Optum at MEDIQA 2021: Abstractive Summarization of Radiology Reports using simple BART Finetuning

Ravi Kondadadi and **Sahil Manchanda** and **Jason Ngo** and **Ronan McCormack**

Optum

Abstract

This paper describes experiments undertaken and their results as part of the BioNLP MEDIQA 2021 challenge. We participated in Task 3: Radiology Report Summarization. Multiple runs were submitted for evaluation, from solutions leveraging transfer learning from pre-trained transformer models, which were then fine tuned on a subset of MIMIC-CXR, for abstractive report summarization. The task was evaluated using ROUGE and our best performing system obtained a ROUGE-2 score of 0.392.

1 Introduction

A BioNLP 2021 shared task, the MEDIQA challenge aims to attract research efforts in NLU across three summarization tasks in the medical domain: multi-answer summarization, and radiology report summarization. We participated in the radiology report summarization and offer experiments and results. A radiology report describes an exam and patient information resulting from trained clinicians(radiologists) interpreting imaging studies during routine clinical care (Zhang et al., 2018). The primary purpose of the report is for radiologists to communicate imaging results to ordering physicians (Gershanik et al., 2011). A standard report will consist of a Background section which will contain details of the patient and describe the examination undertaken, A findings section, in which the radiologist has dictated the initial results into the report, and an Impression section. The Impression section consists of a concise summarization of the most relevant details from the exam based on the dictated findings. Although guidelines for the practice of generating radiology reports are outlined by the American College of Radiology (ACR), there is flexibility in the document in the usage of terms for describing findings and where they are documented. This can lead to referring physicians focusing on just the impressions section of the document (Hall, 2012). Additionally, the process of writing the impressions from the dictation of the findings is time-consuming and repetitive. In this work we propose experiments to automate the generation of the impressions section from the findings of the radiology report, accelerating the radiology workflow and improving the efficiency of clinical communications. Experiments were performed implementing sequence to sequence models with encoder-decoder architecture like BART (Lewis et al., 2019), Pegasus (Zhang et al., 2020a), and T5 (Raffel et al., 2020). These models were then further fine-tuned on a subset of MIMIC-CXR Dataset (Johnson et al., 2019), to generate abstractive summaries from the findings section of the report. MIMIC-CXR is de-identified and Protected health information (PHI) removed, large publicly available dataset of chest radiographs in DICOM format with free-text radiology reports. A subset of MIMIC-CXR and Indiana datasets [1] used for validation carried out using standard ROUGE (Lin, 2004) metrics.

2 Related Work

Initial efforts on summarization were mainly focused on Extractive summarization. Extractive summarization is the process involving extraction of noteworthy words from the text to form a summary. (Luhn, 1958; Kupiec et al., 1995) The advent of Neural network models enabled Abstractive summarization, which involves producing new words to convey the meaning of the text. This involves rephrasing the text in a shorter and more succinct form using similar but not the exact words used in the main text.

Nallapati et al. (2016) proposed an RNN based approach to not only achieve state-of-the-art results in extractive summarization but also enable this model to be trained on abstractive summaries.

[1] https://openi.nlm.nih.gov/faq/collection

Proceedings of the BioNLP 2021 workshop, pages 280–284

Rush et al. (2015) described an attention-based summarization approach where an encoder and a generator model are jointly trained on article pairs. Their work builds on attention-based encoders that are used in neural machine translation (Bahdanau et al. (2016)). Fan et al. (2018) build on the previous work on abstractive summarization to create length constrained summaries and summaries concentrated on particular entities and subjects in the text. Paulus et al. (2017) used intra-temporal attention to produce state-of-the-art results on CNN/Daily Mail dataset.

The work on summarizing radiology reports started with the extraction of information from the text (Friedman et al., 1995; Hassanpour and Langlotz, 2016). For instance, Cornegruta et al. (2016) proposed using clinical language understanding of a radiology report to extract Named entities. A Bidirectional LSTM architecture was used to achieve this. Zhang et al. (2018) describes one of the first attempts at automatic summarization of radiology reports. This work describes an encoder-decoder architecture. Both the encoder and decoder sides are made of Bidirectional LSTMs using the attention framework (Bahdanau et al., 2016).

With the advent of transformers, Pretraining based language generation has been the norm in summarization. Zhang et al. (2019) and Liu (2019) used BERT (Devlin et al., 2019), a pre-trained transformer model on extractive summarization, and achieved state of the art results. Sotudeh et al. (2020) proposed an approach to content selection for abstractive text summarization in clinical notes. Zhang et al. (2020b) presented a general framework and a training strategy to improve the factual correctness of neural abstractive summarization models for radiology reports. In this work, we fine-tune a pre-trained BART architecture (Lewis et al., 2019) for the radiology report summarization task.

3 Task Description & Dataset

The objective of this task is to generate summary of a given radiology report. The training data for the MEDIQA 2021 Radiology report summarization shared task is extracted from a subset from the MIMIC-CXR Dataset (Johnson et al., 2019). The training set contains around 91,544 examples of radiology reports and the corresponding summaries. Each example contains three fields; Findings field contains the original human-written radiology findings text, impression contains the human-written radiology impression text and background contains background information of the study in text format. One can use both the findings and the background fields to generate the summary. There are two development sets that come from two different institutes. The first development set from MIMIC-CXR contains around 2000 examples. There is another development set that also contains 2000 examples from the Indiana University radiology report dataset (Johnson et al., 2019). In all our experiments, we first trained our model on the training set and tested on the validation set. For the actual task submissions, we trained our models by combining training set and both the development sets.

4 Method & Results

Our proposed method leverages pretrained summarization models. We finetuned three types of pretrained models for the radiology report summarization; BART (Lewis et al., 2019), T5 (Raffel et al., 2020) and Pegasus (Zhang et al., 2020a). We used Huggingface Transformers (Wolf et al., 2020) library for finetuning.

BART: Developed by Facebook, BART is a denoising autoencoder. Since it uses the standard transformer-based neural machine translation architecture, it is a generalization of both BERT and GPT3 (Brown et al., 2020). For pretraining, it was trained by shuffling the order of sentence (an extension of next sentence prediction) and text infilling (an extension of the language masking). During text infilling, random spans of text are replaced by masked tokens. The job of the model during training is to recreate this span. Due to its flexible transformer architecture, the inputs to the encoder do not need to be aligned with the outputs of the decoder. This enables the BART model to be trained on a variety of tasks such as token masking, token deletion, sentence permutation, document rotation, etc. Since BART has an autoregressive decoder, it is better suited for sequence generation tasks such as summarization.

T5: T5 stands for Text-To-Text Transfer Transformer. It is a sequence-to-sequence model that takes in text and outputs text. This text-to-text framework enables one to use the same model, loss

function, and hyperparameters on any NLP task, which can range from document summarization to classification. As a result, the way that data is fed into the model is quite different from models like BERT. The task description is used as a prefix to the input. For example, to translate a sentence from English to French, the input would be prefixed with "translate English to French:" Similarly, to summarize a passage, you would add the prefix "summarize:" followed by the text to be summarized. This text-to-text framework uses the same model across a range of tasks. T5 model made improvements on a wide range of categories such as model architecture, and pretraining objectives.
T5 uses the standard transformer architecture (Vaswani et al., 2017). For pretraining, T5 was trained on denoising, where spans of text are replaced with the drop token. The model objective is to reproduce the span of text given the drop token.

Pegasus: PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive SUmmarization Sequence-to- sequence) Pegasus starts with the concept that if the pretraining task and fine-tuning task are closely related, then the model will perform better. As a result, they designed a pretraining task specifically for abstractive summarization. This pretraining task, gap-sentence generation, removes entire sentences from documents. The model's learning objective is to recover these sentences in the concatenated model output. Instead of randomly removing sentences, only the important sentences are removed, so that the model can reproduce these sentences that summarize the text. As a result of this pretraining task, Pegasus can achieve results like T5 with 5% of the parameters.

Table 1 shows the model performance of each participant in the leaderboard for the top 10 teams. Only Rouge-2 F1 is shown because that was the metric used to rank the teams in this task. Our method ranked third on the leaderboard.

System	Rouge-2 F1
Baidu	0.436
IBM	0.408
Optum	0.392
QIAI	0.378
Low-rank-AI	0.331
CMU	0.327
ChicHealth	0.324
healthAI	0.308
DAMO-ALI	0.276
Fudan University	0.274

Table 1: Top 10 teams on the leaderboard

Run	Rouge 1	Rouge 2	Rouge L
1	60.51	48.14	57.65
2	52.35	40.98	50.41
3	35.72	22.69	31.53
4	63.47	51.35	60.54
5	56.14	44.65	53.98
6	37.8	24.73	33.80
7	58.59	46.5	56.01
8	62.85	51.22	60.25

Table 2: Evaluation of Radiology Report Summarization on the development set

4.1 Experiments

We propose eight different runs for this task. Table 2 shows the evaluation of different models we experimented with on the development set. We experimented with different versions of BART, T5 and Pegasus on Huggingface Transformers. We ended up using BART-base, T5-small, T5-base and Pegasus-Pubmed due to memory limitations of our GPUs. The following set of hyperparameters are applied for the following runs. Learning rate=5e-05 , number of epochs=15, gradient accumulation steps=5. The evaluations results of various runs for the radiology report summarization task are summarized in Table 2.

1. Our first proposed method is based on BART-base. We finetuned BART-base on the training set and tested on the development set. We used a batch size of 20 for both training and validation sets.

2. In this run, we used T5-small and finetuned on the training set. We used a batch size of 20 for both training and validation sets.

3. In our third run, we finetuned on pegasus-pubmed. We were able to use only a smaller batch size of 2.

4. The fourth run is similar to the first approach, but we also used the background section in

addition to findings. In this case, we were able to use a batch size of 10.

5. This run is same as the fourth one, but we used T5-small as our base model. A batch size of 10 was used.

6. In this run, but we used Pegasus-pubmed as our base model. A batch size of 1 was used.

7. This run is same as the first run, but we used T5-base as the base model. A batch size of 10 was used.

8. In this run also, we used T5-base as the base model except that we also used background section. A batch size of 2 was used.

Overall, the best results on the test set are achieved using the BART-base as the pre-trained model. The model is trained using just the findings section on the test set. But on the development set, using the background section in addition to the findings helped.

5 Conclusion

In this paper, we present all our experiments of fine-tuning pre-trained models for radiology report summarization. Our experiments demonstrate how an encoder-decoder architecture like BART, which achieved state-of-the-art results in text generation tasks outperforms other architectures in this particular task. Our methods proved effective on the summarization task and were ranked third on the leaderboard.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. 2016. Modelling radiological language with bidirectional long short-term memory networks. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 17–27, Auxtin, TX. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization.

C. Friedman, G. Hripcsak, W. DuMouchel, S. B. Johnson, and P. 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. *AMIA Annu. Symp. Proc.*, 2011:465.

Ferris M. Hall. 2012. Language of the Radiology Report. *Am. J. Roentgenol.*

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

Alistair E. W. Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. 2019. The MIMIC-CXR Database. Type: dataset.

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*, SIGIR '95, page 68–73, New York, NY, USA. Association for Computing Machinery.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu. 2019. Fine-tune bert for extractive summarization.

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

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2016. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization.

Sajad Sotudeh, Nazli Goharian, and Ross W. Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Haoyu Zhang, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

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

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D. Manning, and Curtis P. Langlotz. 2020b. Optimizing the factual correctness of a summary: A study of summarizing radiology reports.