numbers done Q6hrs as ordered and probably swan will come out today
He remains on heparin drip which needed to be decreased to 750u/hr at 11PM for PTT 110

- Domain-specific terminology and abbreviations
- Typos
- Grammatical/lexical variation typical for natural language: can convey similar information in many different ways

The Task

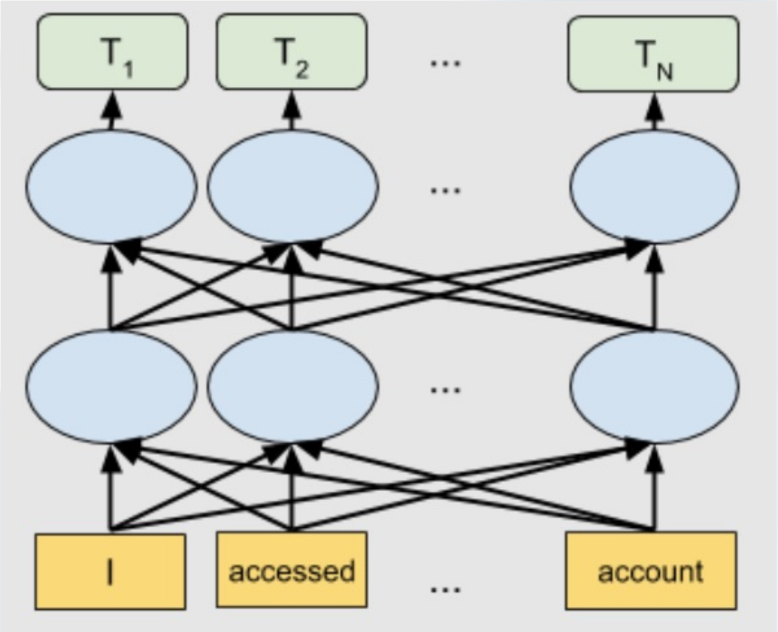
– Predicting Clinical Events Based On Raw Text
  - E.g. *patient readmission within 30 days of discharge*, *patient death within a year, etc.*
– Solution Primarily Focused: pre-trained language models such as *Bert*, *Elmo*, *GPT* or *T5*
– As baseline: classical bags of words classifiers, recurrent/convolutional neural networks, pre-trained word embeddings

Prior Works

- Some successful applications (*ROC > .70* practicality threshold)
- Rarely only raw text, often with *categorical* and *numerical* attributes
- Not clear if any deep neural approaches work better than classical “bag of words” models

Pre-trained Language Models (BERT)

- Impacted **most** applications of *NLP* and *text analytics* in various domains
- Capture the *distribution of word sequences* in a language (or in a specific domain like medical)
- The most popular ones are *attention-based transformers* (e.g. **BERT**)
- Fundamental limitation: number of input *tokens* (roughly 512)
  - which roughly translates to 250 words.
  - 8 times less than we need on average



Combining The Text Segments

- Transformer models use "classification token" (*CLS* = the vector representing the very first input token on the highest layer)
- We split the long texts into **segments** and combine *CLSs* or all top vectors from those segments. Specifically, we've tried:
  – LSTM of CLSs
  – LSTM of all top vectors
  – Concatenation of CLSs
  – Average/min/max of CLSs

Dataset: MIMIC-III

- Very popular dataset
- Sixty-thousand Intensive Care Unit (ICU) admissions
- We used discharge summaries from adult patients
- Roughly 30 thousands records
- No special pre-processing for numbers or abbreviations

Results: compared with baselines

Model	Re-admission AUC	Death within a year AUC	Death or re-admission AUC
Bag-of-words	0.713	0.844	0.763
Deep Neural models:			
Mean-Pooling Word Embeddings	**0.743**	**0.871**	**0.788**
CNN	0.739	0.867	0.787
RNN	0.738	0.873	0.785
Transformer general	0.715	0.776	0.778
Transformer medical	0.741	0.871	0.786

- "AUC" = area under the ROC curve
- Mean-Pooling Word Embeddings AKA "Fast Text", is essentially a deep neural "bag or word" model (ignores the word order, but models word interaction)
- "Transformer general" = BERT
- "Transformer medical" = Clinical BERT by Alsentzer et al.

Results: transformer input size limitation

Model	Re-admission AUC	Death within a year AUC	Death or re-admission AUC
LSTM CLS	**0.741**	**0.871**	**0.786**
LSTM on top layer	-10%	-8%	-9%
Concat top layer	-12%	-11%	-13%
Concat CLS	-2%	-4%	-2%
Average pool CLS	-15%	-10%	-15%
Min pool CLS	-22%	-16%	-21%
Max pool CLS	-17%	-13%	-16%

Conclusions

- All our models reach AUC > .70, which is practically useful level of performance
- Pre-trained transformer-based LM worked only as good as much simpler Fast-Text
- And even only after
  - It was pre-trained on medical texts
  - the input size limitation has been addressed

Ideas For Future

- More datasets
- More experiments with combining several segments
- More advanced pre-trained models (T5, GPT2, *Longformer, Reformer, Performer and BigBird*)
  - Can handle longer texts
  - No medical versions exist yet

Questions ?

- Are welcome by email

dmitri dot roussinov at strath.ac.uk

Or co-authors.