IBMResearch at MEDIQA 2021: Toward Improving Factual Correctness of Radiology Report Abstractive Summarization

Diwakar Mahajan
IBM Research
dmahaja@us.ibm.com

Ching-Huei Tsou
IBM Research
ctsou@us.ibm.com

Jennifer J Liang
IBM Research
jjliang@us.ibm.com

Abstract

Although recent advances in abstractive summarization systems have achieved high scores on standard natural language metrics like ROUGE, their lack of factual consistency remains an open challenge for their use in sensitive real-world settings such as clinical practice. In this work, we propose a novel approach to improve factual correctness of a summarization system by re-ranking the candidate summaries based on a factual vector of the summary. We applied this process during our participation in MEDIQA 2021 Task 3: Radiology Report Summarization, where the task is to generate an impression summary of a radiology report, given findings and background as inputs. In our system, we first used a transformer-based encoder-decoder model to generate top N candidate impression summaries for a report, then trained another transformer-based model to predict a 14-observations-vector of the impression based on the findings and background of the report, and finally, utilized this vector to re-rank the candidate summaries. We also employed a source-specific ensembling technique to accommodate for distinct writing styles from different radiology report sources. Our approach yielded 2nd place in the challenge.

1 Introduction

The radiology report is a crucial instrument in patient care and an essential part of every radiological procedure, serving as the official interpretation of a radiological study and the primary means of communication between the radiologist and referring physician. According to the American College of Radiology, a radiology report should contain certain components, such as relevant clinical information, imaging findings, limitations of the study, and an impression or conclusion (American College of Radiology, 2020). Of these, the impression is the most important component of the radiology report, containing conclusions based on the pertinent findings and suggestions for additional diagnostic studies if warranted (Wallis and McCoubrie, 2011). Previous studies have shown that oftentimes it is the only part of the report that is read; one previous study found that 43% of referring physicians only read the impression if the report was longer than one page (Clinger et al., 1988), while another study found that 23.1% of clinicians agreed with the statement "I usually only read the conclusion of a radiology report" (Bosmans et al., 2011).

In an effort to support radiologists in writing impressions in radiology reports, Zhang et al. (2018) introduced the task of automatic generation of radiology impression statements by summarizing textual findings written by radiologists. MEDIQA 2021 (Asma Ben Abacha, 2021), as part of NAACL-BioNLP 2021 workshop, aims to further research efforts in summarization in the medical domain. Task 3 of the challenge, Radiology Report Summarization (RRS), focuses specifically on radiology impression generation. The basic task setup is as follows: given the findings and background sections of a radiology report, predict the impression or summary.

In this paper, we detail our participation in MEDIQA 2021 RRS challenge. We developed an approach that utilizes a structured label vector of the impression as our proxy for facts for the impression (predicted using findings and background of the report), to re-rank the generated abstractive summaries from a trained encoder-decoder model. We further employed a source-specific ensembling technique utilizing models fine-tuned to each radiology report source to accommodate for distinct language patterns in each source. Our system performed well in the challenge, placing us 2nd on the leaderboard.

2 Related Work

Abstractive Summarization Systems. Abstractive text summarization has been intensively stud-

Proceedings of the BioNLP 2021 workshop, pages 302–310

ied in recent literature. Rush et al. (2015) introduces an attention-based sequence-to-sequence (seq2seq) model for abstractive sentence summarization. Recent models (e.g. Lewis et al. (2019); Zhang et al. (2020)) employ techniques like denoising or Gap Sentence Generation task for pretraining, to help generation tasks including summarization. However, there are a few domain-specific versions of these state-of-the-art models. Other works like Liu and Lapata (2019); Rothe et al. (2020) have demonstrated the effectiveness of initializing encoder-decoder models from pre-trained encoder-only models, such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), for seq2seq tasks providing competitive results in summarization tasks. Our works builds on these findings and utilizes a pre-existing domain-specific pretrained transformer model in an encoder-decoder setting for our summarization task.

Summarization and Factual Correctness in Radiology Reports. Zhang et al. (2018) first studied the problem of automatic generation of radiology impressions by summarizing textual radiology findings, and showed that an augmented pointer-generator model achieves high overlap with human references. They also found that about 30% of the radiology summaries generated from neural models contain factual errors. Research scholars also integrated Radlex ontology into seq2seq models (MacAvaney et al., 2019) to enhance the clinical validity of automated impression prediction systems within the radiology workflows. In their next work, Zhang et al. (2019b) improved upon the problem of factual correctness in radiology reports by optimizing fact scores defined in radiology reports with reinforcement learning methods. They also introduced a new metric Factual F_1 comparing the predicted summaries using a descriptor vector of the gold summary. In our work, we extend the ideas put forward by Zhang et al. (2019b) by utilizing a descriptor vector (generated using off-the-shelf systems like CheXpert (Irvin et al., 2019) or CheXbert (Smit et al., 2020)) to re-rank the automatically generated summaries.

3 Task Description and Dataset

The MEDIQA-2021 RRS task is defined as follows: given a passage of findings represented as a sequence of tokens $x = \{x_1, x_2, \ldots, x_N\}$, with N being the length of the findings, and a passage of background represented as a sequence of tokens $y = \{y_1, y_2, \ldots, y_M\}$ with M being the length of the background, find a sequence of tokens $z = \{z_1, z_2, \ldots, z_L\}$ that best summarizes the salient and clinically significant findings in x, with L being an arbitrary length of the impression or summary[1].

Type	Source-specific Size		Total Size
	MIMIC-CXR	Indiana	
Training	91,544	0	91,544
Validation	2,000	2,000	4,000
Test	?	?	600

Table 1: Dataset statistics.

Datasets for training and validation of summarization models provided by the MEDIQA organizers consisted of radiology reports with findings, background, and impression sections. The training set consists of 91,544 radiology reports from the MIMIC-CXR database (Johnson et al., 2019), while the validation set consists of an additional 4,000 radiology reports - 2,000 from MIMIC-CXR and 2,000 from the Indiana Network for Patient Care (Indiana) (Demner-Fushman et al., 2016). As part of the shared task rules, the rest of the publicly available MIMIC-CXR and Indiana radiology reports were not allowed for use in training or validation. However, the organizers allowed the use of validation data for training. At the conclusion of the shared task, to evaluate participant systems, a test set of 600 radiology reports containing only findings and background sections were released with their sources unknown at the time of the challenge. Dataset statistics are presented in the Table 1.

4 System Description

Our system is a three-step process in which we (1) utilize pre-trained transformer-based language models in an encoder-decoder setting to get our base summarization models, (2) improve the factual correctness of our base models' predictions by incorporating a re-ranking methodology, and (3) utilize a source-specific ensembling technique which identifies the source of a radiology report, and chooses the prediction of the best performing source-specific model accordingly. We detail the above three steps in the following sections.

[1]Throughout this paper we use terms "impression" and "summary" interchangeably.

4.1 Base Models

Previous work by Liu and Lapata (2019); Rothe et al. (2020) have demonstrated the effectiveness of initializing encoder-decoder models from pre-trained encoder-only models, such as BERT and RoBERTa, for seq2seq tasks. Inspired by this work, we experimented with pre-trained transformer models used as both encoder and decoder with parameters shared between encoder and decoder. Using this setup, we experimented with RoBERTa-large, which showed promising results in Rothe et al. (2020), and BioMed-RoBERTa-base, a domain-specific version of RoBERTa that is publicly available[2] from AllenNLP (Gururangan et al., 2020), and fine-tuned both models using the training set of 91,544 MIMIC-CXR reports. Of the two models, BioMed-RoBERTa-base achieved better results and was therefore used as our initial model for subsequent experiments.

Next, we conducted experiments to evaluate the performance of this initial model on different radiology report sources. As the provided training and validation data contains two sources, MIMIC-CXR and Indiana, each with their distinct language (more details in Section 4.3) and official test data could be any source, we further developed two more base models. Using the initial BioMed-RoBERTa-base model fine-tuned on MIMIC-CXR training set, we further fine-tuned the initial model in two settings: (1) with a subset of reports in the Indiana validation dataset, and (2) with a subset of reports in the Indiana and MIMIC-CXR validation dataset.

Our end result is three base models tuned for 3 source categories:

- BioRoBERTa$_{(M)}$: BioMed-RoBERTa-base fine-tuned on MIMIC-CXR training data. This is the base model for MIMIC-CXR source.
- BioRoBERTa$_{(M+I)}$: BioRoBERTa$_{(M)}$ further fine-tuned on Indiana validation data. This is our base model for Indiana source.
- BioRoBERTa$_{(M+M+I)}$: BioRoBERTa$_{(M)}$ further fine-tuned on both MIMIC-CXR and Indiana validation data. This is our base model for unknown sources.

[2] https://huggingface.co/allenai/biomed_roberta_base

4.2 Fact-Aware Re-ranking (FAR)

Previous works in extracting structured labels from free-text radiology reports have identified 14 observations based on clinical relevance and the prevalence in the reports, and have developed automated systems to predict a 14-observations-vector for an impression summary of a radiology report (Irvin et al., 2019; Smit et al., 2020). The 14 observations are: "Atelectasis", "Cardiomegaly", "Consolidation", "Edema", "Enlarged Cardiomediastinum", "Fracture", "Lung Opacity", "Lung Lesion", "No Finding", "Pneumonia", "Pneumothorax", "Pleural Effusion", "Pleural Other", and "Support Devices". "Pneumonia", despite being a clinical diagnosis, was included as a label in order to represent the images that suggested primary infection as the diagnosis. The 13 observations (excluding "No Finding") take on one of the following classes: blank, positive, negative, and uncertain. The 14th observation, "No Finding", is intended to capture the absence of all pathologies, and takes on only one of the two following classes: blank or positive.

Utilizing this 14-observations-vector we developed an approach to improve the factual correctness of our base models by incorporated a factual re-ranking component that re-ranks our N highest scoring summaries predicted from a base model. We achieve this in the following steps, we (1) first fine-tune a transformer-based language model to predict the 14-observation-vector of the impression given the finding and background of a radiology report, (2) obtain top N highest scoring candidate summaries predicted from our base encoder-decoder model (3) use CheXbert to obtain the 14-observation-vector for each of the N candidate summaries, and (4) use a similarity function between predicted 14-observation-vector for impression (obtained in step 1) and each vector for N candidate summaries obtained in step 3 to re-rank these summaries. Finally, we use the highest similarity scoring candidate summary as our impression summary. We detail our impression 14-observation-vector prediction and our similarity function in the following sections.

We apply our FAR methodology on the three base models introduced in section 4.1 to get our three source-specific models, and denote the new models as BioRoBERTa$_{(M),FAR}$, BioRoBERTa$_{(M+I),FAR}$, and BioRoBERTa$_{(M+M+I),FAR}$, respectively.

Source	Finding	Background	Impression
MIMIC-CXR	There is hyperexpansion of both lungs with severe underlying emphysema. Minimal blunting of the right costophrenic angle may reflect underlying atelectasis. No pleural effusion or pneumothorax identified. The size the cardiomediastinal silhouette is within normal limits.	INDICATION: ___ year old woman with COPD exacerbation // evaluate lung sizes, look for PNA TECHNIQUE: AP portable chest radiograph COMPARISON: No prior radiographs available. Comparison is made to the CT torso from ___	No radiographic evidence of acute cardiopulmonary disease. Hyperexpanded lungs with severe underlying emphysema.
Indiana	Heart size and mediastinal contours appear within normal limits. Hyperinflated lungs with flattening of diaphragms, compatible with emphysema. No focal consolidation, pleural effusion or pneumothorax. No acute bony abnormality.	Indication: Short of breath. Comparison: None.	1. Emphysema. 2. No acute cardiopulmonary abnormality.

Table 2: Example depicting the difference in language between MIMIC-CXR and Indiana reports for findings, background and impression sections.

4.2.1 Impression 14-Observations-Vector Prediction

We utilize the 14-observation-vector representation of the impression section of a radiology report predicted by CheXbert as our ground truth label in a prediction task given the finding and background section of the report as inputs. In this process, for each given radiology report that has findings, background and impression section, we (1) first utilize CheXbert to obtain 14-observations-vector representation of the impression section, (2) convert the multiple values of each of the 14 observations to be binary (i.e. presence or absence of the observation)[3], (3) train a transformer-based language model using finding and background (concatenated) as input to predict 14-observations-vector of the impression section.

4.2.2 Similarity Function

Among the 14 observations categories predicted in CheXbert, "No Finding" is intended to capture absence of all pathologies, i.e. if "No Finding" is positive then all other observations must be negative. Therefore, we constructed our similarity function in cases where (1) "No Finding" is not matched, we assign a similarity score of 0, (2) "No Finding" is a match, the similarity score is the cosine similarity between the rest of the vector representing the 13 other observations.

[3]CheXbert outputs for 13 observations one of the following classes: blank, positive, negative, and uncertain. For the 14th observation corresponding to No Finding, the labeler only outputs one of the two following classes: blank or positive. We convert uncertain to positive and blank to negative to get binary positive and negative output for all 14 observations.

4.3 Source-specific Ensemble

We observed in the provided training and validation data that MIMIC-CXR and Indiana reports use distinctly different language when expressing findings, background, and impression, even when the conveyed content is very similar. As shown in Table 2, although both the MIMIC-CXR report and Indiana report convey the same two key findings in their impression, "emphysema" and "no acute cardiopulmonary disease", the MIMIC-CXR report describes these findings with more detail in prose form, while the Indiana report lists the findings more concisely using a numbered list form. This variation in language between different healthcare organizations is common in the clinical NLP domain, resulting in a need to adapt algorithms depending on the applicable dataset (Carrell et al., 2017).

To address this, we trained a BERT-based source-specific classifier which predicts the source given the findings and background as input. We trained this model using a subset MIMIC-CXR and Indiana reports. However, during prediction or evaluation phase, we chose a higher threshold of 0.7 for predicting a source i.e. if an input is predicted to be Indiana or MIMIC-CXR with a probability of 0.7 or higher, we predict it to be Indiana or MIMIC-CXR respectively, otherwise it is marked to be of an unknown source. Based on the predicted source of a test sample (MIMIC-CXR, Indiana or unknown), the source-specific models' output is chosen as the prediction for that sample.

4.4 Evaluation Metrics

We use two sets of metrics to evaluate model performance at the corpus level, ROUGE and Factual

Model	$MIMIC_{200}$				$Indiana_{200}$				$Combined_{400}$			
	R-1	**R-2**	**R-L**	**F-F_1**	**R-1**	**R-2**	**R-L**	**F-F_1**	**R-1**	**R-2**	**R-L**	**F-F_1**
RoBERTa-large$_{(M)}$	0.634	0.509	0.602	0.768	0.425	0.259	0.415	0.634	0.533	0.390	0.516	0.725
BioRoBERTa$_{(M)}$	0.642	0.513	0.617	0.770	0.449	0.273	0.437	0.638	0.541	0.391	0.520	0.729
BioRoBERTa$_{(M),FAR}$	**0.647**	**0.524**	**0.623**	**0.781**	0.455	0.276	0.442	0.665	0.546	0.394	0.523	0.734
BioRoBERTa$_{(M+I)}$	0.499	0.356	0.472	0.694	0.691	0.605	0.677	0.678	0.594	0.480	0.574	0.709
BioRoBERTa$_{(M+I),FAR}$	0.507	0.362	0.481	0.717	**0.701**	**0.626**	**0.685**	**0.685**	0.596	0.480	0.577	0.716
BioRoBERTa$_{(M+M+I)}$	0.585	0.463	0.563	0.712	0.685	0.597	0.671	0.660	0.642	0.539	0.623	0.719
BioRoBERTa$_{(M+M+I),FAR}$	0.592	0.469	0.570	0.719	0.687	0.601	0.676	0.667	0.647	0.544	0.629	0.726
Ensemble	0.632	0.519	0.611	0.768	0.692	0.604	0.672	0.679	**0.670**	**0.568**	**0.650**	**0.741**

Table 3: Results of our base model, factually correct re-ranking and source-specific ensembling experiments on our internal test data of 200 MIMIC-CXR and 200 Indiana radiology reports. Combined presents results for each model when both the sources (400 reports) are considered together. R-1, R-2, R-L and F-F_1 represent ROUGE-1, ROUGE-2, ROUGE-L and Factual F_1 scores respectively.

F_1. The organizers used ROUGE and CheXbert F_1 metrics for evaluation. ROUGE-2 F_1 metric was used for the task leaderboard.

ROUGE We use the standard ROUGE scores (Lin, 2004), and report the F_1 scores for ROUGE-1, ROUGE-2 and ROUGE-L, which compare the word-level unigram, bigram and longest common sequence overlap with the reference summary, respectively.

Factual F_1 For factual correctness evaluation, we use a Factual F_1 score as proposed by Zhang et al. (2019b). The Factual F_1 scores are calculated by 1) running the CheXbert labeler on both the reference and generated summaries to obtain the binary presence values of a collection of disease variables 2) calculating the F_1 score for each of the variables over the entire test set, using reference values as oracle; and 3) obtaining the macro-averaged F_1score over all variables. Following the process in Zhang et al. (2019b), we exclude some variables due to their small sample sizes (with less than 5% positive ratio in the entire dataset). We included only Cardiomegaly, Lung Opacity, Lung Lesion, Pneumonia, Atelectasis, Pleural Effusion and No Finding in our calculation of Factual F_1 scores.

CheXbert F_1 The organizers used CheXbert F_1 score to calculate the factual correctness, which follows the same process as Factual F_1. However, in their calculation they considered a different set of observations which were found prominent in the official test data: Cardiomegaly, Lung Opacity, Edema, Pneumonia, Atelectasis, Pleural Effusion and No Finding.

5 Experiments & Results

5.1 Data

As noted in section 3, training and validation datasets provided in MEDIQA 2021 can be combined and re-split. We set aside 200 radiology reports each, randomly chosen from MIMIC-CXR validation dataset and Indiana validation dataset, to form our combined internal test dataset. The remaining 1,800 reports each from MIMIC-CXR validation data and Indiana validation data, along with 91,544 of MIMIC-CXR training data are utilized for training.

For the clarity of reading, from here onward, we will refer to the original MIMIC-CXR dataset with 91,544 reports as $MIMIC_{train}$. The 200 reports randomly selected each from the original MIMIC-CXR and Indiana validation sets will be denoted as $MIMIC_{200}$ and $Indiana_{200}$, respectively. Together, these 2 new sets formed our internal test set $Combined_{400}$. The remaining reports from the original MIMIC-CXR and Indiana validation sets will be denoted as $MIMIC_{1800}$ and $Indiana_{1800}$, respectively. We present results on this internal test data under 3 settings (1) results on $MIMIC_{200}$, (2) results on $Indiana_{200}$, and (3) results on the combined internal test dataset, $Combined_{400}$. Most of the following results (Tables 3, 4 & 5) are presented on the internal test dataset. The official results presented in Table 6 are on the official external test data of 600 radiology reports.

5.2 Base Models

We conducted four experiments to get our three base models specific to MIMIC-CXR, Indiana and unknown sources. We utilized $MIMIC_{train}$ to train our first two models, RoBERTa-large$_{(M)}$

and BioRoBERTa$_{(M)}$. We used Indiana$_{1800}$ for the model BioRoBERTa$_{(M+I)}$, and used Indiana$_{1800}$ and MIMIC$_{1800}$ for the model BioRoBERTa$_{(M+M+I)}$. In each setting we split the available dataset into 90/10 for training and validation splits. We evaluated all our models on the internal test set of 400 radiology reports. Each of our models uses a seq2seq architecture with encoder and decoder both composed of Transformer layers. For both encoder and decoder, we inherited the RoBERTa Transformer layer implementations. We also added an encoder-decoder attention mechanism. All models were fine-tuned on the target task using Adam optimizer with a learning rate of 0.05. We used Huggingface's transformers library[4] (Wolf et al., 2019) for executing our experiments. In our encoder-decoder setup, our input was capped at 128, output summary at 40, beam size was 10, our length penalty was set as 0.8. Finally, in our summary generation, trigram and higher length phrases were not repeated.

Table 3 presents results of the 4 experiments. Between the 2 models that were trained using only MIMIC$_{train}$, BioRoBERTa$_{(M)}$ consistently outperform RoBERTa-large$_{(M)}$ in this task, likely due to BioRoBERTa$_{(M)}$ utilizing a domain adapted version of RoBERTa. Among the 3 BioMed-RoBERTa-base based models, BioRoBERTa$_{(M)}$ performs better for MIMIC$_{200}$, and BioRoBERTa$_{(M+I)}$ provides better performance for Indiana$_{200}$. BioRoBERTa$_{(M+M+I)}$ fine-tuned on both MIMIC-CXR and Indiana provides better performance on the Combined$_{400}$ but performs poorly when we consider each source separately.

5.3 Fact-aware Re-ranking (FAR)

For the prediction of the 14-observations-vector we combined MIMIC$_{train}$, MIMIC$_{1800}$, and Indiana$_{1800}$ to form our training and validation splits. Table 4 presents our F_1 scores for our impression 14-observations-vector prediction model evaluated on the internal test dataset Combined$_{400}$. We utilized Smit et al. (2020)'s publicly available implementation[5] to train the domain-specific RoBERTa model (BioMed-RoBERTa-base) for predicting impression 14-observations-vector. In this setup, the transformer architecture was modified with 14 linear heads, corresponding to 14 observations. We concatenate Findings and background of a radiology report to be our input, which is then tokenized and the input is capped at 128. The hidden state of the CLS token is fed as input to each of the linear heads. The model is trained using cross-entropy loss and Adam optimization with a learning rate of 2×10^{-5}. The cross-entropy losses for each of 14 observations are added to produce the final loss. During training, the model was periodically evaluated and the best performing model averaged over 14 observations was saved.

Category	**Macro F_1**	**Micro F_1**
Atelectasis	0.839	0.915
Cardiomegaly	0.803	0.943
Consolidation	0.809	0.973
Edema	0.930	0.963
Enlarged Cardiom.	0.634	0.990
Fracture	0.783	0.988
Lung Opacity	0.848	0.911
Lung Lesion	0.829	0.982
No Finding	0.881	0.881
Pneumonia	0.898	0.950
Pneumothorax	0.939	0.996
Pleural Effusion	0.899	0.950
Pleural Other	0.640	0.990
Support Devices	0.918	0.969
Average	0.832	0.957

Table 4: Impression observations-vector prediction results.

Label	**P**	**R**	**F_1**
MIMIC$_{200}$	0.987	0.993	0.989
Indiana$_{200}$	0.993	0.987	0.990

Table 5: Source-specific classifier results

For fact-aware re-ranking we utilize the model trained above to re-rank the top 10 (N=10 was empirically determined) generated summaries from our three base models presented in section 5.2. Table 3 presents results for our following three factually correct re-ranking experiments, BioRoBERTa$_{(M),FAR}$, BioRoBERTa$_{(M+I),FAR}$, and BioRoBERTa$_{(M+M+I),FAR}$. As BioRoBERTa$_{(M),FAR}$ shows best performance for MIMIC-CXR radiology reports (MIMIC$_{200}$), BioRoBERTa$_{(M+I),FAR}$ exhibits best performance for Indiana radiology reports (Indiana$_{200}$) and the combined BioRoBERTa$_{(M+M+I),FAR}$ shows best performance for the combined test data (Combined$_{400}$), these models are chosen to be our source-specific models for MIMIC-CXR, Indiana and unknown sources respectively.

[4] https://github.com/huggingface/transformers

[5] https://github.com/stanfordmlgroup/CheXbert

Model	R-1	R-2	R-L	CheXbert F_1
Ensemble	0.5252	0.4002	0.5060	0.6823
+post-processing	0.5328	0.4082	0.5134	0.6774

Table 6: Official submission and results.

5.4 Source-specific Ensemble

For training our source-specific classifier we used a downsampled subset of $\text{MIMIC}_{\text{train}}$ of 10,000 radiology reports and Indiana_{1800} and formed 90/10 training and validation splits. We evaluated the model on MIMIC_{200} and Indiana_{200} and present our results in Table 5. We again utilized Huggingface transformers library to conduct our experiments. In this setup, we used the BERT-base architecture with a single linear head for our classification of the source. We concatenate Findings and background for a radiology report to be our input, which is then tokenized and input is capped at 512. The model is trained using cross-entropy loss and Adam optimization with a learning rate of 2×10^{-5}. Our model was trained for 3 epochs.

Utilizing the above model we identify the source of a radiology report and apply the source-specific models. Ensemble results in Table 3 presents our results after we apply source-specific ensembling technique to our internal test dataset. Our ensembled results show a slight drop in performance for individual source MIMIC_{200} and Indiana_{200} (due to classification errors), but show best performance on the combined dataset (Combined_{400}).

5.5 Official Submissions & Results

Table 6 presents our top 2 official submission results. Ensemble presents our best performing source-specific ensemble technique applied to the official test data. In our another submission (Ensemble + post-processing) we remove certain tokens (like "1.", "2.", "__") to clean up our source-specific ensemble technique output which slightly improved the rouge scores.

6 Discussion

In this section, we present two major findings of our approach. First, we find that radiology reports from different sources have distinct language, and fine-tuning a model trained on source A with a small amount of data from source B provides significant gains in performance on source B, allowing the model to be transferable. As it can be seen in Table 3, zero-shot application of our model $\text{BioRoBERTa}_{(\text{M})}$, which is fine-tuned only on MIMIC-CXR ($\text{MIMIC}_{\text{train}}$), shows lower performance on the Indiana dataset. However, on further fine-tuning $\text{BioRoBERTa}_{(\text{M})}$ on a small dataset of 1,800 Indiana reports (Indiana_{1800}) leads to huge gains in performance on Indiana dataset (model $\text{BioRoBERTa}_{(\text{M+I})}$ on Indiana_{200}).

Second, fact-aware re-ranking methodology improves performance of the models on natural language metrics (ROUGE) as well as factual correctness of our predictions, but metrics beyond lexical overlap are needed. As shown in Table 3, models using FAR outperform the base models when measured in ROUGE even through FAR's objective is not to optimize ROUGE. Table 7 shows examples of the most probable predictions from base model compared with the predictions after FAR, and the human-generated ground-truth impressions. ROUGE scores for both predictions compared to the ground-truth are shown at the end of each example. In the first example, FAR chooses a better ROUGE scoring prediction over the most probable prediction by the base model. However, in the second example, FAR doesn't choose the higher ROUGE scoring prediction but rather the more factually correct one. With the current evaluation metric ROUGE, this would lead to a drop in performance. Developing and adopting new metric that consider both lexical as well as factual correctness jointly (Mrabet and Demner-Fushman, 2020) is crucial to steer the research community to develop systems that ensure factual correctness as well as readability.

Limitations and Future Work. We acknowledge several limitations to our work. First, we recognize our dependence on an external structured label generator. As we use CheXbert labels as our proxy for ground truth for training our 14-observations-vector predictor, as well as in our similarity function, any errors in CheXbert have a direct impact on our system's performance. Second, though FAR methodology has shown significant gains in performance in Factual F_1 and ROUGE scores, the system is limited by the generated candidate summaries. We aim to build on this approach by incorporating this methodology during training as a modified version of beam search. Third, all of our presented results are evaluated using a relatively small set of internal test data, due to the limitations on data during the challenge. Though

Base Model's Prediction	Prediction after FAR	Human-generated Impression
No acute cardiothoracic process. R-1: 0	No acute cardiopulmonary process. Tiny right pleural effusion. R-1: 0.6	Tiny right pleural effusion.
No acute cardiopulmonary process. R-1: 0.6	Normal chest radiograph. Mild cardiomegaly. R-1: 0.3	Mild cardiomegaly, new since ___. No acute cardiopulmonary process.

Table 7: Examples depicting the most probable prediction from base model, re-ranked prediction using our FAR methodology compared to the ground truth (human-generated impression).

our approach has translated into similar good performance on the official test data, we aim to further evaluate our approach on an increased test data. Finally, as ROUGE has been shown to be an imperfect metric for radiology report summarization evaluation (Zhang et al., 2019b), we aim to further evaluate our system (1) using other automated metrics such as BERTScore (Zhang et al., 2019a), BLEURT (Sellam et al., 2020), and HOLMS (Mrabet and Demner-Fushman, 2020), (2) by conducting qualitative evaluation of our system's predictions by involving human annotators such as radiologists or subject matter experts.

7 Conclusion

We have presented our system developed during our participation in MEDIQA 2021 RRS challenge. We found that radiology reports from different sources have distinct language and fine-tuning a trained model with a small amount of data from another source leads to gains in performance and allows the models to be transferable. Further, techniques like fact-aware re-ranking, which utilizes a factual vector of the summary to re-rank candidate summaries, not only improves factual correctness of the summary but also improves the performance of the model on the traditional natural language metrics like ROUGE. We have also identified limitations of our work, and discussed promising areas of future research.

References

American College of Radiology. 2020. Acr practice parameter for communication of diagnostic imaging findings. Available at www.acr.org Accessed March 2020.

Yuhao Zhang Chaitanya Shivade Curtis Langlotz Dina Demner-Fushman Asma Ben Abacha, Yassine Mrabet. 2021. Overview of the mediqa 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Jan ML Bosmans, Joost J Weyler, Arthur M De Schepper, and Paul M Parizel. 2011. The radiology report as seen by radiologists and referring clinicians: results of the cover and rover surveys. *Radiology*, 259(1):184–95.

David S Carrell, Robert E Schoen, Daniel A Leffler, Michele Morris, Sherri Rose, Andrew Baer, Seth D Crockett, Rebecca A Gourevitch, Katie M Dean, and Ateev Mehrotra. 2017. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *Journal of the American Medical Informatics Association*, 24(5):986–991.

Neal J. Clinger, Tim B. Hunter, and Bruce J. Hillman. 1988. Radiology reporting: attitudes of referring physicians. *Radiology*, 169(3):825–826.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available

database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W Filice. 2019. Ontology-aware clinical abstractive summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1013–1016.

Yassine Mrabet and Dina Demner-Fushman. 2020. Holms: Alternative summary evaluation with large language models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5679–5688.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

A Wallis and P McCoubrie. 2011. The radiology report—are we getting the message across? *Clinical radiology*, 66(11):1015–1022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to summarize radiology findings. *arXiv preprint arXiv:1809.04698*.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D Manning, and Curtis P Langlotz. 2019b. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. *arXiv preprint arXiv:1911.02541*.