Improving Radiology Report Generation Systems by Removing Hallucinated References to Non-existent Priors

Vignav Ramesh[1], Nathan A. Chi[2]*, Pranav Rajpurkar[3]*

*[1] Harvard University   [2] Stanford University   [3] Harvard Medical School   * denotes equal contribution*

Overview

- Current ML models trained to generate radiology reports from chest X-rays all make **hallucinatory references to non-existent prior reports**
- We propose two methods to automatically remove prior references: **(1) FilBERT+GPT-3** and **(2) GILBERT**
- Additionally, we use GILBERT to generate **CXR-PRO,** a novel dataset of chest X-rays and associated radiology reports (derived from the MIMIC-CXR dataset) with prior references omitted
- We retrain CXR-RePaiR, a report generation model, on **CXR-PRO** to form **CXR-ReDonE**
- CXR-ReDonE **outperforms previous report generation methods on clinical metrics**, achieving an average BERTScore of 0.2351 (2.57% absolute improvement)
- Our model-agnostic method is broadly valuable in improving the clinical accuracy of generated reports

Codebase: github.com/rajpurkarlab/CXR-ReDonE
Dataset: physionet.org/content/cxr-pro/1.0.0/

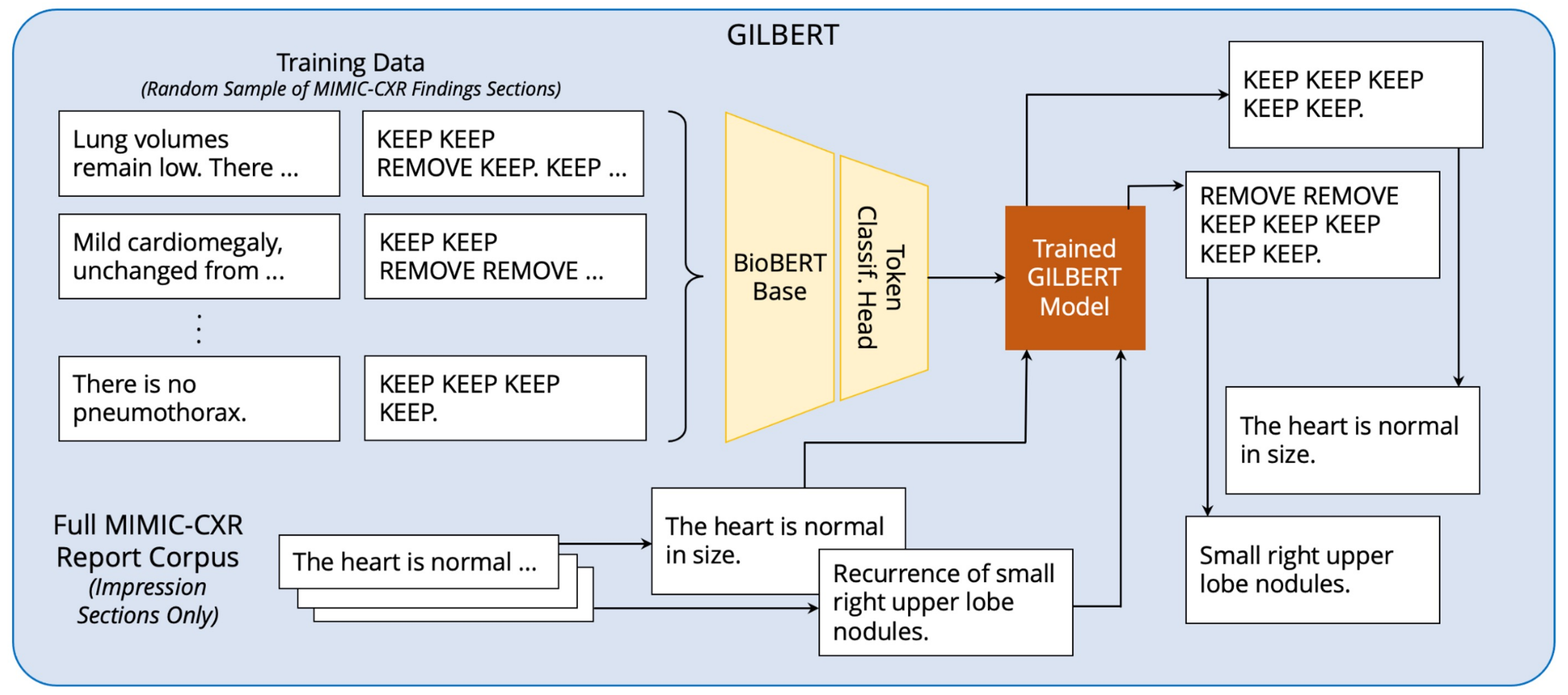
Data and Implementation

- We run GILBERT on **MIMIC-CXR** to generate CXR-PRO
  - Most common references to priors: *change, unchanged, prior*
  - 173,822 (76.3%) of MIMIC-CXR reports contain at least one reference to a prior
- To train FilBERT and GILBERT, we manually create a shared corpus containing tuples of reports and reworded versions with prior references omitted
  - The proportion of references to priors per sentence is intentionally higher in the shared corpus, so as to represent all keywords in a limited space
- We recruit a team of radiologists and medical students to create an expert-edited test set for CXR-ReDonE

Methods

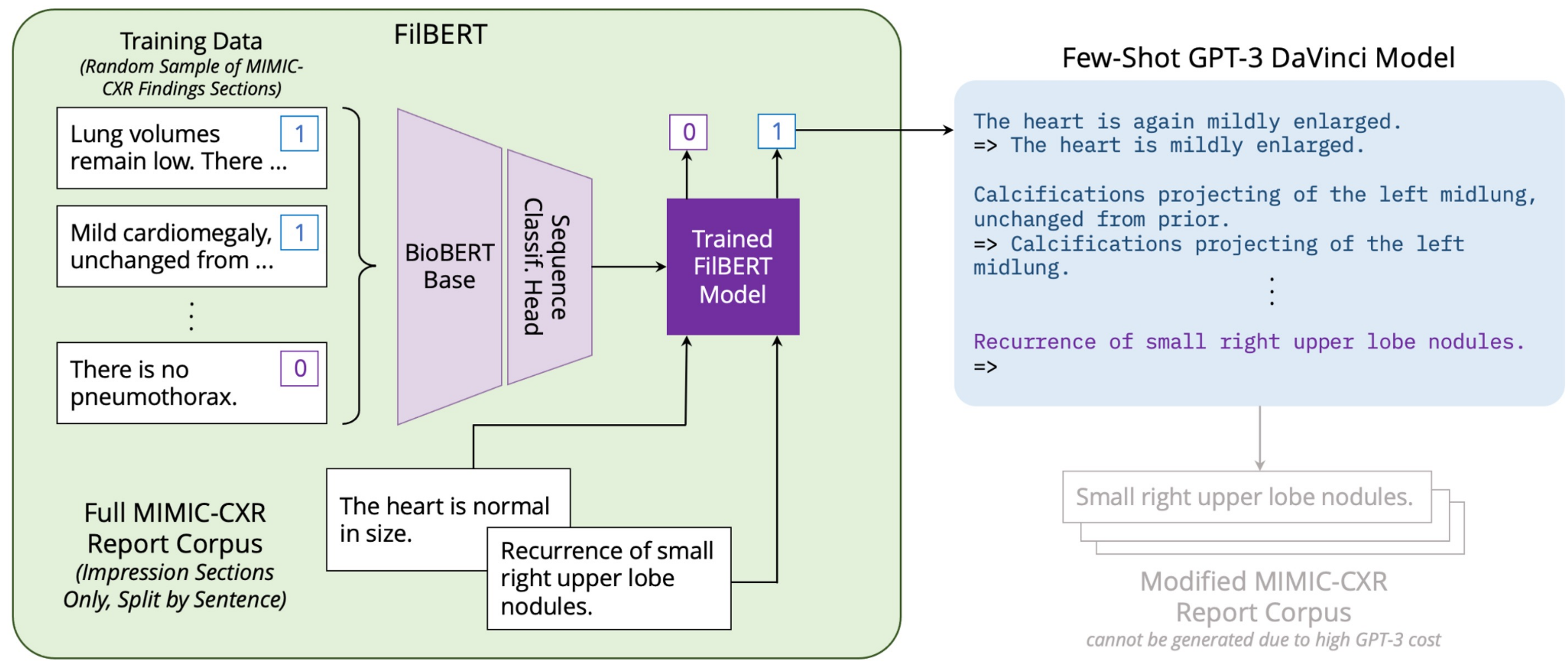
1. GILBERT (<u>G</u>enerating <u>I</u>n-text <u>L</u>abels of References to Priors with Bio<u>BERT</u>)

- We fine-tune a token-level BioBERT model, GILBERT, to classify whether or not a token constitutes a reference to a prior
- Based off standard named entity recognition (NER) task
- Training process: fine-tune BioBERT on pairs of radiology reports and their reworded versions. GILBERT learns to mark tokens as REMOVE or KEEP


GILBERT
Training Data
(Random Sample of MIMIC-CXR Findings Sections)
Lung volumes remain low. There ...
KEEP KEEP REMOVE KEEP. KEEP ...
Mild cardiomegaly, unchanged from ...
KEEP KEEP REMOVE REMOVE ...
There is no pneumothorax.
KEEP KEEP KEEP KEEP.
BioBERT Base
Token Classif. Head
Trained GILBERT Model
KEEP KEEP KEEP KEEP KEEP.
REMOVE REMOVE KEEP KEEP KEEP KEEP KEEP.
The heart is normal in size.
Small right upper lobe nodules.
Full MIMIC-CXR Report Corpus
(Impression Sections Only)
The heart is normal ...
The heart is normal in size.
Recurrence of small right upper lobe nodules.

2. FilBERT+GPT-3: A Two-Step Approach to Prior Reference Removal

- Pipeline of 1) BioBERT for sentence-level classification; and 2) GPT-3 DaVinci for sentence editing
- Motivation: GPT-3 could perform more flexible and fluent editing rather than simply deleting tokens


FilBERT
Training Data
(Random Sample of MIMIC-CXR Findings Sections)
Lung volumes remain low. There ...
1
Mild cardiomegaly, unchanged from ...
1
There is no pneumothorax.
0
BioBERT Base
Sequence Classif. Head
Trained FilBERT Model
0
1
Full MIMIC-CXR Report Corpus
(Impression Sections Only, Split by Sentence)
The heart is normal in size.
Recurrence of small right upper lobe nodules.
Few-Shot GPT-3 DaVinci Model
Small right upper lobe nodules.
Modified MIMIC-CXR Report Corpus
cannot be generated due to high GPT-3 cost

```
The heart is again mildly enlarged.
=> The heart is mildly enlarged.

Calcifications projecting of the left midlung,
unchanged from prior.
=> Calcifications projecting of the left
midlung.
                    ⋮

Recurrence of small right upper lobe nodules.
=>
```

Experiments

GILBERT

- GILBERT attains F1-score = 0.84 on a held-out test set
- Resulting dataset (CXR-PRO) contains far fewer references to priors than MIMIC-CXR, with 259,376 instances in MIMIC-CXR and only 82,074 instances in CXR-PRO: **a >68.3% reduction**

CXR-ReDonE

- Our approach outperforms the baseline on <u>all</u> clinical metrics (see table below) irrespective of *k* (the number of sentences retrieved to compose the final report)
- Qualitatively, note a broad improvement in the factual completeness and consistency of generated reports

k	Training Dataset	Evaluation Metrics		
		BERTScore	s_{emb}	RadGraph F_1
report-level	MIMIC-CXR (Baseline)	0.2083 ± .0023	0.3410 ± .0045	0.0895 ± .0021
	CXR-PRO (Ours)	*0.2160 ± .0025*	*0.3601 ± .0046*	*0.0925 ± .0022*
1	MIMIC-CXR (Baseline)	0.2129 ± .0025	0.3880 ± .0046	0.0838 ± .0023
	CXR-PRO (Ours)	*0.2159 ± .0027*	***0.3967 ± .0048***	*0.0864 ± .0024*
2	MIMIC-CXR (Baseline)	0.2292 ± .0025	0.3822 ± .0045	0.1045 ± .0023
	CXR-PRO (Ours)	***0.2351 ± .0026***	*0.3859 ± .0047*	*0.1056 ± .0024*
3	MIMIC-CXR (Baseline)	0.2179 ± .0025	0.3710 ± .0045	0.1083 ± .0022
	CXR-PRO (Ours)	*0.2254 ± .0025*	*0.3779 ± .0047*	***0.1112 ± .0023***

Ground Truth	Model Trained on MIMIC-CXR	Model Trained on CXR-PRO
A right IJ catheter terminating at the mid right atrium and multiple sternal wires and mediastinal clips are in position.	Various support and monitoring devices removed residual right internal jugular catheter in place and no visible pneumothorax. Cardiomediastinal contours stable. Lungs are remarkable for bibasilar patchy atelectasis the left on the right.	Placement of right internal jugular central venous catheter terminating at the cavoatrial junction without evidence of pneumothorax. Enteric tube grossly courses below the level of the diaphragm, inferior aspect not included on the image.

stable   removed   residual

Acknowledgments

- We would like to thank Dr. Kibo Yoon, Patricia S. Pile, and Pia G. Alfonso for their work on the CXR-PRO test set, as well as Jaehwan Jeong and Ethan Chi for their feedback and technical advice.