Unsupervised Extractive Summarization of EHRs

Sanya Nijhawan

Department of Computer Science, Yale University, New Haven, CT

LILY Lab

Introduction

Electronic Health Records (EHRs) exist in the form of long, unstructured notes that are time-consuming to go through and hard to parse, and have been associated with physician burnout and dissatisfaction. There is a huge need to efficiently parse EHRs and make them more readable and accessible for healthcare professionals. This task is complex for the clinical notes domain because of minimal labelled data and non-existent 'gold-standard' summaries to learn from.

We explore unsupervised extractive summarization using LexRank (Erkan and Radev, 2004) and BERT (Devlin et al., 2019) to evaluate performance in the biomedical domain and reduce the lengthy clinical notes into shorter summaries while retaining any specialized, technical medical language to capture the most critical information. We also implement libraries and scripts for the models to add to LILY's EHRKit library in the hopes of providing easier access for testing these models and aiding research in the future.

Materials and Methods

We first run the LexRank and BERT unsupervised extractive learning on PubMed corpus research papers and evaluate results in the biomedical domain by comparing the summaries to the "gold standard" paper abstracts using ROUGE (C-Y Lin, 2004) scores. We run these on the introduction section of the papers, on the entire body of the papers and finally on further processed introductions without citations.

We then run the better performing model (LexRank) on MIMIC-III de-identified patient clinical notes data to form summaries by entire history as well as by individual notes for each patient, and manually analyze results.

We also write scripts and a summarizers library which includes Lexrank based off the pypi lexrank package and the model described in the original paper (Erkan and Radev, 2004) as well as folder2rouge to calculate ROUGE scores, which extends the capabilities of files2rouge wrapper. We also write several scripts that directly use the pypi bert_extractive_summarizer library (D. Miller, 2019).

Model	R1	R2	RL	Summary Size (sentences)	Runtime (s)
Lexrank (Introduction)	39.93	13.71	18.96	10 (fixed)	2.29
BERT (Introduction)	36.64	11.95	18.57	11 (mean), 0.5 ratio	794.89
Lexrank (Body)	39.87	14.97	19.89	10 (fixed)	26.39
BERT (Body)	36.49	10.63	18.79	8 (mean), 0.05 ratio	2967.21

Figure 1. PubMed corpus summary statistics (100 documents)

Model	R1	R2	RL	Number of documents
Lexrank (Introduction)	41.2	14.36	19.54	24088
Lexrank (Introduction)	41.1	14.25	19.5	16310
Lexrank (Body)	40.1	15.17	20.17	16310

Figure 2. PubMed corpus LexRank summary statistics (all documents)

Model	R1	R2	RL	Summary Size (sentences)
Lexrank no parentheses (Introduction)	42.56	13.57	21.56	10 (fixed)
BERT no parentheses (Introduction)	37.54	11.81	20.11	11 (mean), 0.5 ratio

Figure 3. PubMed corpus without parentheses summary statistics (100 documents)

Method	# Processed	Avg summary length (sentences)	Runtime (s)
By Note	5546 notes	112	35.27
By Entire History	100 patients	33	18.39

Figure 4. MIMIC-III NOTEEVENTS LexRank summary statistics (TF-IDF calculation: 100 patients, summaries produced: 20 patients)

Patient 7 (2 notes, 35 sentences)
Summary by Note

```
Neonatology
Patient is now 36 h old term infant sent to NICU
 for eval of sepsis because of episode of tachyp
nea witrh feeds in NN.
If cbc normal and patient remains asx here will
return to NN for further care and monitoring.
Infant infant stable in RA.
```

Summary by Entire History

```
NPN Septic evaluation
Infant was brought from NBN for evaluation of
RR and septic work up.
[**Name8 (MD) 7**] MD note for details.
Tags checked with NBN and report given.
Infant placed on warmer and exam done and norm
al.
Infant infant stable in RA.
RR, HR and BP stable.
Examined by MD [**First Name (Titles) 12**] [*
*Last Name (Titles) 13**].
CBC and Bld cx sent.
DS 65.
CBC normal, bld cx pending.
Ordered to send to NBN.
Will check tags and give report.
```

Patient 5 (4 notes, 51 sentences)
Summary by Note

```
Maternal fever [**Known lastname **] 100.7.
If blood culture is positive or any clinical si
gns of sepsis are noted, will pursue further wo
rk-up.
Will follow in Newborn Nursery.
If blood culture is positive or any clinical si
gns of sepsis are noted, will pursue further wo
rk-up.
Will follow in Newborn Nursery.
[**Known lastname **] NICU from L&D for septic
workup due [**Known lastname **] maternal temp
100.7, fetal tachycardia.
No antibiotics indicated at this time.
```

Summary by Entire History

```
NICU Nursing Septic Workup Note
[**Name8 (MD) 7**] NNP note for maternal histor
y and delivery room details.
[**Known lastname **] NICU from L&D for septic
workup due [**Known lastname **] maternal temp
100.7, fetal tachycardia.
VS stable as charted.
Voided and stooled here.
D/S 93.
CBC and BC drawn and sent.
baby cares [**Name2 (NI) 8**].
[**Known lastname **] NBN  and continue current
 plan of care.
No antibiotics indicated at this time.
```

Results

LexRank outperforms this BERT summarizer by more than 3 for R1, by roughly 2-4 for R2, and roughly 0.3-0.9 points for RL. This is because the BERT library doesn't allow a fixed summary length and summaries produced vary from 2 to 53 sentences. The results of body versus just introduction are mixed. Moreover, LexRank is extremely fast as compared to BERT. R1 scores for LexRank, are just 1 point behind most extractive summarization models as listed on the NLP-Progress dashboard whereas R2 and RL scores are dismal. LexRank is outperforming the BERT probably because the BERT library used does not allow a fixed summary length as LexRank does and so the summaries produced aren't equivalent in size.

MIMIC-III note event summary results are far from perfect - there needs to be some way of separating the extracted sentences temporally and tying in more context in some cases. Some other summaries produced (not pictured) included long lists of drugs that would probably not be useful in a summary for a physician at the point of care.

Conclusion

LexRank results are mixed where ROUGE-1 results for PubMed are at par with some other models and it outperforms this particular BERT library, but it still falls short with respect to ROUGE-2 and ROUGE-L scores. For next steps, it would be useful to combine other NLP tasks such as text segmentation for medical notes with LexRank to further improve results.

The BERT extractive summarizer should definitely be tested with medical domain specific pre-trained embeddings. Just like cleaning up well structured PubMed text improved ROUGE scores, a lot more work is needed to further clean up MIMIC data for better results. It is my hope that providing my implementation through a Python library would help further such research.

Acknowledgement

I would like to sincerely thank Professor Dragomir Radev for serving as the advisor on this project and Irene Li for her guidance and support.