Labeling Chest X-ray Reports with Markers of Longitudinal Change

Ryan Han, Christina Kwak, Evan Saracay
{ryanhan, kwakc, esaracay}@stanford.edu
Stanford
Computer Science

Problem

Background: Chest X-rays (CXR) are the most common imaging examination and critical to diagnosing and managing many medical conditions. Recently, the use of NLP to extract labels from radiology text reports has enabled the large-scale training of deep learning models for clinical applications focusing on a single point in time.

Problem: Many clinical tasks require comparing multiple points in time to understand disease progression - thus **extracting labels relating to longitudinal change** from radiology text reports would enable the training of AI systems that facilitate tedious comparisons performed by radiologists.

Existing Approach: Little has been done towards characterizing change in imaging datasets.

- Public datasets such as MIMIC-CXR and CheXpert, labeled using NLP, do not contain longitudinal change labels.
- The **only existing** work that focuses on longitudinal change in CXRs uses a rigid text matching approach to match frequent sentences pertaining to disease progression.

Task Proposal

Task Definition: We formulate the report labeling task as a multi-class classification problem where the classes are disease progression, disease stability, and uncertain (no indication).

Datasets: We use 227,827 free text radiology reports from MIMIC-CXR. We randomly selected 1000 reports for manual annotation. Each report was annotated by two human readers, with conflicts determined by committee consensus.

Disease Progression	Disease Stability	Uncertain
194	144	140

Table 1: Class distribution of evaluation set

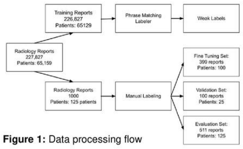

Figure 1: Data processing flow

Proposal: The core idea behind our approach is to utilize both strong and weak supervision in order to maximize the performance and label efficiency of our approach. Our supervision strategies are below, and we explore different ways to combine them in our experiments.

- BERT-phm: Training BERT on a rule-based labeler on all train reports
- BERT-man: Training BERT on a small set of manual annotations
- Distillation: Training BERT on the output of a BERT model on all train reports

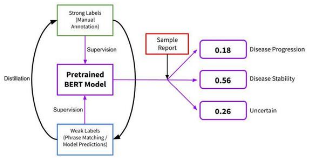

Figure 2: Model training pipeline. Strong and weak labels can be iteratively distilled to maximize label efficiency.

Rule-based baselines:

1. Frequent Sentence Matching (**SM**) - matched 10% of dataset as progression or stability
2. Frequent Phrase Matching (**PHM**) - matched 57% of dataset as progression or stability

Experiment 1 - Supervision Strategies

Approach: We train models using various combinations of strong labels (manual annotations) and weak labels produced from the rule-based phrase matching baseline.

Results:

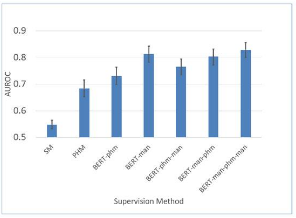

Figure 3: Effect of supervision strategy. All BERT models were fine-tuned from a standard pretrained BERT. Error bars indicate 95% confidence interval calculated using 1000 nonparametric bootstraps.

Analysis:

- We find that the **only existing approach** SM produces **near-random** results on our dataset. Our proposed baseline PHM performs significantly better.
- BERT-phm produces a sizeable yet not statistically significant improvement over PHM. We hypothesize this is due to **pretrained language understanding** rather than simply due to training a model and explore this further in Experiment 2.
- BERT-man significantly outperforms BERT-phm. Attempts to utilize the PHM **weak labels do not produce performance increases** over BERT-man with the slight exception of BERT-man-phm-man. We explore further strategies for incorporating weak labels in Experiment 3.

Experiment 2 - BioMedical Language Representations

Approach: We investigate the effect of pretraining data on model performance.

Results:

Pretraining method	AUROC
General pretraining	
Random initialization	0.697
BERT	0.826
Biomedical pretraining	
BioBERT	0.831
ClinicalBioBERT	0.814
BlueBERT	0.807

Table 2: Effect of pretraining strategy. All BERT models were trained using the manual fine-tuning set. All pretrained models had a statistically significant improvement over random initialization.

Analysis:

- We find that the performance of BERT trained from random initialization is comparable to that of our phrase-matching baseline. This **confirms our hypothesis** that the performance benefit of BERT derives from **pretrained language understanding**.
- Biomedically pretrained BERTs do not significantly outperform default BERT. This may indicate that our model is primarily relying on **low-level pretrained language understanding**, rather than domain-specific concepts. However, this hypothesis needs further exploration which we leave to future work.

Experiment 3 - Distillation

Approach: Taking inspiration from self-distillation, we use our best performing model to produce weak labels on the training set. We then train a BERT model on these labels and fine-tune it using manual annotations.

Results:

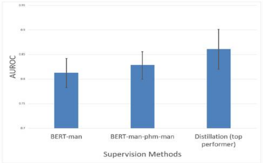

Figure 4: Effect of distillation. Error bars indicate 95% confidence intervals calculated using 1000 nonparametric bootstraps.

Analysis:

- We find that distillation of our BERT-man-phm-man approach produces a sizable (>0.03 AUROC) yet not significant improvement over fine-tuning on manual labels.
- Distillation represents our best performing method that incorporates weak labels.

Experiment 4 - Fine Tuning Training Set Size

Approach: We investigate the effect of changing the number of training samples on fine-tuning performance.

Results:

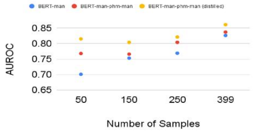

Figure 5: Effect of training set size. Difference between BERT-man-phm-man (distilled) and BERT-man at 50 samples is statistically significant.

Analysis:

- AUROC is roughly linear with respect to number of fine-tuning samples for BERT-man in our observed label domain.
- Weak labeling and distillation provide the largest benefit when labeled examples are scarce

Conclusions

Based on our results, we find that rule based labelers in combination with a small manually labeled set are a viable approach for training a model on our task of detecting longitudinal change in Chest X-Rays. The results of Experiment 3 and 4 show promising avenues for future research that can lead to higher accuracy in model performance, especially in low-label settings. We also see that our Distillation model far outperforms our original sentence matching baseline (SM). Our approach presents a method for accurately extracting labels on medical reports with only a small set of manually labeled reports and a simple rule based labeler.

References:

Dong Yul Oh, Jihang Kim, and Kyong Joon Lee. Longitudinal change detection on chest x-rays using geometric correlation maps. In Dinggang Shen, Tianming Liu, Terry M. Peters,Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pages 748–756, Cham, 2019. Springer International Publishing.

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, 2019.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert, 2020