Customized Impression Prediction from Radiology Reports Using BERT and LSTMs

Batuhan Gundogdu†, Utku Pamuksuz†, Jonathan H. Chung,
Jessica M. Telleria, Peng Liu, Farrukh Khan, Paul J. Chang

***Abstract*—Clinical language processing has become an attractive field with the improvements of deep learning applications and the abundance of large unstructured narratives in the healthcare records. The capability to extract unstructured information from raw text to provide actionable information for healthcare personnel plays a vital role in healthcare workflows. In this study, we introduce a deep learning approach to automate the generation of radiology impressions by analyzing radiology findings and patient background information of each examination. Since the impression section of a radiology report is an essential conclusion, any errors can prove to be detrimental. Thus, we developed a deep learning system to prevent important clinical findings from being overlooked by using almost 1 million de-identified radiology reports obtained from the University of Chicago Medicine over the last twelve years. We propose to automate the generation of radiology reports by incorporating sequence to sequence neural network models with the power of Bidirectional Encoder Representations from Transformers (BERT). We tested our model in a real-time experimental setup with radiologists in a top tier academic institution and statistically validated the performance by using ROUGE metrics. Clinical validations have shown that 76 percent of our predictions are at least as accurate as human-generated impressions by radiologists. Furthermore, statistical validation metrics demonstrated higher ROUGE scores compared to previously published studies over two different test sets.**

***Impact Statement*—The aim of this paper is twofold: 1- Radiologists read as many as 100 exams in one day, therefore ensuring that important findings are not overlooked, while saving time in writing the radiology report and reducing burnout could prove invaluable. The "impression" section of a radiology report is the most important section of the radiology report, it is based on the radiologist's observation of the image that is documented in the "findings" section and is considered the conclusion of the study. Hence, we developed a deep learning system to auto-generate the impression. This was done by making use of large-scale and high quality de-identified reports in system development. Our approach demonstrated strong validations by domain expert practitioners. 2- By integrating an AI-based real-time prediction system, we monitored a 20-25 percent improvement in throughput, more exams can be studied within the same amount of time while projecting a significant reduction in burnout.**

***Index Terms*—BERT, Clinical Language Processing, Deep Learning, LSTM, Neural Networks, Impression, Radiology**

†Batuhan Gundogdu and Utku Pamuksuz contributed equally.

Batuhan Gundogdu, Jonathan H. Chung, Jessica M. Telleria, Peng Liu and Paul J. Chang are with the University of Chicago School of Medicine.

Utku Pamuksuz is with the University of Illinois at Urbana-Champaign and affiliated with Inference Analytics Inc.

Farrukh Khan is with Inference Analytics Inc.

Correspondence e-mail: gundogdu@uchicago.edu

I. Introduction

MAJORITY of the machine learning methods in healthcare applications intend to reduce the healthcare workers' time for analyses, in return allowing a larger number of people access to care. Electronic health records and vast amounts of clinical notes provide large scale unstructured data to extract meaningful insights from clinical narratives. Among such applications, tracking care and revenue cycle management (i.e. ICD-10 code prediction), readmission risk predictions from discharge notes, and annotation of clinical radiology reports [1] could be considered. Machine learning-based approaches provide useful methods for structured data including both disease-centered and patient-centered models. These models include patients' clinical status predictions, identification of prescription needs, disease progression detection, and patient treatment pipelines, even in the event of insufficient medical information [2]. However, majority of the medical information in clinical work cycles are in unstructured text format, and cannot be directly used in common analyses tools. Despite the inherent complexities of de-identified clinical data curation and acquisition, there is potential and benefit in developing and deploying machine-learning based unstructured data solutions. Radiology examinations are important data sources frequently used for diagnosis, therapy assessment, and planning [3]. Medical images such as X-ray radiography, ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI) are analyzed by radiologists . Their FINDINGS are then typically dictated to compose an unstructured text. These customized notes are the crucial part of the clinical process. These detailed clinical findings are later processed, summarized, and concluded by practitioners into a concise section called IMPRESSION. The impression section is typically the most utilized information by the referring physicians when analyzing a patient's clinical status and history. More specifically, according to [4], the IMPRESSION section is the most important part of the radiology report and majority of the physicians solely consult to this section for patient analysis. An example of a FINDINGS-IMPRESSION pair, along with an accompanying BACKGROUND information about a patient is provided in Table I.

Automated prediction of complete, readable, and accurate radiologist IMPRESSIONS from the unstructured FINDINGS, while also making use of some clinical information about the patient, is in increasing demand. An increase in quantity and

speed of information in healthcare, due to the novel Covid-19 pandemic, has highlighted machine learning and natural language processing algorithms as more than tools of a distant future. Such auto-generation and prediction of clinical data allows practitioners to spend less time familiarizing themselves with patients by curtailing lengthy and detailed descriptions to succinct and precise summaries. More importantly, these predicted IMPRESSIONS will prevent important clinical FINDINGS from falling through the cracks, reduce human error stemming from physician burnout, and mitigate the substantial risk of overlooking a relevant actionable finding. The desired task of IMPRESSION prediction involves a great amount of automatic abstractive insight extraction, text summarization and textual representation learning background. Such a natural language processing (NLP) model could save clinicians' time, and potentially provide them with more insights about the impression given that the model is trained with inputs of several experienced doctors. For this reason, in this paper, we propose a high functioning IMPRESSION prediction system based on encoder-decoder based sequence to sequence (seq2seq) models, which also processes the clinical information about the patient. Furthermore, we also make use of BERT representations[5], specifically biobert [6], which gives an enhanced machine understanding of the clinical text. We trained our model with the 957,134 de-identified radiology reports obtained from the University of Chicago (UC) Medicine, curated over the last 12 years until January 1st, 2020. We obtained the state of the art IMPRESSION prediction results, compared to well-established and competitive baselines such as LexRank [7] and pointer generator [8]. Experiments were employed both on the UC Medicine dataset and the publicly available Indiana University (IU) Chest X-ray dataset [9].

TABLE I
AN EXAMPLE OF FINDINGS, BACKGROUND AND IMPRESSION TRIPLET

Background	*Reason: R/O COVID-19, History: Influenza like illness*
Findings	*Multiple bilateral focal groundglass and airspace opacities, more extensive in the lower lungs, highly compatible with atypical pneumonia due to COVID-19. No significant lymphadenopathy. Enlarged main pulmonary artery measuring 34 mm, suggestive of pulmonary hypertension. Approximately normal heart size with no pericardial effusion. Degenerative spurring in the spine. Absence of enteric contrast material limits sensitivity for abdominal pathology. Limited evaluation with no gross abnormalities.*
Impression	*Multifocal groundglass and airspace opacities throughout the lungs, several of which are subpleural. The findings are highly compatible with COVID-19 pneumonia.*

Furthermore, we propose a methodology to fine-tune the IMPRESSION prediction model, to fit the model based on: 1-Modality (CT, X-Ray, Sonogram, MRI and Mammogram) 2-Specific case types (such as Covid-19 exams), and 3- Language style/needs (i.e. private or academic institution language styles, senior or junior experience levels). Such customization is desired because of the language-wise differences across modality, case type and the physician's expertise and style. Thus, as a part of supplementary training, we used an additional 1000 de-identified reports, obtained from a Chicago teleradiology private practice (denoted Telerad) to fine-tune the model. This process provided customized predictions aligning with the practice style. We then evaluated the performance enhancement drawing on this transfer learning approach by testing on a recently composed Covid-19 dataset. Experiments on both the Telerad test set and the Covid-19 dataset showed promising results, even though we tuned the model with limited external healthcare data.

A. Related Work

Healthcare research monitored the success of deep learning applications on mainstream NLP tasks including name-entity recognition, semantic role labeling, and part-of-speech tagging[10]. More recently, contextual embedding structures have gained popularity, which deliver an impactful and customized solution for NLP tasks. Within the healthcare domain, [11] explored generating contextual word embeddings from PubMed articles to better classify tweets during outbreaks. In this work, scholars demonstrated outperforming evidence of domain specific word representations over pre-trained embedding models such as Word2Vec and GloVe. This is consistent with our approach of utilizing radiology specific word embeddings over pre-trained models. In another relevant study, Ji et al. proposed using LSTM-based attentive relation networks in order to embed textual risk indicators based on mental disorders[12]. The essence of the desired output of this work lies in the realm of embedding unstructured text into vectoral representations. With these representations, the concise target text is generated via a generative or probabilistic search algorithm. In our scenario, the input is the FINDINGS and the clinical history (BACKGROUND) sections of the radiologist reports. The target output is the IMPRESSION paragraph/sentence. The frame of such a problem can be referred to as a neural abstractive summarization problem or text sequence generation problem in machine learning literature. Particular to our case, we additionally value the factual correctness of the predictions. Factual correctness plays an important role within our problem definition along with the utilization of natural language accuracy metrics such as ROUGE scores for model improvement efforts.

Text Generation and Summarization: Within the concept of information summarization from unstructured text, several learning techniques have been explored over the past decades (i.e. [13]). Some of the popular techniques are as follows. Surface level approaches leverage from the title and cue-terms for deriving the relevant pieces from sentences (i.e. superlative and comparative adjectives) [14]. Corpus based approaches utilize structural and sequential organization of terms using internal or external corpus (i.e. WordNet [15]). Cohesion-based approaches use cohesive associations and lexical chains among antonyms, repetitions, and synonyms [16]. Graph-based approaches apply each sentence as a node in a graph where the edges form the inter-connections among sentences. Graph-based models, more specifically LexRank [7] and TextRank [17] are some of the most popular extractive high

level text summarization techniques. Over the past decade, NLP research in healthcare has heavily explored several angles to derive valuable information from the unstructured patient narratives. Among those, the most recent studies have applied sentic computing techniques to better detect, interpret, and mine opinions in medical documents [18]. These efforts has been shown to lead to better measurement of patient reported outcomes and healthcare quality[19].

Extractive models generate summaries by cropping important parts from the original text and combining them together to create a coherent summary. Abstractive models, on the other hand, generate de novo summaries without being limited to reuse phrases from the original text. Machine learning models enable abstractive summarization where new words and phrases are generated to derive summaries. For example, in our case, the FINDINGS section contains mainly observations about an imaging exam and the IMPRESSION section can contain follow-ups or recommendations based on those observations which may or may not appear in the FINDINGS section. Although the main focus in this area geared towards extractive methods until recently, advancements in seq2seq and Natural Language Generation techniques are now making it possible to generate reasonable summaries by using abstraction [20], [21], [22]. Those who are among the first, Rush et al. applied an attention-based neural encoder and decoder for this task [23]. Nallapati et al. implemented a recurrent neural network for both the encoder and decoder pairs [24]. Pointer generator models were also proposed to emphasize the limitation that neural network models with a fixed vocabulary cannot tackle with the out of vocabulary words [24], [8]. Kryściński et al. applied a different type of mechanism which is a reinforcement learning architecture to generate summaries [25] and shortly thereafter Chen and Bansal posit improved results with a model that first selects sentences and then rewrites them [26].

Radiology Text Processing: More recently, in the context of clinical text and radiology domain, several other research areas have been explored including automated generation of coherent radiology reports from medical images. Among these studies, the major focus was on the text formation portions of the radiology workflow such as using long-short-term-memory (LSTM) network models to generate the textual paragraphs [27], [28]. Second stream literature focused on identifying actionable findings in radiology reports to help annotation in clinical workflows [29], [1].

Text Prediction and Summarization of Radiology Reports: Our work is closely relevant to recent work that studies impression prediction and evaluation of radiology reports. Hassanpour et al. used rule-based learning to derive name-entities from radiology reports and examined extracting named entities using traditional feature-based classifiers [30]. Subsequently, Goff and Loehfelm developed an open-source natural processing pipeline to identify important disease entities within the impression section of the radiology reports [31]. And finally, Zhang and his colleagues first achieved the automatic prediction of IMPRESSIONS by using FINDINGS and BACKGROUND information from the radiology reports [32]. They developed a neural seq2seq learning model which demonstrated high performance in terms of ROUGE metrics and clinical validation. Zhang et al.'s work is considered as an important milestone such that further research, shortly after, focused on factual correctness of these predicted impressions [33]. Research scholars also integrated Radlex ontology into seq2seq models [34] to enhance the clinical validity of automated IMPRESSION prediction systems within the radiology workflows. Our work attempts to improve the previous studies in two ways: First, the scale of our data and comprehensiveness of our test set enabled us coming up with a more generalized model. Second, we propose a higher accuracy real-time system drawing on a BERT-enhanced encoder-decoder based seq2seq model, which processes not only the FINDINGS section but also clinical BACKGROUND information of the patient as a secondary input. Last but not least, we provide a customization approach to enable the model fine-tuning based on the modality of examination, case types, and user style/needs. Overall, considering these two distinctions, our work positions itself as the pioneer in this direction to the best of our knowledge.

II. METHOD

A. Problem Definition

We would like to predict a sensible and correct estimate of radiologist IMPRESSIONS, using unstructured text of FINDINGS and the patients' BACKGROUND information as input. As discussed in the introduction, this IMPRESSION could be considered an abstractive summary of the fusion of the FINDINGS, which is possibly collected by several physicians. The proposed system takes as input the text of FINDINGS $w_1^{(f)}, w_2^{(f)}, \cdots, w_F^{(f)}$ and the text of BACKGROUND $w_1^{(b)}, w_2^{(b)}, \cdots, w_B^{(b)}$ and produces the IMPRESSION $w_1^{(i)}, w_2^{(i)}, \cdots, w_I^{(i)}$. Here, the words w can be any word, number, abbreviation, acronym, name or a mistyped sequence of characters. The lengths of the text sequences F, B and I are arbitrary and vary for each report along the corpus.

Since this is a sequence prediction task, the typical approach is using an enCODer-DECoder (CODEC) based model. In CODEC-based methodologies, input sequence is modeled as sequences of vectoral embeddings and the output is obtained to give the most likely sequence, generated using this embedding. Therefore, the problem definition can be segmented in two inter-related groups : (1) representing the input text as embeddings, and (2) modeling the CODEC architecture. As it will be explained in the next section, we utilize BERT for the former, and LSTMs for the latter. The input sentence is tokenized into a sequence of words/subwords then, each token is represented by real-valued vector embeddings, henceforth denoted as $\mathbf{x} \in \mathcal{R}^D$, where D is the dimensionality of the representation. These representations are typically rule-based or pre-trained embeddings that have a constant and bounded *vocabulary* ($\mathcal{V}$) of tokens (words/subwords). The impression prediction is then the sequence of tokens, within the vocabulary ($\mathcal{V}$), that maximizes the joint conditional probability given

the input. The performance of the model is evaluated by the averaged similarity of the predicted IMPRESSION sequence and the actual human IMPRESSION that correspond to the input FINDINGS. In creation of the truth annotation of our dataset, i.e. the FINDINGS, IMPRESSION and BACKGROUND information triplets for the training and test sets, UC Medicine Radiologists provided the required pre-processing, in order to ensure the validity of human generated labels after several rounds of quality check. This study was approved by the Institutional Review Board (IRB) with number IRB20-1853 University of Chicago Medicine.

B. System Description

The crux of our approach is in harvesting the power of BERT for the seq2seq abstractive text summarization. Specifically, we utilize BERT-enhanced embeddings for representing the input text, and adopt the high performance background-augmented pointer-generator network [32] for the seq2seq CODEC network. The overview of the proposed methodology is demonstrated in Fig. 1. In the general scheme, the FINDINGS and BACKGROUND text sequences are processed with a BERT-enhanced tokenization and the corresponding vectoral representations are obtained. These representations are then fed to a seq2seq IMPRESSION prediction network to obtain the most likely IMPRESSION sequence.

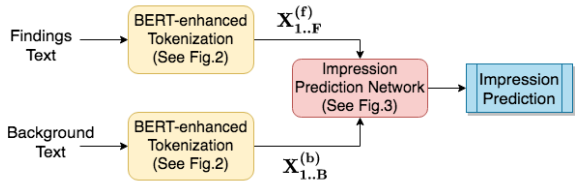

Fig. 1. General Flowchart of the impression prediction system.

C. BERT-enhanced Tokenization

In this paper, we approach to improve the the representation power and word coverage of the token embeddings inputted to the IMPRESSION prediction model. Hence in this work, we extend the state of the art background-augmented pointer-generator network [32] to harvest the language and word representation power of BERT, which is, pre-trained on billions of tokens of in-domain text. Specifically, both the FINDINGS and the BACKGROUND vectors inputted to the model are enhanced by biobert embeddings [6]. The flowchart of the BERT-enhanced tokenization and embedding extraction section is depicted in Figure 2. The input texts are tokenized with Stanford CoreNLP word tokenizer [35] into sequences of *word tokens*. The vectoral representation for each token is obtained by a simple table look-up from the pre-trained embeddings, if the token exists within the vocabulary. Otherwise, such a word is represented with a random vector. For this first set of representations, we used the pre-trained Glove[36] embeddings, open sourced by the authors of [32].

The tokens obtained by the CoreNLP tokenization is further processed to investigate the root words that lie within, in an

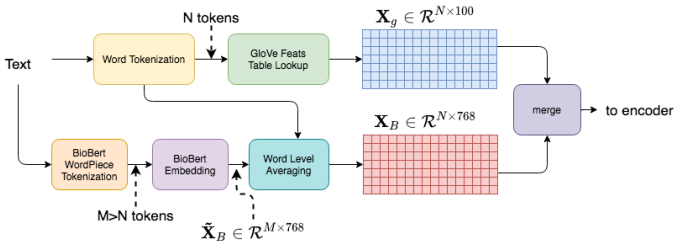

Fig. 2. BERT-enhanced tokenization prior to impression prediction.

effort to leverage for the missed spaces that occur during merging of reports. This procedure is done by maximizing the probabilities of such root words, if any, and inferring the position of the spaces that would split them. The idea behind this split comes from the naive assumption that words in a dictionary are independently distributed and follow the Zipf distribution. Once the probability distributions are estimated, the most probable split is obtained by maximizing the product of the probability of individual words. We used the implementation provided here[1].

For the input sentence $\mathcal{W}$, the word tokenization and the further splitting will yield N tokens:

$$\texttt{word-tokenize}(\mathcal{W}) = [w_1, w_2, \cdots, w_N] \quad (1)$$

The sequence length N, i.e. the sentence length, varies for different inputs. The D dimensional embeddings ($\mathbf{X}_g \in \mathcal{R}^{D\times N}$) are obtained by a simple table look-up procedure.

$$\texttt{embed}([w_1, w_2, \cdots, w_N]) = [\mathbf{x}_{g1}, \mathbf{x}_{g2}, \cdots, \mathbf{x}_{gN}] = \mathbf{X}_g \quad (2)$$

The subscript g denotes the GloVe embeddings with $D = 100$. The major addition we contribute is the incorporation of the BERT's context dependent power into the schema. For this, we obtained biobert embeddings to be inputted to the encoder. Unlike the other pre-trained vocabularies like GloVe or ConceptNet Numberbatch, BERT acts on WordPiece tokens. WordPiece tokens (denoted wp) have shared *word parts* across several different words in the language. This advantageous feature limits the vocabulary size of the network, while covering a great deal of the unseen words when compared to other vocabulary and tokenization techniques. For example, the sentence "`he's playing`" is tokenized as "`[he, ', s, play, ##ing]`. Also, with the WordPiece tokenization, any additional words can be composed of by already seen subwords (for example "`Immunoglobulin = I, ##mm, ##uno, ##g, ##lo, ##bul, ##in`).

$$\texttt{bert-tokenize}(\mathcal{W}) = [wp_1, wp_2, \cdots, wp_M] \quad (3)$$

Naturally, the sequence length of the WordPiece tokens for the input text will be larger than the word tokenization, i.e. $M \geq N$. Therefore we employ the following procedure. We first obtain the 768 dimensional biobert embeddings for each WordPiece token in the input text. Then, we obtain word level representations to match the GloVe vectoral representation in

[1]https://pypi.org/project/wordninja/

terms of sequence length, by averaging the WordPiece biobert representations that correspond to each word in the CoreNLP sequence.

$$\begin{aligned}\texttt{biobert}([wp_1, wp_2, \cdots, wp_M]) &= [\tilde{\mathbf{x}}_{B1}, \tilde{\mathbf{x}}_{B2}, \cdots, \tilde{\mathbf{x}}_{BM}] \\ &= \tilde{\mathbf{X}}_B \in \mathcal{R}^{768\times M}\end{aligned} \tag{4}$$

The biobert embedding vector that corresponds to the word w, $\mathbf{x}_{B,w}$ is obtained from the WordPiece representations as thus:

$$\mathbf{x}_{B,w} = \frac{1}{|wp \in w|}\sum_{wp\in w}\tilde{\mathbf{x}}_{B,wp} \tag{5}$$

The embeddings for FINDINGS and BACKGROUND, i.e. $\mathbf{X}^{(f)}$ and $\mathbf{X}^{(b)}$ are obtained by concatenating the GloVe and BERT embeddings.

D. Impression Prediction Network

For the impression prediction network we extended the background-augmented pointer generator network architecture. The general overview of this network is given in Figure 3. We will be providing a high level description of the model here, demonstrating the input/output relations for completeness, though readers interested in detailed functional descriptions are referred to [32]. The FINDINGS and BACKGROUND embeddings ($\mathbf{X}^{(f)}$ and $\mathbf{X}^{(b)}$) are encoded into a hidden state representation via their corresponding ENCODERs. Since the FINDINGS are lengthy dictations by the physicians and the BACKGROUND is merely the patient specific information like age, sex, clinical history etc., the lengths and information content of the two texts vary drastically. Therefore, the FINDINGS encoder has a deeper architecture than the BACKGROUND encoder, both of which are bidirectional LSTMs.

$$\begin{aligned}\mathbf{h}^{(f)} &= \texttt{encoder-f}(\mathbf{X}^{(f)}) \\ \mathbf{h}^{(b)} &= \texttt{encoder-b}(\mathbf{h}^{(f)})\end{aligned} \tag{6}$$

The background embedding is calculated with the background attention mechanism that takes the last hidden layer activation of the FINDINGS ($\mathbf{h}_F^{(f)}$) and the background activations($\mathbf{h}^{(b)}$) and generates a weight distribution.

$$a' = \texttt{attention-b}(\mathbf{h}_F^{(f)}, \mathbf{h}^{(b)}) \tag{7}$$

The single vector BACKGROUND embedding is a weighted average of the BACKGROUND activations. a' is as long as the number of tokens in $\mathbf{h}^{(b)}$ and the dot operation denotes the weighting the tokens of $\mathbf{h}^{(b)}$ with a' and summing.

$$\mathbf{b} = \mathbf{h}^{(b)} \cdot a' \tag{8}$$

The decoder at time t takes the previous IMPRESSION prediction (y_{t-1}), the previous decoder state (s_{t-1}) and the BACKGROUND embedding ($\mathbf{b}$) to produce the current state s_t.

$$s_t = \texttt{decoder}(s_{t-1}, y_{t-1}, \mathbf{b}) \tag{9}$$

The decoder state at time t (s_t) is used to obtain the FINDINGS attention, which is used for two goals : (1) decide how much of the FINDINGS tokens to be directly reflected in the prediction of the next IMPRESSION token, and (2) calculate the attended FINDINGS embedding ($\mathbf{h}_t^*$):

$$a = \texttt{attention-f}(\mathbf{h}^{(f)}, s_t) \tag{10}$$

The attended FINDINGS embedding at time t ($\mathbf{h}_t^*$) is obtained as a weighted average of the FINDINGS activations ($\mathbf{h}^{(f)}$):

$$\mathbf{h}_t^* = \mathbf{h}^{(f)} \cdot a \tag{11}$$

The IMPRESSION predictions, that are conditioned on $\mathbf{X}^{(f)}$, $\mathbf{X}^{(b)}$ and the preceding model outputs are obtained by maximizing the likelihood of each token activation using $\mathbf{h}_t^*$, s_t and a as input.

$$p(y_t|\mathbf{X}^{(f)}; \mathbf{X}^{(b)}; \hat{w}_{t-1}^{(i)}) = \texttt{prediction}(\mathbf{h}_t^*, s_t, a) \tag{12}$$

The sequence generation is conducted until a special END OF SENTENCE token is predicted or a pre-defined maximum output length is reached.

$$\hat{w}_t^{(i)} = \arg\max_{y\in\mathcal{V}} p(y_t^{(i)}|\mathbf{X}^{(f)}; \mathbf{X}^{(b)}; \hat{w}_{t-1}^{(i)}) \tag{13}$$

Using the new input representations enhanced with BERT embeddings, we follow the model explained in [32], in terms of the CODEC architecture. However, we increased the model depth for the FINDINGS encoder from 2 layers to 4 layers, and the hidden state dimensionality from 200 to 1024, in order the fully harvest the increased dataset size by almost 10-times, when compared to the work explained there.

E. Out of Vocabulary Token Handling

The clinical text of interest contains a relatively large number out of vocabulary (OOV) terms, due to several aspects like typing errors, or merged words occurred while combining separate reports into one report, dictated by different radiologists. During the table look-up phase in obtaining the vectoral representations, the words that exist in the representation vocabulary are replaced with the pre-trained D dimensional vectors. All others, i.e OOV tokens, are replaced with randomly generated vectors. Handling OOV terms is important because, it is naturally desired to use pre-trained GloVe or BERT embeddings to represent input tokens, rather than random vectors. It should be noted that the input word representations play a major role in the performance of the task, and enhancing this aspect is one of the contributions of this study. We augment the 100 dimensional GloVe vectors with BERT embeddings.

The initial OOV rate we observed after Stanford CoreNLP tokenization, measured for the GloVe vocabulary was more than 1%, which yields a word coverage (ratio between the covered unique tokens to all unique tokes discovered) of about 40%. This means that, although more than 98% of the tokens are covered by GloVe, more than half of the unique tokens are represented by random vectors. The further segmentation by the root words, eliminated the word merging errors, possibly caused by the dictation software used by radiologists or combination of dictations of several radiologists. As a result,

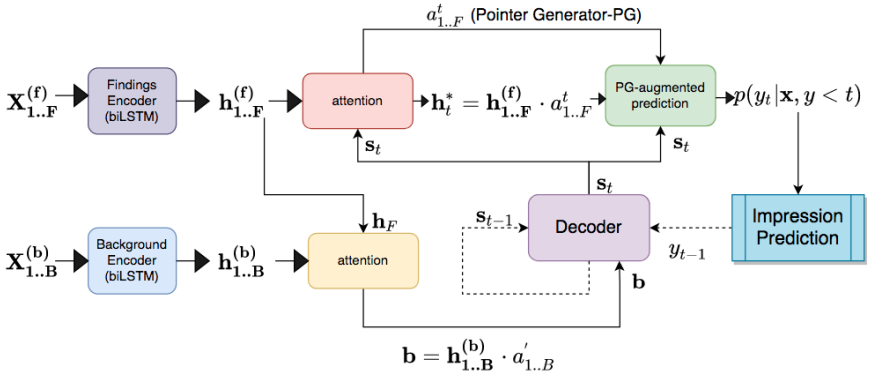

Fig. 3. Background-augmented PG network for impression prediction.

the coverage increased to 75% and the OOV rate reduced to 0.07%. The GloVe vocabulary and the corresponding vectors used in this work and in [32], were reported to have been trained on about $4M$ radiology reports. It is not difficult to see why an in-domain yet smaller training set would be preferable to an out-of-domain larger dataset for GloVe training. To this very point we advocate that utilizing BERT embeddings, which use WordPiece tokens and have been trained on millions of in-domain text, would practically reduce the OOV rate drastically and provide considerable help with the OOV tokens.

Therefore, we can accentuate the contribution of the BERT embeddings in two aspects: 1. The WordPiece tokenization of the BERT helps with the OOV words inherently. 2. Compared to the in-domain training size of the initial embedding representation, BERT is trained on a drastically larger corpus of medical text, hence probably possesses a better representation power than GloVe.

F. Over Generation Problem

One very common problem observed with the neural network-based machine translation and summarization systems is *over generation*, i.e. the repetitions of some words or phrases [37], [38], [39]. With *over generation*, the repetitions could occur several times and the length of repeating phrases could be as long as 8-9 words. We have observed that the background-augmented pointer generator model we use is also susceptible to such a phenomenon, due to the its decoding scheme. We observe that some of the impression prediction outputs exhibited several repetitions of their sub-sequences. The conundrum with this problem is that the recall, hence the ROUGE score, is not effected while the readability is seriously damaged. An impression prediction from the test set that suffers from over generation can be exemplified with the following text: *"1. Borderline* ***evidence of cholecystitis constipation evidence of cholecystitis constipation evidence of cholecystitis constipation****. 2. Negative chronicity not applicable, multiplicity not applicable most proximal not applicable rv strain not applicable"*

In order to address this issue, we propose the algorithm visualized in Figure 4 as a post processing step. For each token in the output sequence, the algorithm looks back T tokens,

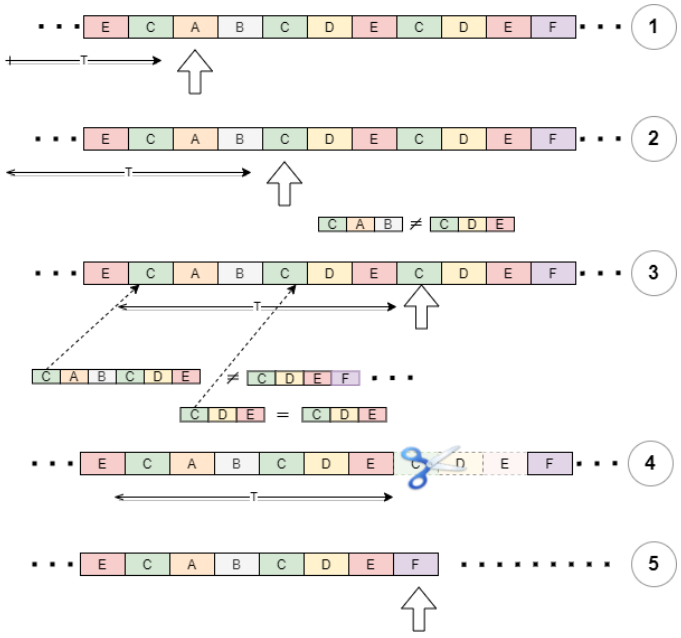

Fig. 4. Algorithm to remove repetitions on linear time: 1. The current cursor moves forward doing nothing, if it doesn't see the same token within the window of the past T tokens. 2-3: For all locations (m) of the occurrences of $w[t]$ in the past window of T tokens, checks if $w[m : t-1]$ matches $w[t : t + (t-m)]$, moves forward if they don't match. 4. Removes $w[t : t + (t-m)]$ for such m that exhibits a match. 5. Algorithm goes until it reaches the end of the shortened sequence

and removes any subsequence starting from the current frame, if the subsequence already exists in the previous T frames. This procedure makes sure to keep the recall unchanged while increasing the precision (and hence the readability) of the generated text.

III. Experimental Results

We conducted our experiments on the UC Medicine dataset and the publicly available IU Chest X-ray dataset. We evaluate the performance of the proposed model (1) by comparing the ROUGE scores of the system with the state of the art system outputs and (2) employing clinical validations with radiologists using both generalized and customized system outputs.

A. Dataset

The training dataset we used during model development was obtained from the UC Medicine over the past 12 years and has 957,134 de-identified radiology reports including their corresponding physician impressions. All of the reports used in our analysis are de-identified to meet HIPAA standards and do not contain any PHI information. The dataset spreads over a variety of imaging modalities such as CT, MRI, X-ray, sonogram and mammogram. The distribution of the imaging modalities of the UC Medicine dataset is given in Figure 5. The dataset is distributed over several body parts, yet over 95% is covered by chest, abdomen, head/neck, spine, vessel, joints and knee.

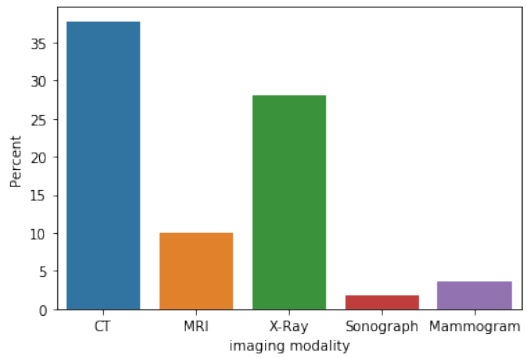

Fig. 5. The distribution of several imaging modalities in the training dataset

One of the most critical aspects in NLP-based machine learning applications is the scale and the content of the training material. This issue is the main reason why several recent NLP applications with fascinating performances could exhibit incorrect or biased features in their language usage. The reason for this, is not the way the models are trained, but the dataset that they are trained on. That being said, we find it important to mention the size and the extent (modality/body part), the reporting content (positive/negative cases) of the dataset we used. For practical impression prediction applications, it is desired for a dataset to have more than "no acute findings" and longer impression summaries. Out of the $950K$ impressions in the UC Medicine dataset, $520K$ of cases have multiple indexes/sentences of impressions and are partially positive. $310K$ of the cases have only one sentence per impression and are positive. Only $120K$ cases are negative; exhibiting no evidence of findings. The dataset has on average 119.4 tokens in the FINDINGS section and 27.3 tokens in the IMPRESSION section, whereas the corresponding average values for the IU Chest X-ray dataset is $45/10.5$. The distribution of the report lengths can be seen in Figure 6.

The scale of the training size used in this work should also be taken into consideration. Not only the content of the data, but also the size of the dataset used in this work is significantly larger than most of datasets used in similar works in the literature (impression prediction). Just like the increase in the dataset size in ImageNet challenge made the usage of deep networks like alexnet [40] feasible for image recognition, compared to similar and shallower networks like LeNet [41], the scale of data made the extension of input representation feasible and fruitful for this work.

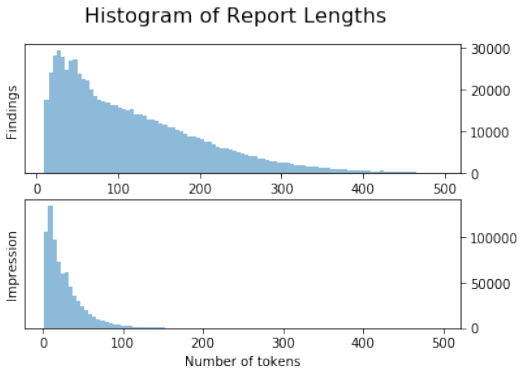

Fig. 6. The distribution of report lengths for FINDINGS and IMPRESSION

B. Experimental Results

For the evaluation of the proposed system, we used the widely used ROUGE metric, which stands for "Recall-Oriented Understudy for Gisting Evaluation". We report the ROUGE-1, ROUGE-2 and ROUGE-L performances, defined in [42]. For experimental analysis, we randomly segmented the dataset into training, development and test sets with ratios proportional to 70-15-15 percent of the dataset. The ROUGE-L results are monitored on the development set at the end of each epoch to decide the early stopping during the training. The trained model is then evaluated on the test set and compared with the following well established extractive summarization and the state-of-the art abstractive summarization baselines.

- LexRank[2] [7] : Extractive summarization, based on computing relative importance of textual units using stochastic graphs
- Background-augmented pointer-generator (PG)[3] [32]: The abstractive baseline that uses the same CODEC structure as this work with solely GloVe embeddings.

For the sake of completeness of system descriptions, the proposed system can be briefly defined as BERT-enhanced pointer-generator (PG $\oplus$ BERT).

The ROUGE performances of the proposed system compared with the above-mentioned baselines, calculated on UC Medicine test set is provided in Table II. The performance difference observed between PG and PG $\oplus$ BERT demonstrates the improvement brought by the BERT's representation power. On another note, in order to be able to assess the effect of the training with the large UC Medicine dataset, we evaluated the UC Medicine test set with the model pre-trained on Stanford Dataset, which was made publicly available by the authors of PG[3]. We name this model as PG@S.

To observe the generalization power of the proposed model to data from another institution and also contribute to the benchmark created with the IU Chest X-ray dataset, we evaluated our models on this publicly available data. The IU Chest dataset results are given in Table III.

[2]For LexRank we used the Sumy library

[3]For PG, we used the implementation provided in author's repository

TABLE II
SYSTEM PERFORMANCE ON UC MEDICINE TEST SET

System	ROUGE-1	ROUGE-2	ROUGE-L
LexRank	15.11	8.61	13.21
PG@S	33.16	18.48	32.23
PG	41.07	28.84	40.62
PG $\oplus$ BERT (this work)	47.17	32.89	45.91

TABLE III
SYSTEM PERFORMANCE ON THE UI CHEST SET

System	ROUGE-1	ROUGE-2	ROUGE-L
LexRank	15.42	5.65	14.60
PG@S [32]	35.02	20.79	34.56
PG	41.58	28.64	41.21
PG $\oplus$ BERT (this work)	43.95	31.45	43.56

C. Clinical Validation

In addition to the statistical validation obtained by the ROUGE analyses, we employed radiologists evaluations, assessing the clinical validity of the predicted impressions. Since the ROUGE metric is only measuring word and sequence level similarity of the predicted and actual impressions, it comprises the visible deficiency of missing the factual correctness and the utility of the inference. Two inferences conveying the same information may result in a low ROUGE score if they have word mismatches. For this reason, we also employed human assessments with one UC Medicine radiologist and one independent external board certified radiologist to understand the clinical validity of the model predictions. We randomly sampled 1000 reports from the UC Medicine test set and asked the experts to rank each report by comparing the original impression with the predicted impression. If the original impression is better than the system output, they voted "negative". If the system output appears to be better than the original human generated impression, they voted "positive". We monitored that this incident occurred when there are some incidental findings in the findings section that were not originally reflected to the impression but the system successfully captures them. And finally the outputs are voted as "neutral" when both the original human impression and the system output roughly equal to each other.

As a result, we found that 669 of the reports were marked as "neutral", 83 of the reports were marked as "positive" and 235 of them were marked "negative". For 13 reports, the inter-coder reliability could not be established so that we excluded them from the final evaluation. As a result, we achieved 76 percent of the samples were comprising the neutral and positive votes meaning that 76% of the predictions were stated by radiologists to be at least as accurate as human generated impressions.

D. Customization per Modality, Institution and Experience

Narratives of distinct imaging modalities posit structural differences. For instance, Ultrasound impressions are more structured, which include specific "BIRADS" and "RECOMMENDATION" sections whereas CT impressions are more free form unstructured texts and more lengthy on average. Examples for Ultrasound and CT impression narratives are given in Table IV.

TABLE IV
AN EXAMPLE OF FINDINGS, BACKGROUND AND IMPRESSION TRIPLET

ULTRASOUND	*Continued follow-up with repeat bilateral breast ultrasound in one year is recommended. BIRADS: 2 - Benign finding. RECOMMENDATION: T - Take Appropriate Action - No Letter.*
CT	*1. Stable pulmonary nodule.2. Stable intrathoracic lymphadenopathy. Stable small upper abdominal node which was PET+ in the past.3. Stable osseous metastases.4. Mass in the neck is incompletely evaluated.*

Hence, modality specific fine-tuning is applied to generate more accurate results. While we realize such structural differences among modalities, we also observed that the impression length exhibits differences based on the institution type. For example, a teleradiology private practice may prefer a more succinct and brief wording, whereas as an academic institution provides more detailed impressions about an exam. Such difference even occurs among the radiologists from the same institution based on their style and tenure in the field. For a given finding, we monitored two separate impressions generated by two different radiologists. One of them preferred to include a minor incidental finding into the impression, whereas the other one dictated only major/actionable findings within the impression. Thus, we have decided to apply a sample customization for an additional institution. We obtained 1000 CT reports from a Chicago based teleradiology private practice (Chicago Telerad, PLLC) and fine-tuned our existing model using 80% of the reports and tested with the remaining 20%. Based on ROUGE scores, we observed 10.3 percent higher precision on the test set since our model better captured the private institution's style and length choices. In the context of clinical validity, practitioners also rated the fine-tuned model higher on randomly selected 100 reports. After fine-tuning, they voted 79 system outputs to be at least as good as the original impression, whereas only 74 were as good as the original before fine-tuning.

Furthermore, since we fine-tuned the model with a specific CT modality type, we also observed higher scores on a small sample of Covid-19 CT data. As a result, we propose the customization module as an important feature for our proposed solution.

IV. DISCUSSION AND LIMITATIONS OF WORK

While we showed the improved performance of our proposed method on predicting radiology impressions, we also recognize several limitations of our work. First, as all machine learning systems, our model suffers from out-of-domain test data. More explicitly, our proposed training strategy relies on the limits of samples and exposes questions if the model is transferable to unseen body parts. For instance, "pleural effusions" and "opacity in lobe" are common observations in chest examinations, yet do not exist in musculosketal exams. A typical healthcare institution may not have sufficient

examinations that focus on rare body parts. Thus, we monitor to what extent our model can perform on unseen body parts during training. As a result, we sought out answers to the above-mentioned concern through the following steps.

We collaborated with additional academic radiologists from Columbus, Ohio who are experts in reading musculosketal radiology exams, to test the performance of our system on rare and more specific cases. Musculoskeletal radiologists (MSK) use X-ray, MRI, CT, and ultrasound to diagnose and assess issues with the musculoskeletal system in the most detail possible. We obtained additional 100 reports to monitor the performance on these new rare samples. Since our model has been originally trained with the mixed 950K UC Medicine dataset, the test performance on this musculoskeletal data was naturally lower than the in-domain test performance. In order to 'customize' the impression prediction, as our paper title suggests, we used 89 percent of the additional data to fine tune the model and performed tests on the remaining 11 percent. We observed approximately 4 absolute point increases in statistical metrics (ROUGE) after fine-tuning the model with the in-domain data, resulting in a 38.4 ROUGE score on average. The fine-tuning process only served a small increase in performance due to the limitation in data size for such rare cases. However, we estimate that an additional 3000 or 4000 samples for rare body parts could help improve the system performance up to the accuracy level of the original model.

The second limitation is that the follow-up recommendations from the physicians are missing from the predicted impressions. Since our dataset is fully de-identified and does not contain any PHI information, the model suffered in providing follow-up recommendations, which are not included in findings section. Future research may overcome this bottleneck by introducing institution-specific recommendation styles and the use of in-house reports where doctor names are included in the training set, as long as HIPAA standards are met.

Finally, due to the computational resource and time constraints, the training of the model containing the 950K dataset takes about 1 month with a GPU of 16GB of memory. In order for the length of input text to be adequate, we had to reduce the batch size drastically, which in turn increased the training time. For the scholars interested in replicating our model on their own data, we suggest they consider paralleling this model into multiple GPU's to utilize larger inputs with faster training.

V. Conclusion

In the context of intelligent recommender systems in public health decisions, accurate and automated information generation from clinical reports could save a clinician's time, improve communication, and reduce errors. In this work, we presented a fully automated impression prediction system which outperformed previously proposed models in terms of both statistical validation (ROUGE metric) and clinical validation (physician evaluations). From research implications view: Our work attempted to enhance the previous models in multiple ways. First, the comprehensiveness of our development and test set enabled us to develop a more generalized model. We utilized almost all applicable radiology modality samples, used data from a large academic institution and tested both on another academic institution and private practice data. Second, drawing on a BERT enhanced encoder-decoder based seq2seq neural network model, we were able to eliminate the rule-based, static word-embedding usage problems. Specifically, we leveraged from contextual clinical embeddings to better process poorly written reports. Third, we utilized a neural network architecture where we could process not only the findings section, but also the clinical background information about the patient as a second input. Thus, scholars can make use of our architecture to use other types of inputs such as lengthy reason of study, protocol/technique, and even image extracted annotations to improve the prediction scope for future research. Finally, we provided a customization module to enable model fine-tuning based on: 1- Modality including CT, X-Ray, Sonogram, MRI and Mammogram. 2- Specific case types such as Covid-19 exams. And 3- Radiologist language style/needs (i.e. private or academic institution language styles, senior or junior experience levels). From a practitioner standpoint, our model can suggest: 1- Real-time tailored impression for dictations that reduce the time spent on each report, allowing them to spend more time on patient care and research. 2- Mitigate the risk of overlooking crucial findings due to physician mental overload and burnout. 3- Practitioners can utilize the model to audit previous impressions by using batch processing capabilities, fix errors, and even use as a support system to train inexperienced radiologists for the impression generation process. Overall, considering the above-mentioned contributions, our work positions itself as a facilitator towards AI-driven radiology functionalities. We hope that our study attracts the attention of AI scholars to the clinical text processing/generation problems, automation of clinical documentation, which eventually may reduce human-based errors in healthcare workflows and inspires future work in this field.

Acknowledgment

Authors would like to thank Inference Analytics personnel for their continuous support of this research during the entire process, Dr. Vineet Khanna, Dr. Sami Faisal and Chicago Telerad for their cooperation, Yuhao Zhang for making the code for PG publicly available and his fruitful discussions, Alican Gok for his contribution with the repetition removal algorithm, and finally our dear friends and colleagues Dingchao Zhang and Yuntao Li for their valuable contributions with the initial model.

References

[1] John Zech, Margaret Pain, Joseph Titano, Marcus Badgeley, Javin Schefflein, Andres Su, Anthony Costa, Joshua Bederson, Joseph Lehar, and Eric Karl Oermann, "Natural language–based machine learning models for the annotation of clinical radiology reports," *Radiology*, vol. 287, no. 2, pp. 570–580, 2018.

[2] Z Berkay Celik, David Lopez-Paz, and Patrick McDaniel, "Patient-driven privacy control through generalized distillation," in *2017 IEEE Symposium on Privacy-Aware Computing (PAC)*. IEEE, 2017, pp. 1–12.

[3] Ashwin Belle, Raghuram Thiagarajan, SM Soroushmehr, Fatemeh Navidi, Daniel A Beard, and Kayvan Najarian, "Big data analytics in healthcare," *BioMed research international*, vol. 2015, 2015.

[4] M Lafortune, G Breton, and JL Baudouin, "The radiological report: what is useful for the referring physician?," *Canadian Association of Radiologists journal= Journal l'Association canadienne des radiologistes*, vol. 39, no. 2, pp. 140, 1988.

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

[6] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[7] Günes Erkan and Dragomir R Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.

[8] Abigail See, Peter J Liu, and Christopher D Manning, "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.

[9] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.

[10] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[11] Aparup Khatua, Apalak Khatua, and Erik Cambria, "A tale of two epidemics: Contextual word2vec for classifying twitter streams during outbreaks," *Information Processing & Management*, vol. 56, no. 1, pp. 247–257, 2019.

[12] Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Transactions on Computational Social Systems*, 2020.

[13] Josef Steinberger and Karel Ježek, "Evaluation measures for text summarization," *Computing and Informatics*, vol. 28, no. 2, pp. 251–275, 2012.

[14] Vivi Nastase, "Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 763–772.

[15] George A Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[16] Regina Barzilay and Michael Elhadad, "Using lexical chains for text summarization," *Advances in automatic text summarization*, pp. 111–121, 1999.

[17] Rada Mihalcea and Paul Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[18] Md Shad Akhtar, Asif Ekbal, and Erik Cambria, "How intense are you? predicting intensities of emotions and sentiments using stacked ensemble [application notes]," *IEEE Computational Intelligence Magazine*, vol. 15, no. 1, pp. 64–75, 2020.

[19] Erik Cambria, Tim Benson, Chris Eckl, and Amir Hussain, "Sentic proms: Application of sentic computing to the development of a novel unified framework for measuring health-care quality," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10533–10543, 2012.

[20] Tal Baumel, Matan Eyal, and Michael Elhadad, "Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models," *arXiv preprint arXiv:1801.07704*, 2018.

[21] Sumit Chopra, Michael Auli, and Alexander M Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 93–98.

[22] Peter Liu and Xin Pan, *Improving abstraction in text summarization*, 2016 (accessed October 21, 2020), http://goo.gl/16RNEu.

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

[24] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al., "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv preprint arXiv:1602.06023*, 2016.

[25] Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher, "Improving abstraction in text summarization," *arXiv preprint arXiv:1808.07913*, 2018.

[26] Yen-Chun Chen and Mohit Bansal, "Fast abstractive summarization with reinforce-selected sentence rewriting," *arXiv preprint arXiv:1805.11080*, 2018.

[27] Mehdi Moradi, Ali Madani, Yaniv Gur, Yufan Guo, and Tanveer Syeda-Mahmood, "Bimodal network architectures for automatic generation of image annotation from text," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 449–456.

[28] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Advances in neural information processing systems*, 2018, pp. 1530–1540.

[29] Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors, "Natural language processing in radiology: a systematic review," *Radiology*, vol. 279, no. 2, pp. 329–343, 2016.

[30] Saeed Hassanpour and Curtis P Langlotz, "Information extraction from multi-institutional radiology reports," *Artificial intelligence in medicine*, vol. 66, pp. 29–39, 2016.

[31] Daniel J Goff and Thomas W Loehfelm, "Automated radiology report summarization using an open-source natural language processing pipeline," *Journal of digital imaging*, vol. 31, no. 2, pp. 185–192, 2018.

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

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

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

[35] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[37] Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah, "Coverage embedding models for neural machine translation," *arXiv preprint arXiv:1605.03148*, 2016.

[38] Rohit Gupta, Patrik Lambert, Raj Nath Patel, and John Tinsley, "Improving robustness in real-world neural machine translation engines," *arXiv preprint arXiv:1907.01279*, 2019.

[39] Long Zhou, Jiajun Zhang, and Chengqing Zong, "Look-ahead attention for generation in neural machine translation," in *National CCF Conference on Natural Language Processing and Chinese Computing*. Springer, 2017, pp. 211–223.

[40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[41] Yann LeCun, Yoshua Bengio, et al., "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, pp. 1995, 1995.

[42] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain, July 2004, pp. 74–81, Association for Computational Linguistics.