Learning to Summarize Radiology Findings

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian,
Christopher D. Manning, Curtis P. Langlotz
Stanford University
Stanford, CA 94305
{yuhaozhang, dingd, tianpei}@stanford.edu
{manning, langlotz}@stanford.edu

Abstract

The Impression section of a radiology report summarizes crucial radiology findings in natural language and plays a central role in communicating these findings to physicians. However, the process of generating impressions by summarizing findings is time-consuming for radiologists and prone to errors. We propose to automate the generation of radiology impressions with neural sequence-to-sequence learning. We further propose a customized neural model for this task which learns to encode the study background information and use this information to guide the decoding process. On a large dataset of radiology reports collected from actual hospital studies, our model outperforms existing non-neural and neural baselines under the ROUGE metrics. In a blind experiment, a board-certified radiologist indicated that 67% of sampled system summaries are at least as good as the corresponding human-written summaries, suggesting significant clinical validity. To our knowledge our work represents the first attempt in this direction.

Background: history: swelling; pain. technique: 3 views of the left ankle were acquired. comparison: no prior study available.

Findings: there is normal mineralization and alignment. no fracture or osseous lesion is identified. the ankle mortise and hindfoot joint spaces are maintained. there is no joint effusion. the soft tissues are normal.

Human Impression:
normal left ankle radiographs.

Extractive Baseline:
there is no joint effusion.

Pointer-Generator:
normal right ankle.

Our model:
normal radiographs of the left ankle.

Figure 1: An example radiology report with study background information organized into a **Background** Section, and radiology findings in a **Findings** Section. The human-written summary (or impression) and predicted summaries from different models are also shown. The extractive baseline does not summarize well, the baseline pointer-generator model generates spurious sequence, while our model gives correct summary by incorporating the background information.

1 Introduction

The radiology report documents and communicates crucial findings in a radiology study. As shown in Figure 1, a standard radiology report usually consists of a Background section that describes the exam and patient information, a Findings section, and an Impression section (Kahn Jr et al., 2009). In a typical workflow, a radiologist first dictates the detailed findings into the report, and then summarizes the salient findings into the more concise Impression section based also on the condition of the patient.

The impressions are the most significant part of a radiology report that communicate the findings. Previous studies have shown that over 50% of referring physicians read only the impression statements in a report (Lafortune et al., 1988; Bosmans et al., 2011). Despite its importance, the generation of the impression statements is error-prone. For example, crucial findings may be forgotten, which would cause significant miscommunications (Gershanik et al., 2011). Additionally, the process of writing the impression statements is time-consuming and highly repetitive with the dictation of the findings. This suggests a crucial need to automate the radiology impression generation process.

In this work, we propose to automate the generation of radiology impressions with neural sequence-to-sequence learning. In particular, we argue that this task could be viewed as a text summarization problem, where the source sequence is the radiology findings and the target sequence the

Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pages 204–213
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics

impression statements. We collect a dataset of radiology reports from actual hospital radiographic studies, and find that this task involves both *extractive summarization* where descriptions of radiology observations can be taken directly from the findings, and *abstractive summarization* where new words and phrases, such as conclusions of the study, need to be generated from scratch. We empirically evaluate existing popular summarization systems on this task and find that, while existing neural models such as the pointer-generater network can generate plausible summaries, they sometimes fail to model the study background information and thus generate spurious results. To solve this problem, we propose a customized summarization model that properly encodes the study background information and uses the encoded information to guide the decoding process.

We show that our model outperforms existing non-neural and neural baselines on our dataset measured by the standard ROUGE metrics. Moreover, in a blind experiment, a board-certified radiologist indicated that 67% of sampled system summaries are at least as good as the reference summaries written by well-trained radiologists, suggesting significant clinical validity of the resulting system. We further show through detailed analysis that our model could be reliably transferred to radiology reports from another organization, and that the model can sometimes summarize radiology studies for body parts unseen during training.

To review, our main contributions in this paper include: (i) we propose to summarize radiology findings into impression statements with neural sequence-to-sequence learning, and to our knowledge our work represents the first attempt in this direction; (ii) we propose a new customized summarization model to this task that improves over existing methods by better leveraging study background information; (iii) we further show via a radiologist evaluation that the summaries generated by our model have significant clinical validity.

2 Related Work

Early Summarization Systems. Early work on summarization systems mainly focused on extractive approaches, where the summaries are generated by scoring and selecting sentences from the input. Luhn (1958) proposed to represent the input by topic words and score each sentence by the amount of topic words it contains. Kupiec et al. (1995) scored sentences with a feature-based statistical classifier. Steinberger and Jezek (2004) applied latent semantic analysis to cluster the topics and then select sentences that cover the most topics. Meanwhile, various graph-based methods, such as the LexRank (Mihalcea and Tarau, 2004) and the TextRank algorithm (Erkan and Radev, 2004), were applied to model sentence dependency by representing sentences as vertices and similarities as edges. Sentences are then scored and selected via modeling of the graph properties.

Neural Summarization Systems. Summarization systems based on neural network models enable abstractive summarization, where new words and phrases are generated to form the summaries. Rush et al. (2015) first applied an attention-based neural encoder and a neural language model decoder to this task. Nallapati et al. (2016) used recurrent neural networks for both the encoder and the decoder. To address the limitation that neural models with a fixed vocabulary cannot handle out-of-vocabulary words, a pointer-generator model was proposed which uses an attention mechanism that copies elements directly from the input (Nallapati et al., 2016; Merity et al., 2017; See et al., 2017). See et al. (2017) further proposed a coverage mechanism to address the repetition problem in the generated summaries. Paulus et al. (2018) applied reinforcement learning to summarization and more recently, Chen and Bansal (2018) obtained improved result with a model that first selects sentences and then rewrites them.

Summarization of Radiology Reports. Most prior work that attempts to "summarize" radiology reports focused on classifying and extracting information from the report text (Friedman et al., 1995; Hripcsak et al., 1998; Elkins et al., 2000; Hripcsak et al., 2002). More recently, Hassanpour and Langlotz (2016) studied extracting named entities from multi-institutional radiology reports using traditional feature-based classifiers. Goff and Loehfelm (2018) built an NLP pipeline to identify asserted and negated disease entities in the impression section of radiology reports as a step towards report summarization. Cornegruta et al. (2016) proposed to use a recurrent neural network architecture to model radiological language in solving the medical named entity recognition and negation detection tasks on radiology reports. To our knowledge, our work represents the first attempt at automatic summarization of radiology findings into natural language impression statements.

3 Task Definition

We now give a formal definition of the task of summarizing radiology findings. Given a passage of findings represented as a sequence of tokens $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, with N being the length of the findings, our goal is to find a sequence of tokens $\mathbf{y} = \{y_1, y_2, \ldots, y_L\}$ that best summarizes the salient and clinically significant findings in $\mathbf{x}$, with L being an arbitrary length of the summary.[1] Note that the mapping between $\mathbf{x}$ and $\mathbf{y}$ can either be modeled in an unsupervised way (as done in unsupervised summarization systems), or be learned from a dataset of findings-summary pairs.

4 Models

In this section we introduce our model for the task of summarizing radiology findings. As our model builds on top of existing work on neural sequence-to-sequence learning and the pointer-generator model, we start by introducing them.

4.1 Neural Sequence-to-Sequence Model

At a high-level, our model implements the summarization task with an encoder-decoder architecture, where the encoder learns hidden state representations of the input, and the decoder decodes the input representations into an output sequence.

For the encoder, we use a Bi-directional Long Short-Term Memory (Bi-LSTM) network. Given the findings sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, we encode $\mathbf{x}$ into hidden state vectors with:

$$\mathbf{h} = \text{Bi-LSTM}(\mathbf{x}), \tag{1}$$

where $\mathbf{h} = \{h_1, h_2, \ldots, h_N\}$. Here h_N combines the last hidden states from both directions in the encoder.

After the entire input sequence is encoded, we generate the output sequence step by step with a separate LSTM decoder. Formally, at the t-th step, given the previously generated token y_{t-1} and the previous decoder state s_{t-1}, the decoder calculates the current state s_t with:

$$s_t = \text{LSTM}(s_{t-1}, y_{t-1}). \tag{2}$$

We then use s_t to predict the output word. For the initial decoder state we set $s_0 = h_N$.

The vanilla sequence-to-sequence model that uses only s_t to predict the output word has a major limitation: it generates the entire output sequence based solely on a vector representation of the input (i.e., h_N), which may result in significant information loss. For better decoding we therefore employ the attention mechanism (Bahdanau et al., 2015; Luong et al., 2015), which uses a weighted sum of all input states at every decoding step.

Given the decoder state s_t and an input hidden state h_i, we calculate an input distribution a^t as:

$$e_i^t = v^\top \tanh(W_h h_i + W_s s_t), \tag{3}$$

$$a^t = \text{softmax}(e^t), \tag{4}$$

where W_h, W_s and v are learnable parameters.[2] We then calculate a weighted input vector as:

$$h_t^* = \sum_i a_i^t h_i. \tag{5}$$

h_t^* encodes the salient input information that is useful at decoding step t. Lastly, we obtain the output vocabulary distribution at step t as:

$$P(y_t|\mathbf{x}, y_{<t}) = \text{softmax}(V' \tanh(V[s_t; h_t^*])), \tag{6}$$

where V' and V are learnable parameters.

4.2 Pointer-Generator Network

While the encoder-decoder framework described above can generate impressions from a fixed vocabulary, the model can clearly benefit from being able to "copy" salient observations directly from the input findings. To add such "copying" capacity into the model, we use a pointer-generator network similar to the one described in See et al. (2017).

The main idea is that at each decoding step t, we allow the model to either generate a word from the vocabulary with a generation probability p_{gen}, or copy a word directly from the input sequence with probability $1 - p_{\text{gen}}$. We model p_{gen} as:

$$p_{\text{gen}} = \sigma(w_{h^*}^\top h_t^* + w_s^\top s_t + w_y y_{t-1}), \tag{7}$$

where y_{t-1} denotes the previous decoder output, w_{h^*}, w_s and w_y learnable parameters and σ a sigmoid function. For the copy distribution, we reuse the attention distribution a^t calculated in (4). Therefore, the overall output distribution in the pointer-generator network is:

$$P(y_t|\mathbf{x}, y_{<t}) = p_{\text{gen}} P_{\text{vocab}}(y_t) + (1 - p_{\text{gen}}) \sum_{i:x_i = y_t} a_i^t, \tag{8}$$

[1] While the name "impression" is often used in clinical settings, we use "summary" and "impression" interchangably.

[2] For clarity we leave out the bias terms in all linear layers.

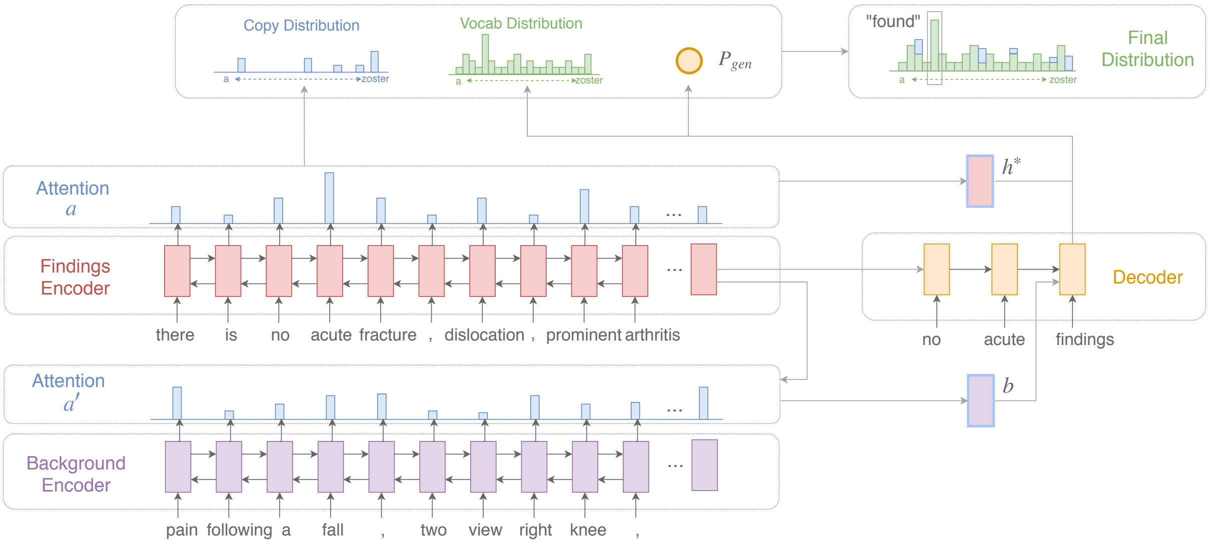

Figure 2: Overall architecture of our summarization model.

where $P_{\text{vocab}}(y_t)$ is the same as the output distribution in (6).

4.3 Incorporating Study Background Information

The background part of a radiology report is also important, since crucial information such as the purpose of the study, the body part involved and the condition of the patient are often mentioned only in the background. A straightforward way of incorporating the background information is to prepend all the background text to the findings, and treat the entire sequence as input to the pointer-generator network. However, as we show in Section 6, this naive method in fact hurts the summarization quality, presumably because the model cannot sufficiently distinguish between the findings and the background information, which as a result leads to insufficient modeling of both the findings and the background. To solve this, we propose to encode the background text with a separate attentional encoder, and use the resulting background representation to guide the decoding process in the summarization model (Figure 2).

For clarity we now use $\mathbf{x}^b$ to denote the background token sequence, and $\mathbf{x}$ to denote the actual findings section. Our goal is then to find $\mathbf{y}$ that maximizes $P(\mathbf{y}|\mathbf{x}, \mathbf{x}^b)$. To do this, we again obtain the hidden state vectors $\mathbf{h}$ of the findings section as in (1). Similarly, we obtain the hidden state vectors of the background text with $\mathbf{x}^b$ as input using a separate Bi-LSTM encoder:

$$\mathbf{h}^b = \text{Bi-LSTM}^b(\mathbf{x}^b). \tag{9}$$

Next, we calculate a distribution over $\mathbf{h}^b$ as:

$$e'_i = v'^{\top} \tanh(W_b h^b_i + W_h h_N), \tag{10}$$

$$a' = \text{softmax}(e'), \tag{11}$$

where W_b and W_h are learnable parameters and h_N the last hidden state of the findings encoder. The distribution a' models the importance of tokens in the background section. We then obtain a weighted representation of the background text as:

$$b = \sum_i a'_i h^b_i, \tag{12}$$

where vector b has the same size as h^b, and encodes the salient background information.

Lastly, we use the background vector b to guide the decoding process, by modifying the recurrent kernel of the decoder LSTM in (2) to be:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ u_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} W \cdot \begin{bmatrix} s_{t-1} \\ y_{t-1} \\ b \end{bmatrix}, \tag{13}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot u_t, \tag{14}$$

$$s_t = o_t \cdot \tanh(c_t), \tag{15}$$

where i_t, f_t, o_t denotes the input, forget, and output gates, W the weight matrix and c_t the internal cell of the LSTM repectively, and $\cdot$ represents an element-wise multiplication. Again for clarity we leave out the bias terms in (13). As a result, every state in the decoding process is directly influenced by the information encoded by the background vector b. The rest of the model, including

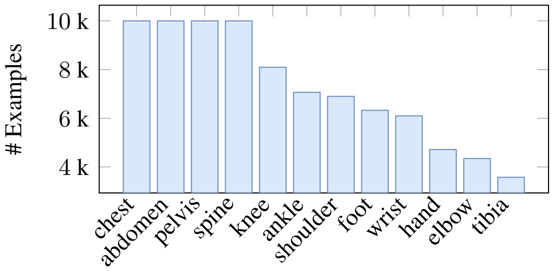

Figure 3: Number of examples split by body part in the collected Stanford Hospital dataset.

the calculation of the vocabulary distribution and the copy distribution, remains the same.

5 Experiments

To test the effectiveness of our summarization model, we collected reports of radiographic studies from the picture archiving and communication system (PACS) at the Stanford Hospital. We describe our data collection process, baseline models and experimental setup in this section, and present the results and discussions in Section 6.

5.1 Data Collection

Reports of all radiographic studies from 2000 to 2014 were collected. We first tokenized all reports with Stanford CoreNLP (Manning et al., 2014), and filtered the dataset by excluding reports where (1) no findings or impression section can be found; (2) multiple findings or impression sections can be found but cannot be aligned; or (3) the findings have fewer than 10 words or the impression has fewer than 2 words.

We removed body parts where only a small number of cases are available, and included reports of the top 12 body parts in the PACS system to maintain generalizability. For common body parts with more than 10k reports (e.g., chest), we subsampled 10k reports from them.

This results in a dataset with a total of 87,127 reports. We further randomly split the dataset into a 70% training (60,990), a 10% development (8,712) and a 20% test set (17,425). We show the dataset statistics split by body part in Figure 3.

5.2 Baseline Models

For our main experiments, we compare our model against several competitive non-neural and neural systems on the collected dataset. Unless otherwise stated, the baseline models take only the findings section as input.[3]

S&J-LSA. This is an extractive approach described by Steinberger and Jezek (2004), which applies Latent Semantic Analysis (LSA) to summarization. It first scores "concept" clusters by applying singular value decomposition to the term-by-sentence co-occurence matrix derived from the passage. Sentences with the top scored concepts are then kept as the summaries.

LexRank. LexRank is another popular extractive model introduced by Erkan and Radev (2004). In LexRank, a passage is represented as a graph of sentences, and a connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph. Sentences are scored by the eigenvector centrality in the graph, and the highest scored sentences are kept.

Pointer-Generator. We also run the baseline pointer-generator model introduced by See et al. (2017). We find the "coverage" mechanism described in the paper did not improve summary quality in our task and therefore did not use it for simplicity. We compare our model with two versions of the pointer-generator model: one with only the findings section as input and another one with the background sections prepended to the findings section as input.

5.3 Experimental Setup

Evaluation Metrics. In our main experiments we evaluate the models with the widely-used ROUGE metric (Lin, 2004). We report the F_1 scores for ROUGE-1, ROUGE-2 and ROUGE-L, which measure the word-level unigram-overlap, bigram-overlap and the longest common sequence between the reference summary and the system predicted summary respectively.

Word Vectors. To enable knowledge transfer from a larger corpus, we applied the GloVe algorithm (Pennington et al., 2014) to a corpus of 4.5 million radiology reports of all modalities (e.g., X-ray, CT) and body parts. We used the resulting 100-dimensional word vectors to initialize all word embedding layers in our neural models, and empirically found this to improve the performance of our neural models by about 1 ROUGE-L score.

[3] We find that when the background section is prepended to the input, the extractive baseline models may select sentences in the background part as the summary, resulting in deteriorated performance.

System	ROUGE-1	ROUGE-2	ROUGE-L
Extractive Baseline: S&J-LSA	29.39	16.27	27.38
Extractive Baseline: LexRank	30.48	17.09	28.49
Pointer-Generator	46.51	33.39	45.07
Pointer-Generator ($\oplus$ Background)	45.39	32.60	44.05
Our model	**48.56**	**35.25**	**47.06**

Table 1: Main results on the test set of the Stanford reports. "$\oplus$ Background" represents prepending the background section to the findings section to form the input to the model. All the ROUGE scores have a 95% confidence interval of at most ± 0.50 as calculated by the official ROUGE script.

Implementations & Model Details. For the two non-neural extractive baselines, we use their open implementations.[4] For both of them, we select the top N scored sentences to form the summary and treat N as a hyperparameter. We use $N = 3$ in our experiments as it yields best scores on the dev set. We implemented all neural models with PyTorch.[5] To train the neural models we append a special `<EOS>` token to the end of every reference summary. We then employ the standard teacher-forcing with the reference summaries and optimize the negative log-likelihood loss using the Adam optimizer (Kingma and Ba, 2015). We tune all hyperparameters on the dev set. We use 2-layer Bi-LSTM for all encoders, and set the hidden size to be 100 for each direction; 1-layer LSTM for the decoder and set the hidden size to be 200. During inference, we employ the standard beam search with a beam size of 5. We stop decoding whenever a `<EOS>` token is predicted, and otherwise use a maximum output sequence length of 100.

6 Results & Analysis

6.1 Main Results

We present results of our main experiments in Table 1. We find that the two non-neural extractive models perform comparably, and both are able to obtain non-trivial subsequence overlap with the reference summaries as measured by ROUGE scores. However, a baseline neural pointer-generator that combines the sequence generation and the copy mechanism beats the non-neural baselines substantially on all metrics. We confirm that naively incorporating the study background information by prepending the background section directly to the input findings in the pointer-generator model in fact hurts the performance (noted by $\oplus$ Background). In comparison, our model benefits from using the separately encoded background vector to guide the decoding process, and achieves best scores on all ROUGE metrics.

We also present sampled test examples and system output in Figure 4. We find that compared to the non-neural extractive baselines, the neural models are not limited by sentences in the findings section and therefore generate summaries of better quality. For example, the neural models learn to compose the summary by combining observation phrases from different sentences, or by generating new conclusive phrases such as "negative study". Compared to the pointer-generator model, our model learns to correctly utilize relevant background information (e.g., previous study or exam information) to improve the summary.

6.2 Clinical Validity with Radiologist Evaluation

One potential shortcoming of the ROUGE metrics is that they only measure the similarity between the predicted summary and the reference summary, but do not sufficiently reflect the overall grammaticality or utility of the predictions. Therefore, we also conducted evaluations with a board-certified radiologist to understand the clinical validity of our system generated summaries.

In this evaluation, we randomly sampled 100 examples from our test set. We ran our best model over these 100 examples, and presented each example along with the corresponding system predicted summary and reference human-written summary to the radiologist. We randomly ordered the predicted and reference summary such that the correspondence cannot be guessed from the order. The radiologist was asked to select which of the two summaries was better, or that they have roughly equal quality.

Table 2 presents the result. For 51 examples, the

[4] https://github.com/miso-belica/sumy

[5] https://pytorch.org/

Background: radiographic examination of the abdomen. clinical history: xx years of age, male, please obtain upright and lateral decub. comparison: abdominal x-ray <date>. procedure comments: two views of the abdomen. **Findings:** median sternotomy wires are seen in the anterior chest wall in addition to several mediastinal clips and an aicd. trace bilateral pleural effusions are noted. interval increase in small bowel dilatation compared to previous study with multiple air-fluid levels, consistent with small bowel obstruction. there is a paucity of colonic gas. no pneumoperitoneum.	**Background:** three views of the right shoulder and three views of the left shoulder: <date>. clinical history: an xx-year-old female with bilateral shoulder pain. **Findings:** three views of the right shoulder consisting of external rotation, axillary, and scapular views demonstrate no evidence of fracture or dislocation. the joint spaces are well-maintained without evidence of degenerative change. there is normal mineralization throughout. three views of the left shoulder . . . are well-maintained without evidence of degenerative change. mineralization is normal throughout.	**Background:** three views of the abdomen: <date>. comparison: <date>. clinical history: a xx-year-old male status post hirschsprung's disease repair. **Findings:** the supine, left-sided decubitus and erect two views of the abdomen show increased dilatation of the small bowel since the prior exam on <date>. there are multiple air-fluid levels, suggesting bowel obstruction. no free intraperitoneal gas is present.
Human: small bowel dilatation with multiple air-fluid levels and colonic decompression consistent with small bowel obstruction.	**Human:** unremarkable radiographs of bilateral shoulders.	**Human:** increased dilatation of the small bowel with multiple air-fluid levels, suggesting bowel obstruction. no free intraperitoneal gas.
Extractive Baseline: median sternotomy wires are seen in the anterior chest wall in addition to several mediastinal clips and an aicd.	**Extractive Baseline:** three views of the right shoulder consisting of external rotation, axillary, and scapular views demonstrate no evidence of fracture or dislocation.	**Extractive Baseline:** the supine, left sided decubitus and erect two views of the abdomen show increased dilatation of the small bowel since the prior exam on <data>.
Pointer-Generator: interval increase in bowel dilatation, consistent with bowel obstruction.	**Pointer-Generator:** no evidence of fracture or dislocation of the right shoulder.	**Pointer-Generator:** increased dilatation of small bowel, suggesting small bowel obstruction.
Our model: interval increase in small bowel dilatation compared to abdominal x-ray dated <date> with multiple air-fluid levels, consistent with small bowel obstruction.	**Our model:** unremarkable bilateral shoulders.	**Our model:** increased dilatation of small bowel, suggesting bowel obstruction. no free intraperitoneal gas.

Figure 4: Sampled test examples and system predictions from the Stanford dataset. First example: our model learns to relate the summary with a previous study mentioned only in the background section. Second: our model correctly summarizes the body part involved in the study. Third: our model correctly includes more crucial information as found in the human summary.

Category	Percentage
Human Summary Wins	33
System Prediction Wins	16
Roughly Equal Quality	51

Table 2: Radiologist evaluation result on 100 sampled test examples. For a total of 67 examples, the radiologist indicated that the system summary is at least as good as the human-written summary.

System	ROUGE-1	ROUGE-2	ROUGE-L
LexRank	15.42	5.65	14.60
Our model	35.02	20.79	34.56

Table 3: Cross-organization evaluation results on the Indiana University chest x-ray dataset. All the ROUGE scores have a 95% confidence interval of at most ±1.10 as calculated by the official ROUGE script.

radiologist indicated that the human-written and system-generated summaries are equivalent. For 16 examples, the radiologist preferred the system summary, and for the remaining 33 examples, the radiologist preferred the human-written summary. Note that under our setting, a randomly generated sequence would have almost zero chance to be indicated as good as the human-written summary. We therefore believe the result suggests significant clinical validity of our system.

6.3 Does the model transfer to reports from another organization?

Deploying a clinical NLP system at an organization different from the one where the training data comes from is a common need. However, this is challenging in that medical practitioners including radiologists from different organizations tend to go through different training and follow different templates or styles when writing medical text. Here we aim to understand the cross-organization transferability of our summarization model.

We use the publicly available Indiana University Chest X-ray Dataset (Demner-Fushman et al., 2015), which consists of chest X-ray images paired with the corresponding radiology reports. We filtered the reports with the same set of rules and arrived at a collection of 2,691 unique reports. We used this dataset as the test set, and ran our best model trained on our own dataset directly on it. The results are shown in Table 3 and sampled examples are shown in the first two columns of Figure 5. We find that our model again outperforms the baseline extractive model substantially in this transfer setting, and the generated summaries are both grammatical and clinically meaningful.

Cross-organization	Cross-organization	Cross-body part: Knee
Background: indication: xxxx year old male with end-stage renal disease on hemodialysis **Findings:** the heart size is mildly enlarged. there is tortuosity of the thoracic aorta. no focal airspace consolidation, pleural effusions or pneumothorax. no acute bony abnormalities.	**Background:** indication: xxxx year old female, hypoxia. comparison: pa lateral views of the chest dated xxxx. **Findings:** bilateral emphysematous again noted and lower lobe fibrotic changes. postsurgical changes of the chest including cabg procedure, stable. stable valve artifact. there are no focal areas of consolidation. no large pleural effusions. no evidence of pneumothorax. . . . contour abnormality of the posterior aspect of the right 7th rib again noted, stable.	**Background:** radiographic examination of the knee: <date> <time>. clinical history: xx-year-old man with right knee pain. comparison: none. procedure comments: 2 views of the right knee were performed. **Findings:** there is no visible fracture or malalignment. likely small joint effusion. mild fullness in the popliteal region of the right knee may represent a baker 's cyst. mild soft tissue swelling along the medial aspect of the knee is present.
Human: cardiomegaly without acute pulmonary findings.	**Human:** no acute cardiopulmonary abnormality. stable bilateral emphysematous and lower lobe fibrotic changes.	**Human:** no acute bony abnormality. likely joint effusion and soft tissue swelling along the medial aspect of the knee.
Our model: mild cardiomegaly. no radiographic evidence of acute cardiopulmonary process.	**Our model:** stable postsurgical changes of the chest as described above. no evidence of pneumothorax.	**Our model:** mild soft tissue swelling along the medial aspect of the knee. no fracture or malalignment.

Figure 5: First two columns: sampled examples from the Indiana University dataset and system output in the cross-organization evaluation. Last column: sampled test example of a "knee" study in our cross-body part evaluation.

Body Part	ROUGE-1	ROUGE-2	ROUGE-L
Chest	31.24	17.99	30.38
Abdomen	28.90	17.23	27.83
Knee	48.78	35.07	47.49

Table 4: Cross-body part evaluation results of our neural model on the Stanford dataset. All the ROUGE scores have a 95% confidence interval of at most ±0.75 as calculated by the official ROUGE script.

Category	Percentage
Good Summary	63
Missing Critical Info.	24
Inaccurate/Spurious Info.	8
Redundant	4
Ungrammatical	6

Table 5: Error analysis on 100 sampled dev examples from the Stanford dataset.

6.4 Does the model transfer to body parts unseen during training?

Radiology studies conducted on different body parts often include vastly different observations and diagnosis. For example, while "lung base opacity" is a common observation in chest radiographic studies, it does not exist in musculoskeletal studies. In practice, an organization may not have adequate report data that covers some rare body parts. It is therefore interesting to test to what extent our summarization model can generalize to reports for body parts unseen during training.

We study this by simulating the condition where a specific body part is not present in the training data. Given the entire dataset $\mathcal{D}$, and a subset of the dataset $\mathcal{D}_B$ that corresponds to a body part B, we reserved the entire subset $\mathcal{D}_B$ as test data, and used $\mathcal{D} - \mathcal{D}_B$ for training (90%) and validation (10%). Table 4 presents the evaluation results for body part "chest", "abdomen" and "knee". We find that for "chest" and "abdomen", the system summaries degrade substantially when the corresponding data were not seen during training. However, the predicted summaries degrade less for "knee" when reports of it were not seen during training, presumably because the model can learn to summarize reasonably well from reports of other close musculoskeletal studies such as "ankle" or "elbow" studies. We confirm this by examining the model predictions: in the example shown in the last column of Figure 5, the model learns to compose the summary with salient observations such as "tissue swelling" and "fracture", while being able to copy the anatomy "knee" (unseen during training) from the findings section.

6.5 What is the model missing on?

Lastly, we run a detailed error analysis on 100 sampled dev examples. We focus on four types of errors: (1) missing critical information, if the predicted summary fails to include some clinically important information; (2) inaccuate/spurious information, if the predicted summary contains observations or conclusions that are inaccurate, or that do not exist in the findings; (3) redundant summary, if the predicted summary is repetitive or over-verbose; and (4) ungrammatical summary, if the predicted summary contains significant gram-

Error type: missing critical information	Error type: redundant summary	Error type: ungrammatical summary
Background: radiographic examination of the lumbar spine: <time>. clinical history: <age>, lower back pain. comparison: none. procedure comments: 4 views of the lumbar spine. **Findings:** five non-rib bearing lumbar type vertebral bodies are present. there is trace retrolisthesis of l5 on s1. there is no evidence of instability on flexion and extension views. the spinal alignment is otherwise normal. the disc spaces and vertebral body heights are preserved. there is no visible fracture. no visible facet joint arthropathy or pars defects.	**Background:** radiographic examination of the shoulder: <time>. clinical history: <age> years of age, pain in joint involving shoulder region. comparison: outside study dated <date>. procedure comments: single axillary view of the left shoulder. **Findings:** single axillary view of the shoulder again demonstrates a highly comminuted fracture of the humeral head and likely fracture of the scapular body. the humeral head appears located on the glenoid.	**Background:** radiographic examination of the shoulder: <time>. clinical history: <age> years of age, xray exam of lower spine 2 or 3 views. x-ray exam of right shoulder complete. comparison: none. procedure comments: three views of the right shoulder. **Findings:** a calcification of the rotator cuff is seen above the greater tuberosity. there is no fracture or malalignment. the soft tissues and visualized lung are unremarkable.
Human: trace retrolisthesis of l5 on s1 with no evidence of instability with motion. otherwise normal lumbar spine.	**Human:** redemonstration of a highly comminuted fracture of the humeral head and likely fracture of the scapular body . the humeral head appears to be located on the glenoid .	**Human:** no acute bony or joint abnormality, but there is calcification of the rotator cuff that may be due to calcific tendinitis.
Our model: no acute bony or articular abnormality.	**Our model:** highly comminuted fracture of the scapular body and likely fracture of the scapular body.	**Our model:** calcification acute bony or joint abnormality.

Figure 6: Examples of different types of errors that our system makes on the Standord dataset. Words that are missing from or are erroneously included in the model predictions are highlighted in red.

matical errors. For each example, we examine whether it contains any of the errors by comparing it with the reference summary; otherwise we classify it as a good summary. Note that an example can be assigned to more than one error categories.

We include examples of different error types in Figure 6, and present the result of error analysis in Table 5. We find that 63% examples are qualitatively close to the reference summary, which aligns well with the radiologist evaluation result. Among the four error categories, missing critical information is the most common error with 24% examples, suggesting that the summaries may be improved with explicit modeling of the importance of different radiology findings. We also find through qualitative analysis that the model tends to miss on followup procedures recommended by the human radiologist, since these procedures are often not included in the findings section and generating them needs significant understanding of the study and domain knowledge.

7 Conclusion

In this paper we proposed to generate radiology impressions from findings via neural sequence-to-sequence learning. We proposed a customized neural model for this task which uses encoded background information to guide the decoding process. We collected a dataset from actual hospital studies and showed that our model not only outperforms non-neural and neural baselines, but also generates summaries with significant clinical validity and cross-organization transferability.

Acknowledgments

We thank Peng Qi and the anonymous reviewers for their helpful suggestions.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *The 2015 International Conference on Learning Representations*.

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):184–195.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *The 2018 Annual Meeting of the Association of Computational Linguistics (ACL 2018)*.

Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. 2016. Modelling radiological language with bidirectional long short-term memory networks. *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI)*.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Jacob S Elkins, Carol Friedman, Bernadette Boden-Albala, Ralph L Sacco, and George Hripcsak. 2000.

Coding neuroradiology reports for the northern manhattan stroke study: a comparison of natural language processing and manual review. *Computers and Biomedical Research*, 33(1):1–10.

Günes Erkan and Dragomir R Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Carol Friedman, George Hripcsak, William DuMouchel, Stephen B Johnson, and Paul D Clayton. 1995. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(1):83–108.

Esteban F Gershanik, Ronilda Lacson, and Ramin Khorasani. 2011. Critical finding capture in the impression section of radiology reports. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

Daniel J Goff and Thomas W Loehfelm. 2018. Automated radiology report summarization using an open-source natural language processing pipeline. *Journal of Digital Imaging*, 31(2):185–192.

Saeed Hassanpour and Curtis P Langlotz. 2016. Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine*, 66:29–39.

George Hripcsak, John HM Austin, Philip O Alderson, and Carol Friedman. 2002. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology*, 224(1):157–163.

George Hripcsak, Gilad J Kuperman, and Carol Friedman. 1998. Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods of Information in Medicine*, 37(01):01–07.

Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *The 2015 International Conference for Learning Representations*.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*.

M Lafortune, G Breton, and JL Baudouin. 1988. The radiological report: What is useful for the referring physician? *Canadian Association of Radiologists*, 39(2):140–143.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out: ACL Workshop*.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. *The 2017 International Conference on Learning Representations*.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. *The SIGNLL Conference on Computational Natural Language Learning (CoNLL), 2016*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. *The 2018 International Conference on Learning Representations*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *The 2017 Annual Meeting of the Association of Computational Linguistics (ACL 2017)*.

Josef Steinberger and Karel Jezek. 2004. Using latent semantic analysis in text summarization and summary evaluation. *Proceedings of the 2004 International Conference on Information System Implementation and Modeling*.