BDKG at MEDIQA 2021: System Report for the Radiology Report Summarization Task

Songtai Dai, Quan Wang, Yajuan Lyu, Yong Zhu
Baidu Inc., Beijing, China
{daisongtai,wangquan05,lvyajuan,zhuyong}@baidu.com

Abstract

This paper presents our winning system at the Radiology Report Summarization track of the MEDIQA 2021 shared task. Radiology report summarization automatically summarizes radiology findings into free-text impressions. This year's task emphasizes the generalization and transfer ability of participating systems. Our system is built upon a pre-trained Transformer encoder-decoder architecture, i.e., PEGASUS, deployed with an additional domain adaptation module to particularly handle the transfer and generalization issue. Heuristics like ensemble and text normalization are also used. Our system is conceptually simple yet highly effective, achieving a ROUGE-2 score of 0.436 on test set and ranked the 1st place among all participating systems.

Background: Examination: chest (portable AP) indication: history: ___m with acute coronary syndrome technique: upright AP view of the chest comparison: chest radiograph ___
Findings: Patient is status post median sternotomy and CABG. Heart size remains mildly enlarged. The aorta is tortuous. Mild pulmonary edema is new in the interval. Small bilateral pleural effusions are present. Patchy bibasilar airspace opacities likely reflect areas of atelectasis ...

Impression: Mild pulmonary edema and trace bilateral pleural effusions.

Figure 1: A radiology report sampled from MEDIQA 2021 training set, where the impression is a summarization of the findings taking the background into account.

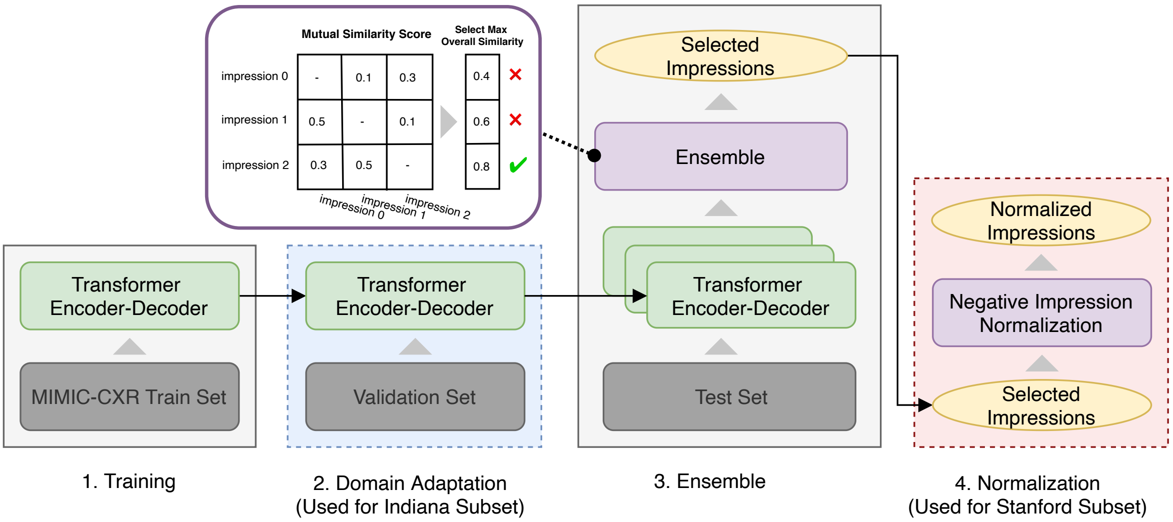
1 Introduction

Radiology reports are documents that record and interpret radiological examinations. A typical radiology report usually consists of three sections: (1) a *background* section that describes general information about the patient and exam, (2) a *findings* section that presents details of the examination, and (3) an *impression* section that summarizes the findings against the background (Kahn Jr et al., 2009). Figure 1 provides an example of such a radiology report. In a standard radiology reporting process, a radiologist first dictates detailed findings into the report, and then summarizes the findings into a concise impression based also on general background of the patient (Zhang et al., 2018). The impression section, which provides the most valuable information to make clinical decisions, is the most crucial part of a radiology report for both doctors and patients. However, manually summarizing radiology findings into impressions are time-consuming and error-prone (Gershanik et al., 2011), which necessitates the need to automatically generate radiology impressions.

The MEDIQA 2021 shared task (Abacha et al., 2021) at the NAACL-BioNLP workshop sets up a *Radiology Report Summarization* subtask, the aim of which is to build advanced systems to automatically summarize radiology findings (along with the background) into concise impressions. A key feature of this task is that radiology reports used for training and evaluation are collected from different sources, e.g., training instances are sampled from the MIMIC-CXR database (Johnson et al., 2019) and some evaluation instances come from the Indiana chest X-ray collection (Demner-Fushman et al., 2016). This inevitably results in significant discrepancies between training and evaluation, posing new challenges to the generalization and transfer ability of participating systems.

Zhang et al. (2018) presented the first sequence-to-sequence attempt at automatic summarization of radiology findings into natural language impressions. After that, several extensions and improvements have been proposed, e.g., to take into account the factual correctness (Zhang et al., 2019) or the ontologies (MacAvaney et al., 2019; Gharebagh

Proceedings of the BioNLP 2021 workshop, pages 103–111
June 11, 2021. ©2021 Association for Computational Linguistics


Mutual Similarity Score
Select Max Overall Similarity
Selected Impressions
Ensemble
Transformer Encoder-Decoder
MIMIC-CXR Train Set
Transformer Encoder-Decoder
Validation Set
Transformer Encoder-Decoder
Test Set
Normalized Impressions
Negative Impression Normalization
Selected Impressions
1. Training
2. Domain Adaptation (Used for Indiana Subset)
3. Ensemble
4. Normalization (Used for Stanford Subset)

	impression 0	impression 1	impression 2	Select Max Overall Similarity	
impression 0	-	0.1	0.3	0.4	✗
impression 1	0.5	-	0.1	0.6	✗
impression 2	0.3	0.5	-	0.8	✓

Figure 2: An overview of our system, which consists of (1) a Transformer encoder-decoder tuning module, (2) a domain adaptation module, (3) an ensemble module, (4) a negative impression normalization module. The domain adaptation module is activated only for test instances in the Indiana subset, and the final normalization module is activated only for test instances in the Stanford subset.

et al., 2020). These prior studies, however, are all based on traditional sequence-to-sequence models like RNN, BiLSTM, as well as pointer-generator network (See et al., 2017), and none of them actually touches the generalization or transfer issue.

In the past few years, pre-training Transformer-based encoder-decoder architectures from large-scale text corpora has been proposed and quickly received massive attention (Radford et al., 2018; Dong et al., 2019; Xiao et al., 2020). Quite a number of such pre-trained models, e.g., MASS (Song et al., 2019), BART (Lewis et al., 2020), and T5 (Raffel et al., 2020), have been devised and proved extremely effective in various language generation tasks. Against this background, we choose PEGASUS (Zhang et al., 2020), a pre-trained model that reports state-of-the-art performance on abstractive text summarization, as the backbone of our system. Since radiology report summarization is a special form of abstractive text summarization, we expect this choice to yield optimal performance. Besides, we employ a simple yet effective domain adaptation strategy, by further fine-tuning on a small amount of in-domain data to improve generalization and transfer abilities. We also use model ensemble and negative impression normalization strategies to further enhance the performance. Figure 2 provides an overview of our system.

With all these strategies, our system achieves an overall ROUGE-2 score of 0.436 on the whole test set, ranked at the 1st place among all participating systems. We will discuss later in the experimental section the performance of different pre-trained models and the effect of each individual strategy.

2 Task Description

This section gives a formal definition of the radiology report summarization task, and introduces data and evaluation metrics used for the task.

2.1 Task Definition

The MEDIQA 2021 Radiology Report Summarization task aims to automatically summarize radiology findings into natural language impression statements. Figure 1 provides an example of a standard radiology report, which consists of a *background*, *findings*, and *impression* section, detailed as below:

- **Background:** This section provides general information about the patient and exam, e.g., clinical history of the patient, type of the exam, and examination techniques. This kind of information helps diagnose diseases when combined with specific findings.

- **Findings:** This section records notable details in each part of the body observed in the exam, after reading an X-ray image. It describes the normality and abnormality a radiologist found in each part of the body. If a specific part was examined but not mentioned, there is probably no obvious abnormality found in that part.

- **Impression:** This section is a concise summarization of the findings written by a radiologist. It lists the patient's symptoms and sometimes with suggested diagnoses. This section is the most crucial part of a radiology report, providing valuable information for doctors to make clinical decisions.

Radiology Report Summarization is to generate the impression given the background and findings. Formally, given a passage of findings represented as a sequence of tokens $\mathbf{x} = \{x_1, x_2, \cdots, x_L\}$ along with the background represented as a sequence of tokens $\mathbf{y} = \{y_1, y_2, \cdots, y_M\}$, the goal is to generate another sequence of tokens $\mathbf{z} = \{z_1, z_2, \cdots, z_N\}$ that best summarizes salient and clinically significant findings in $\mathbf{x}$. Here, L, M, N are the lengths of the findings, the background, and the impression, respectively.

2.2 Official Data

The official data consists of a training split, two validation splits, and two test splits collected from different sources, detailed as follows:

- **Training split:** The training split is composed of 91,544 chest radiology reports picked from MIMIC-CXR database (Johnson et al., 2019). These reports are collected from patients presenting to the Beth Israel Deaconess Medical Center Emergency Department between 2011 and 2016.

- **Validation split I:** The first validation split consists of 2,000 chest radiology reports sampled also from MIMIC-CXR. It therefore has the same distribution with the training split.

- **Validation split II:** The second validation split consists of 2,000 radiology reports sampled from the Indiana chest X-ray collection (Demner-Fushman et al., 2016). These reports are collected from the Indiana Network for Patient Care, thus bearing a risk of inconsistency with the training split.

- **Test split I:** The first test split is also extracted from the Indiana chest X-ray collection, composed of 300 radiology reports in total.

- **Test split II:** The second test split comprises another 300 chest radiology reports collected from the picture archiving and communication system at the Stanford Hospital.

Split	**# Reports**	**Source**
Training	91,544	MIMIC-CXR database
Validation I	2,000	MIMIC-CXR database
Validation II	2,000	Indiana collection
Test I	300	Indiana collection
Test II	300	Stanford collection

Table 1: Statistics and sources of the official data.

The statistics and sources of the data splits are summarized in Table 1. As we can see, both test splits come from different sources with the training split. This poses significant challenges to the generalization and transfer ability of participating systems.

2.3 Evaluation Metrics

The task uses ROUGE (Lin, 2004) to evaluate the performance of participating systems. F1 scores for ROUGE-1, ROUGE-2 and ROUGE-L are reported on the whole test set, and also on the Indiana and Stanford splits. The metrics measure the word-level unigram-overlap, bigram-overlap and the longest common sequence between reference summaries and system predicted summaries respectively. The overall **ROUGE-2** on the whole test set is selected as the primary metric to rank participating systems.

3 Our Approach

We employ a Transformer-based encoder-decoder architecture for radiology report summarization. Our system, as illustrated in Figure 2, consists of four consecutive modules:

- a *Transformer encoder-decoder training* module that fine-tunes a pre-trained language generation model, e.g., PEGASUS (Zhang et al., 2020), on the training split;

- a *domain adaptation* module that further fine-tunes the model on a small amount of validation data coming from the same source with the test split, designed specifically to enhance generalization and transfer ability to unseen data;

- an *ensemble* module that combines diverse predictions from multiple models to generate robust summarization;

- a final *normalization* module that normalizes system predicted negative impressions into a specific form.

Our system is simple yet highly effective, ranked at the 1st place among all participating systems. In the rest of this section, we detail key modules of the system.

3.1 Transformer Encoder-Decoder Training

Transformer-based encoder-decoder architectures pre-trained from large-scale text corpora have recently stood out as the most promising techniques for natural language generation, outperforming the traditional RNN- or LSTM-based opponents in a wide range of language generation tasks (Radford et al., 2018; Raffel et al., 2020). We thereby choose a pre-trained Transformer encoder-decoder model as the backbone of our system, and fine-tunes the model on the training split.

During the fine-tuning process, for each training radiology report, we concatenate the findings $\mathbf{x}$ and background $\mathbf{y}$ into a single sequence, and pair that sequence with the impression $\mathbf{z}$, i.e.,

- Source: $x_1, x_2, \cdots, x_L, [\text{SEP}], y_1, y_2, \cdots, y_M$
- Target: $z_1, z_2, \cdots, z_N$

where $[\text{SEP}]$ is a special token separating the findings and the background. The source sequence is fed into the encoder, and the decoder autoregressively decodes the next token conditioned on the encoder output and previous tokens.

We are free to use any pre-trained Transformer encoder-decoder models. We investigate three representatives: BART, ERNIE-GEN, and PEGASUS, detailed as below.

- **BART** (Lewis et al., 2020) is a denoising autoencoder for sequence-to-sequence learning. It is trained by corrupting text with a noising function, and learning a model to reconstruct the original text. It achieves promising results on a range of abstractive dialogue, question answering, and summarization tasks.
- **ERNIE-GEN** (Xiao et al., 2020) is a multi-flow sequence-to-sequence model that mitigates exposure bias with an infilling generation mechanism and a noise-aware generation method. It achieves comparable results with a smaller number of parameters on several abstractive summarization, question generation, and dialogue response generation tasks.
- **PEGASUS** (Zhang et al., 2020) is a Transformer encoder-decoder model specifically designed for abstractive text summarization. It is trained by masking out important sentences from an input document and generating the masked sentences together from the remaining sentences, similar to an extractive summary. It achieves state-of-the-art performance on 12 summarization tasks spanning across news, science, stories, instructions, emails, patents, and legislative bills.

Model	# Parameters	Corpus Size
BART	400M	160GB
ERNIE-GEN	340M	430GB
PEGASUS	568M	3.8TB + 750GB

Table 2: Number of parameters and size of pre-training corpus of the three models.

Table 2 compares number of parameters and size of pre-training corpus of the three models. PEGASUS gets the largest number of parameters and is trained on the largest amount of data.

3.2 Domain Adaptation

As the test splits (Indiana and Stanford) are collected from different sources with the training split (MIMIC-CXR), participating systems need to address the generalization and transfer issue. Inspired by (Gururangan et al., 2020), we employ a domain adaptation strategy. Specifically, after fine-tuning a pre-trained model on the MIMIC-CXR training set, we further fine-tune the model on a small amount of data similar to the test splits. In this way, we can effectively adapt the model trained from MIMIC-CXR to target test domains.

For the Indiana test split where there is a validation split sampled from the same source, we simply use this validation split for further fine-tuning. After a few epochs over the Indiana validation split, we use the resultant model to make predictions for reports in this test split. As we will show later in the experiments, this adaptation strategy, though conceptually simple, is highly effective, leading to a remarkable boost in ROUGE-2 on this test split.

For the Stanford test split, there is no validation split sampled from the same source. Therefore we construct a subset from the training split to conduct domain adaptation. For each case in this test split (a radiology report without impression), we exploit

Negative Impression	Indiana Freq.	MIMIC-CXR Freq.	Overall Freq.
No acute cardiopulmonary abnormality.	14.2%	4.9%	9.6%
No acute cardiopulmonary process.	3.0%	15.0%	9.0%
No acute cardiopulmonary findings.	6.0%	0.1%	3.1%
No acute cardiopulmonary disease.	0.2%	4.9%	2.6%
No acute cardiopulmonary abnormalities.	4.9%	0.1%	2.5%

Table 3: Top 5 frequent negative impressions and their frequencies on the validation splits.

ElasticSearch[1] to retrieve the top 10 reports from the MIMIC-CXR training split that share the most similar findings. We obtain 2,618 such radiology reports in total after removing duplicates. Then we conduct further fine-tuning on these reports, which, however, downgrades the performance. So we just use the model trained from training split to predict for reports in this test split.

3.3 Model Ensemble

We further employ ensemble that combines diverse predictions from multiple models for robust summarization. Suppose we have T candidate models, e.g., multiple runs with different seeds, each producing a predicted impression $\hat{\mathbf{z}}^i$ $(1 \leq i \leq T)$ for the given findings along with the background. We first compute the mutual similarity score $\mathrm{Sim}(\hat{\mathbf{z}}^i, \hat{\mathbf{z}}^j)$ between each pair of predictions, and aggregate these scores to measure the overall similarity of a specific prediction against all the other predictions:

$$s(\hat{\mathbf{x}}^i) = \sum_{j \neq i} \mathrm{Sim}(\hat{\mathbf{z}}^i, \hat{\mathbf{z}}^j), \quad i = 1, \cdots, T.$$

Then we select the prediction $\hat{\mathbf{z}}^i$ with the highest overall similarity $s(\hat{\mathbf{x}}^i)$ as our final prediction. Figure 2 visualizes this ensemble process. We have tried various similarity scoring functions $\mathrm{Sim}(\cdot, \cdot)$, e.g., ROUGE-1, ROUGE-2, ROUGE-L, and token-level F1, but observed no significant differences between their performance. We finally use ROUGE-1 as the similarity scoring function.

3.4 Negative Impression Normalization

The final normalization module normalizes system predicted negative impressions into a specific form. Roughly speaking, the impression of a radiology report can be divided into two categories: *positive* or *negative*. A positive impression typically reveals symptoms observed during the exam, e.g., "*Mild pulmonary edema and tracebilateral pleural effusions*", whereas a negative impression indicates no symptoms at all, e.g., "*No acute cardiopulmonary abnormality*". Unlike positive impressions which vary drastically w.r.t. input findings, negative impressions tend to be expressed in specific forms. Table 3 presents the top 5 frequent negative impressions and their frequencies on the validation splits. Though expressed in different forms, these negative impressions are all of the same meaning. The choice of a particular form is just a matter of writing style. As the writing style usually varies across organizations, predicting negative impressions by a complex model trained from another organization is prone to over-fitting and may not work well. In contrast, simple heuristics based on basic statistics may lead to less over-fitting and perform better.

Based on this observation, we introduce a heuristic strategy, i.e., for any negative prediction starting with "No acute", we normalize it into "No acute cardiopulmonary abnormality", which is the most frequent negative impression in the validation sets. This normalization process is carried out only for the Stanford test split, for which there is no training or validation set from same organization.

4 Experiments and Results

This section presents experiments and results of our system on the official data.

4.1 Experimental Setups

Our system is built upon a pre-trained Transformer encoder-decoder architecture, PEGASUS (Zhang et al., 2020). The maximum lengths of source and target sequences are restricted to 512 and 128 respectively, covering 99% of the cases in the training and validation splits. Throughout all experiments, we employ a decoding process with beam size of 5, length penalty of 0.8, and early stopping.

Fine-tuning Setup We first fine-tune PEGASUS-large[2] on the MIMIC-CXR training split. We tune

[1] https://www.elastic.co

[2] https://huggingface.co/google/pegasus-large

Rank	Team	All Test Set ROUGE-1/-2/-L			Indiana Test Set ROUGE-1/-2/-L			Stanford Test Set ROUGE-1/-2/-L		
1	**BDKG (Ours)**	**.5573**	**.4362**	**.5366**	**.6834**	**.5956**	**.6717**	**.4312**	.2769	**.4014**
2	IBMResearch	.5328	.4082	.5134	.6772	.5881	.6657	.3884	.2284	.3611
3	optumize	.5186	.3918	.4957	.6188	.5182	.6050	.4183	.2655	.3864
4	JB	.4955	.3778	.4794	.5895	.5039	.5824	.4015	.2517	.3763
5	low_rank_AI	.4716	.3311	.4487	.5129	.3846	.5026	.4302	**.2777**	.3948
6	med_qa_group	.4642	.3265	.4440	.5051	.3774	.4965	.4233	.2757	.3916
7	ChicHealth	.4606	.3236	.4411	.5070	.3782	.4984	.4143	.2690	.3838
8	hEALTHai	.4481	.3084	.4273	.4845	.3527	.4752	.4118	.2641	.3794
9	DAMO_ali	.4330	.2763	.4116	.4371	.2839	.4278	.4289	.2687	.3954
10	I_have_no_flash	.4303	.2743	.4092	.4351	.2826	.4258	.4256	.2661	.3926

Table 4: Official results of top 10 systems on the test splits. Systems ranked by ROUGE-2 on the whole test set.

the initial learning rate $\in \{1e{-}5, 3e{-}5, 6e{-}5, 1e{-}4\}$, batch size $\in \{8, 16, 32\}$, and number of epochs $\in \{5, 10, 15, 25\}$. Other hyper-parameters are fixed to their default values. The optimal configuration is determined by ROUGE-2 on the whole validation set (a combination of the MIMIC-CXR and Indiana splits), which is learning rate $= 6e{-}5$, batch size $= 8$, and number of epochs $= 15$.

Domain Adaptation Setup We further fine-tune the model derived above on the Indiana validation split, so as to adapt the model from MIMIC-CXR to our target test domain. Specifically, we split the Indiana validation set into 1700 : 300 subsets. We tune the model with initial learning rate $\in \{1e{-}4, 2e{-}4, 4e{-}4\}$, batch size $\in \{8, 16\}$, and number of epochs $\in \{10, 20, 50, 100\}$ on the former, and determine the optimal configuration on the latter (by ROUGE-2). The optimal configuration is initial learning rate $= 2e{-}4$, batch size $= 8$, and number of epochs $= 100$, with other hyper-parameters set, again, to their default values. After determining the optimal configuration, we re-tune the model on the whole Indiana validation set.

Ensemble Setup We ensemble 16 models further fine-tuned with in-domain data for the Indiana test split. These models are obtained with the same optimal configuration determined during domain adaptation, but different random seeds. We ensemble another 15 models trained from MIMIC-CXR training split for the Stanford test split. These models are obtained, again, with the same configuration but different seeds.

4.2 MEDIQA 2021 Official Results

Table 4 shows the official results of top 10 participating systems on the test splits, where systems are ranked by ROUGE-2 score on the whole test set. Our system, though conceptually simple, is highly effective, ranked the 1st place among participating systems. Notably, it consistently outperforms the other systems across all three test splits and almost in all metrics.

4.3 Further Analyses

This section provides in-depth analyses to show the effect of each individual module in our system.

Effect of Pre-trained Models We first examine the effect of different pre-trained models. Specifically, besides PEGASUS-large, we consider other pre-trained models including BART[3], DistilBART[4], ERNIE-GEN[5], and PEGASUS-xsum[6], all in the "large" setting. We tune their hyper-parameters in the same ranges as in PEGASUS-large, and report optimal results on the validation splits. The results are summarized in Table 5, where (S) scores denote results for single models averaged over five runs. Among these models, the two PEGASUS variants (-large and -xsum), which are designed specifically for abstractive text summarization, consistently perform better. And the -large variant performs even better than the -xsum one. The reason may be that the -xsum variant has been further tuned on XSum

[3] https://huggingface.co/facebook/bart-large

[4] https://huggingface.co/sshleifer/distilbart-xsum-12-6

[5] https://github.com/PaddlePaddle/ERNIE/tree/repro/ernie-gen

[6] https://huggingface.co/google/pegasus-xsum

	All Valid Set			MIMC-CXR Valid Set			Indiana Valid Set		
Model	**ROUGE-1/-2/-L**			**ROUGE-1/-2/-L**			**ROUGE-1/-2/-L**		
BART (S)	.5352	.3871	.5103	.6209	.4902	.5865	.4495	.2840	.4340
BART (E)	.5535	.4057	.5284	.6425	.5125	.6077	.4644	.2989	.4491
DistilBART (S)	.5456	.3987	.5214	.6385	.5109	.6055	.4526	.2865	.4372
DistilBART (E)	.5604	.4144	.5360	.6516	.5244	.6189	.4691	.3043	.4531
ERNIE-GEN (S)	.5385	.3951	.5167	.6237	.4996	.5926	.4532	.2905	.4409
ERNIE-GEN (E)	.5476	.4035	.5229	.6313	.5070	.6002	.4638	.3000	.4515
PEGASUS-xsum (S)	.5506	.4107	.5303	.6413	.5233	.6117	.4600	.2981	.4489
PEGASUS-xsum (E)	.5566	.4172	.5361	.6441	.5266	.6141	.4691	.3078	.4581
PEGASUS-large (S)	.5559	.4129	.5330	.6511	.5290	.6188	.4608	.2968	.4471
PEGASUS-large (E)	**.5649**	**.4224**	**.5413**	**.6572**	**.5329**	**.6235**	**.4725**	**.3088**	**.4591**

Table 5: Results of different pre-trained models on validation splits. We run each model five times with different seeds under its optimal configuration. (S)/(E) respectively denotes the averaged/ensemble results of the five runs.

	All Test Set			Indiana Test Set			Stanford Test Set		
Ablation	**ROUGE-1/-2/-L**			**ROUGE-1/-2/-L**			**ROUGE-1/-2/-L**		
Full Model	**.5573**	**.4362**	**.5366**	**.6834**	**.5956**	**.6717**	**.4312**	**.2769**	**.4014**
– Domain Adaptation	.4539	.2916	.4333	.4766	.3062	.4652	**.4312**	**.2769**	**.4014**
– Normalization	.5487	.4221	.5281	**.6834**	**.5956**	**.6717**	.4139	.2486	.3844

Table 6: Ablation results of domain adaptation and negative impression normalization on test splits.

(Narayan et al., 2018), which consists of articles from the British Broadcasting Corporation and exhibits drastic distinctions from radiology reports. This thereby may result in catastrophic forgetting.

Effect of Ensemble We further investigate the effect of model ensemble. To this end, for each of the pre-trained models considered above, we run the model five times with its optimal configuration but different seeds. We then compare performance of the single model (S) and the ensemble (E) on the validation splits, and report the results in Table 5. We can see that ensemble is a generally effective strategy, leading to about 1% to 2% gains across all data splits and metrics, not matter which pre-trained model is used.

Effect of Domain Adaptation We then evaluate the effect of our domain adaptation module, which is applied solely to the Indiana test split. We consider an ablation that uses the model trained from MIMIC-CXR to predict on both Indiana and Stanford test splits, without further fine-tuning on the in-domain Indiana validation split. Table 6 reports the performance of this ablation on the test splits, and makes comparisons to the full model. We can see that the adaptation module, though conceptually simple, is extremely useful, pushing the ROUGE-2 score drastically from 0.3062 to 0.5956 on Indiana test split.

Effect of Normalization We finally evaluate the effect of negative impression normalization, which is applied solely to the Stanford test split. Table 6 compares performance with and without this final normalization strategy on the test splits. We can see that this simple strategy brings meaningful gains, pushing the ROUGE-2 score from 0.2486 to 0.2769 on Stanford test split.

5 Conclusion

This paper presents our winning system at the Radiology Report Summarization track of the MEDIQA 2021 shared task. Participating systems in this track are required to summarize radiology findings into natural language impressions, and be able to generalize or transfer to reports collected from previously unseen hospitals. We build our system on the basis of a pre-trained Transformer encoder-decoder architecture, namely PEGASUS. We further employ a domain adaptation module to enhance generalization and transfer ability. Heuristics such as ensemble and negative impression normalization are also used. Our system finally achieves a ROUGE-2 score of 0.436 on the test set, ranked the 1st place among all participating systems.

Acknowledgements

We would like to thank the MEDIQA 2021 organizers for the challenging and exciting tasks. This work is supported by the National Key Research and Development Program of China No.2020AAA0109400.

References

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th SIG-BioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021.*

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

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197.*

Esteban F Gershanik, Ronilda Lacson, and Ramin Khorasani. 2011. Critical finding capture in the impression section of radiology reports. In *AMIA Annual Symposium Proceedings*, pages 465–469.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1899–1905.

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

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):1–8.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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

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

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *EMNLP*, pages 1797–1807.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, Technical report, OpenAI.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5926–5936.

Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE-GEN: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3997–4003.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213.

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