Building a Natural Language Processing System to Characterize the Disease Progression in Radiology Reports

Gautham Raghupathi, Hamza El Boudali, Yusef Qazi
{gautham, hamza410, yqazi27}@stanford.edu
Stanford University

Problem

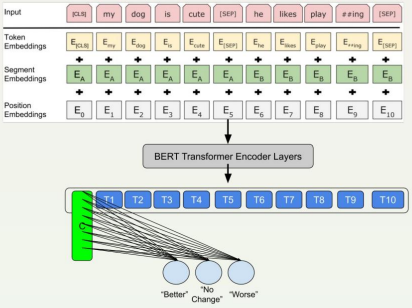
How do we use NLP to extract disease progression from chest x-ray radiology reports?

- Radiology reports are used to train medical image classifiers
- Chest X-rays are the most common radiographic examination
- **Challenge**: Lack of annotated reports

Background

- Our work is similar to CheXbert [1], a state-of-the-art BERT-based model that extracts the presence of clinically important observations from free text radiology reports.
  - They use *Impression* section of reports to extract observations
  - We use *Findings* section of reports to extract disease progression
- We use the CheXpert [2] rule-based labeler as a baseline
- We use the MIMIC-CXR dataset

Model



Experiments

We are using three different BERT-based model variations to test our hypotheses:

- BlueBERT
- BERT
- Bio_ClinicalBERT

Baseline: We are using the CheXpert labeler

- Phrase extraction, aggregation, and classification

Our Deep Learning implementation uses the following:

- Input: Tokenized free-text radiology reports
- Middle: Selected BERT system
- Output: Classification of "Better", "No Change", or "Worse"

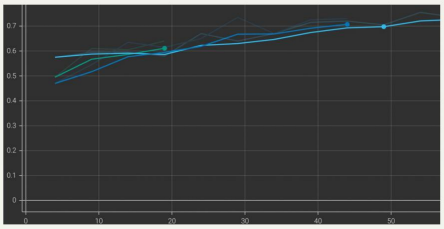
BlueBERT performed the best in Finetune, and BERT performed the best in Linear Evaluation

- However, no real statistical significance

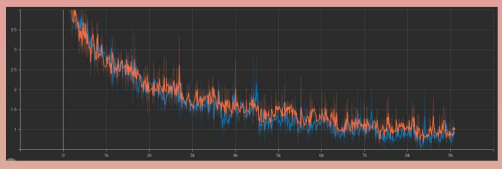
Backtranslation model received BLEU scores of **32.99** (DE-EN) and **31.67** (EN-DE)

BERT	Configuration	
	Finetune	Linear Evaluation
BlueBERT	**0.536**	0.672
BERT (Regular)	0.532	**0.679**
Bio_ClinicalBERT	0.527	0.657
Baseline	0.474	—

Table 2: **Finetune and Linear Evaluation F1 Metrics for Different Pretraining Methods**



F1 Score on Linear Evaluation: Dark Blue: BlueBERT Light Blue: BERT Green: Bio_ClinicalBERT



Validation Loss on Backtranslation: Orange - DE-EN Blue - EN-DE

Analysis

Backtranslation is translating an input into another language and then back to the source

- Helps augment training data with more examples

Backtranslation is supposed to be a key enhancer for our BERT-based models

- Translations actually seem to be nonsensical
- Our Multi30K dataset and Spacy vocabulary doesn't have medical jargon

We need a medical dataset with clinical notes in another language

Findings	Backtranslations
the cardiomediastinal silhouette is normal. there is no pleural effusion or pneumothorax. there is no focal lung consolidation. views of the upper abdomen are normal.	church members <unk> <unk> , <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> .

Example of Backtranslation

Conclusions

Findings:

- BlueBERT (domain-specific pretrained BERT encoder) outperforms the traditionally pretrained BERT encoder when finetuning
- BERT-based models outperform rule-based labelers, as expected

Improvements:

- Acquiring more labeled data and vetting it with radiologists would improve the performances of our methods

Future Work:

- Create a dataset that characterizes the disease progression from prior data points rather than single time point
- Acquire translation data for the medical domain to train back translation system

References

[1] Akshay Smit, Saahil Jain, and Pranav Rajpurkar. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In Empirical Methods in Natural Language Processing (EMNLP), 2020.
[2] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In Association for the Advancement of Artificial Intelligence, 2019.