A New Dataset for Summarizing Radiology Reports

Anonymous ACL submission

Abstract

The radiology report summarization is an important technology in smart healthcare. Compared with medical image processing and disease recognition which have been comprehensively studied, the research on radiology report summarization is much limited, which is mainly due to the lack of a high-quality benchmark dataset. In this paper, we present a dataset called CRRsum for radiology report summarization, where it is constructed from over 10K real radiology reports that contains diagnostic findings and diagnostic opinions. An extensive evaluation is performed with the current state-of-the-art methods for radiology report summarization on our proposed dataset. Our experiments reveal the challenges of radiology report summarization and provide many opportunities for research going forward. We also show that the CRRsum can be used in medical classification to facilitate the research in this task.

1 Introduction

The application of smart healthcare technology, such as medical Q&A (He et al., 2020; Wang et al., 2020), disease recognition (Ji et al., 2021), medical image processing (Yang et al., 2021), etc., can effectively alleviate the medical resource shortage. As a crucial component of smart healthcare, the radiology report summarization has important implications: it can automatically summarize critical findings in the radiology report using summary generation technology to provide an accurate and concise description of the patient's disease. An important clinical value can be derived from this task since it has the potential to speed up radiology workflow, decrease repetitive human labor, and positively alleviate healthcare resource shortages (Zhang et al., 2019).

A standardized radiology report is made up of a Finding section and an Opinion section, as shown in Table 1. A typical workflow requires that the radiologist first dictate the radiology examinations' detailed findings into the Finding section and then summarizes the salient findings into the more concise Opinion section (Kahn Jr et al., 2009). This is similar to the traditional summary generation model, where it compresses the finding into the opinion that is a concise description covering its key facts (Zhang et al., 2020a; Liu et al., 2019b). However, compared with the traditional summary generation, which has been comprehensively studied, the research on radiology report summarization is limited, mainly because of the absence of a reliable benchmark dataset.

Diagnostic findings: 左足CT平扫+三维重建左足第3、4跖骨远端骨皮质断裂、皱褶;余诸骨未见明显骨折。诸小关节在位。 (The left foot CT plain scan + three-dimensional reconstruction of the left foot 3rd and 4th metatarsal distal bone cortical fractures, folds; no obvious fractures in the remaining bones. The small joints are in place.)
Diagnostic opinions: 左足第3、4跖骨远端骨折。(The third and fourth metatarsals of the left foot were fractured.)

Table 1: An example of radiology report summarization, which is the standard form of radiology report in China.

A high-quality dataset can significantly facilitate the research in an area, such as ImageNet for image classification (Deng et al., 2009) and Microsoft COCO Captions for image captioning (Chen et al., 2015). There are several public datasets for traditional summary generation tasks, such as LCSTS (Hu et al., 2015) and Gigaword (Nallapati et al., 2016) datasets. Based on these datasets, many well-known summary generation methods have been developed. However, existing studies on radiology report summarization are much fewer, and many of them are conducted on proprietary datasets. Thus, a public high-quality radiology report summarization dataset is of great value for the research in this area.

To this end, our paper proposes a novel dataset for radiology report summarization (called CRRsum), which is collected from real radiology reports. It contains more than 10K reports, and each report includes diagnostic findings and diagnostic opinions. We implement many state-of-the-art summary generation methods originally developed on different publicly datasets, and compare their performances on the CRRsum dataset to provide a benchmark for radiology report summarization research. The experimental results of different state-of-the-art summary generation models show that a deep understanding of diagnostic reports through NLP techniques is important for radiology report summarization. Both effective diagnostic findings representation approaches and pre-trained language models can contribute to the performance improvement of the radiology report summarization. We hope CRRsum can serve as a benchmark dataset for radiology report summarization and facilitate the research in this area.

In summary, our contributions are listed as follows:

- We release a radiology report summarization dataset, which includes more than 10K real radiology reports, and covers 15 categories of body part diseases. CRRsum is the only Chinese radiology report summarization dataset currently open access.
- We report results for several summary generation approaches on the CRRsum, and compare their performance using automatic metrics. Through experiments, we find that the NEZHA models can significantly improve performance on radiology report summary generation task.
- In addition to the radiology report summary generation task, the CRRsum dataset can also be used for the disease classification task, and we report the results.
- We demonstrate the feasibility and prospect of the NLP technologies in the domain of radiology and smart healthcare.

2 Related Work

Most prior studies attempt to classify and extract diseases information from the diagnostic findings to “summarize" radiology reports (Hripcsak et al., 2002). In recent studies, Hassanpour and Langlotz (2016) investigated which named entities can be extracted from multi-institutional radiology reports using traditional feature-based classification methods. Goff and Loehfelm (2018) developed an NLP model to identify the description of the disease entities in the Opinion section of radiology reports to support the report summarization. Cornegruta et al. (2016) used a BiLSTM neural network architecture to address questions about the disease negation detection and entity recognition on radiology reports. Zhang et al. (2018) first attempted the generation of diagnostic opinions based on the summary generation technology and showed that their model is highly correlated with the reference opinions. MacAvaney et al. (2019) proposed a radiology report summary model based on the ontology-aware network and demonstrated better diagnostic opinions. Liu et al. (2019a) proposed an RL-based model to generate textual descriptions of diagnostic findings from medical images. Zhang et al. (2018) showed that the radiology summaries generated from NLP models contain many factual errors, improving factual correctness in radiology summaries by reinforcement learning. Zhang et al. (2020a) explored using question-focused dual attention to summarize medical answers. Cai et al. (2021) proposed the ChestXRAYBERT model to summarize chest report summaries automatically. In addition, some radiology report datasets combining images are worthy of attention, such as MIMIC-CRX (Johnson et al., 2019), ME-DIA (Abacha et al., 2021), Padchest (Bustos et al., 2020), Rad-SpRL (Datta and Roberts, 2020), and others (Wang et al., 2018; Demner-Fushman et al., 2016). These datasets contribute significantly to the study of the radiology report summarization.

To our knowledge, most of the existing studies on radiology report summarization are based on English datasets and are not publicly available. Our work has made the first attempt at automatic summarization of Chinese radiology reports and is freely available. The lack of datasets has hampered progress in developing radiology report summarization research, and we hope that our CRRsum dataset will facilitate this progress.

3 CRRsum Dataset

In this section, we first present the CRRsum dataset that includes data creation and processing procedures. Then, we also report statistical analyses and a human evaluation.

3.1 Dataset Creation

In order to facilitate the research in radiology report summarization, we built the radiology report summarization dataset (CRRsum)[1]. It was created by real radiology reports and collected from the hospital radiology department[2]. All reports were collected in 2021, and the radiological examination method of patients is Computed Tomography (CT), which included 15 body parts, such as the head and lumbar spine. Radiologists marked the body part of the CT examination in each data to the distinction between different report categories. This means that each piece of data in CRRsum will be constructed by a diagnostic finding, a diagnostic opinion, and a category.

Diagnostic findings. As the input of the model, the following should be considered for coverage in the diagnostic findings: 1) the examination method used by the radiologist; 2) the body parts of the patient examined by the radiologist; 3) a description of the findings of the examined disease; 4) a focused description of the abnormalities.

Diagnostic opinions. As the model's output, the diagnostic opinions need to cover the major facts in the diagnostic findings. According to standards and specifications of radiology report writing (Zhihui Shen and Ruimin, 2019), the diagnostic opinions provide a judgment on the disease condition. It generates a reasonable recommendation to patients, such as recommending further examination and requesting a diagnosis in the context of the clinic.

Category. The CRRsum dataset contains 15 categories, covering the main body parts for radiological examinations.

The diagnostic finding, diagnostic opinion, and category in each radiology report are written and annotated by radiologists, making them clinically useful.

3.2 Data Processing

We carefully construct the CRRsum dataset to maximize its usability. The build process includes: 1) hiding the personal information; 2) extracting the radiology report content; 3) cleaning the data.

- The preprocessing of each radiology report is necessary to protect the patient's privacy. Also, to prevent the influence of irrelevant information, we removed personal information and kept only the diagnostic findings and the opinions, as shown in Table 1. In other words, the radiology report we received contained only diagnostic findings and opinions. These two sections are limited to the patient's condition and do not involve patient privacy.

- Efficient text extraction is crucial to the construction of the CRRsum dataset, as it affects the quality of the diagnostic opinions generated by the model. Tencent's OCR technology was selected after comparison.

- Following the standards and specifications for writing radiology reports (Niederkohr et al., 2013), we perform the review and verification of data through medical professionals. The purpose is to deal with meaningless characters and correct errors.

It is worth noting that all medical datasets inevitably involve patient privacy issues, such as Standford reports containing patient background information. In contrast, in the CRRsum dataset, all data include only diagnostic findings and diagnostic opinions and do not contain any patient's privacy. Therefore, the CRRsum dataset does not have any risk of revealing patients' privacy. Moreover, two medical professionals re-checked the data to ensure that the processed data were available.

3.3 Dataset Statistics and Analysis

The detailed statistics of the CRRsum dataset are summarized in Table 2 and Fig. 1. This dataset contains 10,066 real radiology reports. There are 8,136 (80.83%) samples in the training set, 901 (8.95%) samples in the validation set, and 1,029 (10.22%) samples in the test set. In more detail, the different categories of reports we divide in the ratio of close to 8:1:1, which can empower the training of the radiology report summarization models.

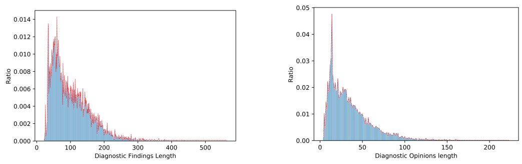
Figs. 1(a) and 1(b) show the length distributions of diagnostic findings and opinions. We can see that the average lengths of the diagnostic findings and opinions are 100 and 35, respectively. Most of the radiology reports are under 300 characters, and

[1] The dataset that support the findings of this study are available from the corresponding author upon reasonable request.

[2] We purchase through individuals and are committed to protecting patient privacy.

Traing set	8,135	Validation set	901
Test set	1,029	Category	15
Max find. len.	563	Min find. len.	22
Avg. find. len.	**100.7**	Find. S.D.	57.23
Max opin. len.	223	Min opin. len.	4
Avg. opin. len.	**35.6**	Opin. S.D.	25.48
New word	**32.22%**	Split M.	Random

Table 2: Detailed statistics of the CRRsum dataset.



(a) findings length (b) opinions length

Figure 1: Key statistics of the CRRsum dataset.

the diagnostic opinions are under 100 characters, which is in line with the radiology report writing standards (Zhihui Shen and Ruimin, 2019). It is necessary to note that in Table 2, we present the percentage of new words appearing in the diagnostic opinions as 32.2% (words that do not appear in the same finding are considered new), which suggests that the CRRsum dataset is more suitable for abstractive approaches (Lu et al., 2020).

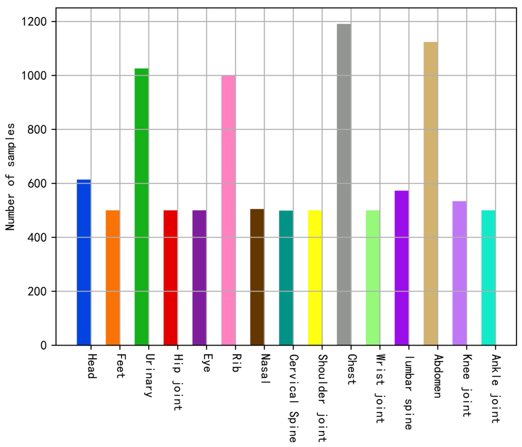


Figure 2: The distribution of the radiology report types in CRRsum.

In Fig. 2, we show the distribution of the radiology report types in CRRsum. As shown in Fig. 2, the number of Chest, Urology, and Abdomen reports is higher than other reports. In addition, we further show the distribution of each type of disease in the training set, validation set, and test set. As shown in Table 3, the training set, validation set, and test set of the CRRSum dataset have similar distributions, which is beneficial to test the performance of the radiology report summarization model and promote the development of this task.

Class	Total	Train	Val.	Test
头 Head	614 6.10%	509 6.25%	53 5.88%	52 5.05%
脚部 Feet	500 4.97%	397 4.88%	45 4.99%	58 5.63%
泌尿 Urology	1026 10.19%	841 10.33%	90 9.98%	95 9.23%
髋关节 Hip.	500 4.97%	379 4.65%	52 5.77%	69 6.70%
眼部 Eye	500 4.97%	423 5.19%	35 3.88%	42 4.08%
肋骨 Ribs	1000 9.93%	774 9.51%	104 11.54%	122 11.85%
鼻腔 Nose	505 5.02%	381 4.68%	53 5.88%	71 6.90%
颈椎 Cervical.	499 4.96%	396 4.86%	51 5.66%	52 5.05%
肩关节 Shoulder.	500 4.97%	386 4.74%	68 7.54%	46 4.47%
胸腔 Chest	1191 11.83%	1001 12.30%	95 10.54%	95 9.23%
腕关节 Wrist.	500 4.97%	410 5.03%	45 4.99%	45 4.37%
腰椎 Lumbar.	573 5.69%	467 5.74%	50 5.54%	56 5.44%
腹部 Abdomen	1124 11.17%	935 11.49%	54 5.99%	135 13.12%
膝关节 Knee.	534 5.3%	432 5.31%	51 5.66%	51 4.95%
踝关节 Ankle.	500 4.97%	405 4.97%	55 6.10%	40 3.887%

Table 3: The number and percentage of different types of radiology reports in training set, validation set, test set.

In addition, to get a clearer picture of the composition of the CRRsum dataset, we show a heat map of the length distribution of different categories of radiology reports. In Fig. 3, we observe that in different categories of diagnostic findings, the length is usually under 200 characters. Also, the Abdominal and Chest diagnostic findings are longer than other diagnostic findings because the examination of this body part contains more diseases, which correspond to the actual situation.

3.4 Human Evaluation of Datasets

We randomly selected 30 radiology reports from CRRsum and evaluated the disease description consistency between the diagnostic findings and opin-

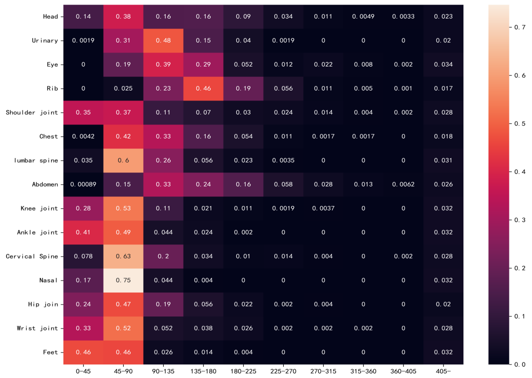


Figure 3: Heat map of the length distributions of different categories of diagnostic findings

ions by medical professionals. Each report was scored using the measure in Table 4.

Consistency	Criteria	Score
perfect consistent	75% - 100%	4
major consistent	50% -75%	3
partial consistent	25% - 50%	2
poor consistent	less than 25%	1

Table 4: Human evaluation criteria. When the description of the diagnostic opinion was perfectly consistent with the diagnostic finding (75%-100%), this report scored 4.

By evaluation, we obtained the average quality score of CRRsum is 3.51. There is a high consistency between the reference opinions and the diagnostic findings based on this score, highlighting that the diagnostic findings are covered despite only using the diagnostic opinions, which can empower the CRRsum dataset to serve as a benchmark. (Lu et al., 2020).

4 Experiments

In this section, several state-of-the-art models have been evaluated using the CRRsum dataset to determine their performance. An in-depth analysis of the quality of the opinion is also provided, including both quantitative and qualitative analysis in addition to the statistical analysis.

4.1 Model

For extractive, we used four commonly models, LDA (Blei et al., 2003), Lead-3, Textrank (Mihalcea and Tarau, 2004) and BERTSUM (Liu, 2019), as baselines. About the abstractive model, we test LSTM (Su, 2018) and Pointer-Generator (See et al., 2017), where the LSTM model used a bidirectional long-short term memory network as the encoder. Furthermore, we apply several state-of-the-art pre-trained models for radiology report summary generation, including BERT (Kenton and Toutanova, 2019), ALBERT (Lan et al., 2020), NEZHA (Junqiu Wei, 2019), MT5 (Xue et al., 2021), BERT-wwm (Cui et al., 2020), WoBERT (Su, 2020), RoBERTa-wwm (Liu et al., 2019c), and MC-BERT (Zhang et al., 2020b), where MC-BERT is a pre-trained model based on medical data. We hope that the experiments with pre-trained language models can provide a useful benchmark for diagnostic report summarization.

4.2 Experimental Setting

In our experiments, we verified and compared all the models presented in Section 4.1 on the CRRsum dataset. Adam (Kingma and Ba, 2014), EMA-Adam (Yu et al., 2018) and Adagrad (Duchi et al., 2011) are used as optimizers. In the decoding stage, beam search is used. The maximum input and output sequence lengths of the model are 512 and 64. In the pre-trained language model, the early stopping strategy is used, the maximum training epoch of the model is 35, the learning rate is 10^{-5}. We validate the model at the end of each epoch to save the best checkpoint. The diagnostic opinions quality evaluation metrics are used ROUGE (Lin and Hovy, 2003) and BLEU (Papineni et al., 2002).[3]

4.3 Result Analysis

We report the ROUGE and BLEU Scores for different models on the CRRsum dataset in Table 5. We note that, when we compare abstractive models to extractive ones, all abstractive models are superior to extractive models—LDA, Lead-3, Textrank, and BERTSUM—by wide margins. Additionally, in terms of ROUGE-L, each of the abstractive models outperformed the extractive oracle significantly. This is consistent with the analysis in Section 3.3, which further shows the suitability of CRRsum for abstractive approaches.

Pre-trained language models such as MT5, NEZHA, and WoBERT usually perform better than Pointer-Generator model. This is because these models are pretrained on a large collection of corpora before being finetuned on CRRsum. Pretraining enables the model to better capture the linguistic structure among words, which yields higher ROUGE and BLEU Scores. In addition, we also

[3]The code used in this study will be open source.

	Model	Optimizer	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Extractive	LDA	None	28.99	19.32	27.70	11.51
	Lead-3	None	34.82	22.92	32.53	14.76
	Textrank	None	37.34	25.76	35.26	17.42
	BERTSUM	Adam	42.01	29.65	38.52	17.34
Abstractive	Point-Gen.	Adagrad	64.28	50.95	62.05	26.33
	LSTM	Adam	68.31	57.26	68.59	47.05
	ALBERT-small	Adam	68.40	58.32	69.09	47.28
	ALBERT-Xlarge	Adam	75.75	66.24	70.48	55.19
	MC-BERT	Adam	76.73	67.63	75.65	56.90
	BERT	Adam	76.78	67.61	75.57	56.82
	BERT-wwm	Adam	76.96	67.85	75.98	56.55
	RoBERTa-wwm	Adam	77.36	68.20	76.29	57.62
	WoBERT	Adam	77.87	68.86	76.60	**58.07**
	MT5	Adam	**77.88**	67.87	74.75	56.58
	NAZHA	Adam	77.79	**68.88**	**76.76**	57.67
	ALBERT-small	EMA-Adam	69.73	59.46	69.93	48.31
	ALBERT-Xlarge	EMA-Adam	76.62	67.25	75.13	56.26
	MC-BERT	EMA-Adam	76.40	67.27	75.36	55.79
	BERT	EMA-Adam	76.72	67.53	75.35	56.44
	BERT-wwm	EMA-Adam	76.87	67.89	75.72	57.13
	RoBERTa-wwm	EMA-Adam	77.84	68.78	76.50	**57.82**
	WoBERT	EMA-Adam	77.93	68.82	76.70	57.55
	MT5	EMA-Adam	76.72	66.93	74.38	55.42
	NAZHA	EMA-Adam	**77.96**	**69.03**	**76.86**	57.62

Table 5: ROUGE and BLEU results on CRRsum test set.

compare the models under different optimizers. It is not difficult to find that Adam with an exponential moving average works better than Adam in most pre-trained models. To our surprise, the

Class	R.-1	R.-2	R.-L	BLEU
Head	73.09	65.07	73.36	53.45
Feet	78.63	71.87	79.67	63.26
Urology	70.62	59.16	71.33	46.30
Hip.	62.65	52.19	64.85	42.17
Eye	59.45	51.62	67.76	40.26
Ribs	53.39	43.07	55.97	31.96
Nose	70.78	61.43	73.13	49.98
Cervical.	73.60	62.52	73.77	51.07
Shoulder.	66.74	57.58	66.86	46.45
Chest	41.34	31.94	52.73	21.16
Wrist.	78.76	70.44	79.79	59.02
Lumbar.	73.24	64.41	74.97	52.71
Abdomen	66.26	54.19	65.89	39.90
Knee.	71.77	61.46	73.23	49.54
Ankle.	79.76	72.88	78.40	65.18

Table 6: ROUGE and BLEU results on single-category radiology reports.

performance of LSTM is close to the ALBERT-small model. Although ALBERT has a significant advantage over other pre-trained language models in decoding rate, generating high-quality diagnostic opinions is challenging when the model size is small. Moreover, as the model size increases, the performance improves. As shown in Table 5, ALBERT-Xlarge outperforms LSTM.

We report the experimental results for single-category radiology reports in Table 6. For the pre-trained language model, we used BERT. We found that although the numbers of samples for the Abdomen and Chest are larger than other reports, its effect was not outstanding. The reason for this fact is, as described in Section 3.3, that the Abdomen and Chest reports contain multiple diseases and the diagnostic findings are longer, which is a challenge for the model to generate diagnostic opinions. In contrast, the shorter diagnostic findings are easier to generate high-quality opinions. As shown in Fig. 4, the ROUGE-1 score showed a decreasing trend as the length of the diagnostic finding increased.

To get a step further analysis of the quality of diagnostic opinions, we show a radiology report summarization example in Table 8. Since the extractive model is copied from the diagnostic findings, the generated diagnostic opinions fail to resemble the writing standards despite capturing the correct content. In contrast, the abstractive models can adhere to the radiology report writing standards, and their

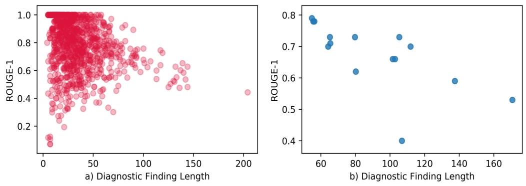


Figure 4: ROUGE-1 scores and diagnostic findings length distribution of the test set. Subplot a) represents the ROUGE-1 of diagnostic findings of different lengths; subplot b) represents the average length and ROUGE-1 of diagnostic findings of different categories.

diagnostic opinions are also the correct content.

5 Extensions of CRRsum dataset

We focus on diagnostic opinions from the diagnostic findings, but our dataset could also be used for another task: disease classification. Disease classification has the potential clinical value of accelerating the patient access process. In the CRRsum dataset, we use diagnostic findings as input to the classification model, and the output of the model is the disease category.

We apply several benchmark classification models to the CRRsum dataset and briefly report the results. The classification models include RNN (Liu et al., 2016), Transformer (Vaswani et al., 2017), BERT (Kenton and Toutanova, 2019), RoBERTa (Liu et al., 2019c), NEZHA (Junqiu Wei, 2019), ALBERT (Lan et al., 2020) and MC-BERT (Zhang et al., 2020b).

Model	Validation set	Test set
RNN	80.69%	82.51%
Transformer	87.35%	89.02%
ALBERT	92.45%	91.83%
RoBERTa	92.45%	91.64%
BERT	**93.56%**	92.22%
NEZHA	92.89%	92.80%
MC-BERT	93.23%	**93.58%**

Table 7: Disease classification results on CRRsum dataset.

As shown in Table 7, all the pre-trained language models outperform the RNN, and the MC-BERT achieves the best results. The results also show that our CRRsum dataset can be used for disease classification tasks. As we expect, CRRsum could advance the development of smart medical-related tasks.

Diagnostic findings: 左肺上叶(薄层Im44)见一直径约2mm小结节影，境界清晰;余两肺纹理增多。气管、支气管通畅。纵隔内未见明显肿大淋巴结。心影大小、形态正常。两侧胸腔未见明显积液。(A small nodular shadow of about 2 mm in diameter was seen in the upper lobe of the left lung (thin layer Im44) with clear boundaries; the remaining two lungs had increased texture. The trachea and bronchi were patented. No obvious enlarged lymph nodes were seen in the mediastinum. The heart shadow was normal in size and shape. There was no obvious fluid accumulation in both chest cavities.)
Diagnostic opinions: 左肺上叶微小结节，建议12个月复查。(Small nodule in the upper lobe of the left lung, recommended for review at 12 months.)
Textrank: 左肺上叶(薄层Im44)见一直径约2mm小结节，影境界清晰。两侧胸腔未见明显积液。(A small nodular shadow of about 2 mm in diameter was seen in the upper lobe of the left lung (thin layer Im44) with clear boundaries. There was no obvious fluid accumulation in both chest cavities.)
RoBERTa-wwm: 左肺上叶小结节，随诊复查。(A small nodule in the upper lobe of the left lung is recommended for follow-up review.)
NEZHA: 左肺上叶小结节，建议6-9个月复查。(Small nodules in the upper lobe of the left lung, with a 6- to 9-month review recommended.)

Table 8: Examples of radiology report summarization.

6 Conclusion

The lack of a dataset has impeded progress in radiology report summarization research. This paper introduced CRRsum, a dataset for radiology report summarization. We extensively evaluated several state-of-the-art models for diagnostic opinions generation on the CRRsum dataset. Experimental results show that our dataset can be an important benchmark in developing and evaluating summary generation approaches to radiology reports. We also show that the CRRsum can be used as a dataset for disease classification. More importantly, our work demonstrates the feasibility and promise of the language model to the domain of radiology and smart healthcare fields.

A further study focusing on improving the accurate description of the disease in the summary of radiology reports is suggested.

References

Asma Ben Abacha, Yassine M'rabet, 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 Workshop on Biomedical Language Processing*, pages 74–85.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Xiaoyan Cai, Sen Liu, Junwei Han, Libin Yang, Zhenguo Liu, and Tianming Liu. 2021. Chestxraybert: A pretrained language model for chest radiology report summarization. *IEEE Transactions on Multimedia*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Piotr Gupta, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *IEEE Conference on Computer Vision and Pattern Recognition*.

Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. 2016. Modelling radiological language with bidirectional long short-term memory networks. *EMNLP 2016*, page 17.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Surabhi Datta and Kirk Roberts. 2020. A dataset of chest x-ray reports annotated with spatial role labeling annotations. *Data in Brief*, 32:106056.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

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

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

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614.

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

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972.

Zongcheng Ji, Tian Xia, Mei Han, and Jing Xiao. 2021. A neural transition-based joint model for disease named entity recognition and normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2819–2827.

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.

Xiaoguang Li Junqiu Wei, Xiaozhe Ren. 2019. Nezha: Neural contextualized representation for chinese language understanding. *arXiv preprint arXiv:1909.00204*.

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

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ICLR2014*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *ICLR 2020*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Guanxiong Liu, Tzu-Ming Harry Hsu, and McDermott. 2019a. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *International Joint Conference on Artificial Intelligence*.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Yang Liu, Ivan Titov, and Mirella Lapata. 2019b. Single document summarization as tree induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1745–1755.

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

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074.

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.

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

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Ryan D Niederkohr, Bennett S Greenspan, John O Prior, Heiko Schöder, Marc A Seltzer, and Zukotynski. 2013. Reporting guidance for oncologic 18f-fdg pet/ct imaging. *Journal of Nuclear Medicine*, 54(5):756–761.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *Proceedings of Association for Computational Linguistics*.

Jianlin Su. 2018. keras example of seq2seq, auto title. Technical report.

Jianlin Su. 2020. Open language pre-trained model zoo - zhuiyiai. Technical report.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, and Kaiser. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058.

Xing David Wang, Leon Weber, and Ulf Leser. 2020. Biomedical event extraction as multi-turn question answering. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 88–96.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xingyi Yang, Muchao Ye, Quanzeng You, and Fenglong Ma. 2021. Writing by memorizing: Hierarchical retrieval-based medical report generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *International Conference on Learning Representations*.

Ningyu Zhang, Shumin Deng, Juan Li, Xi Chen, Wei Zhang, and Huajun Chen. 2020a. Summarizing chinese medical answer with graph convolution networks and question-focused dual attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 15–24.

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020b. Conceptualized representation learning for chinese biomedical text mining. *ACM International Conference on Web Search and Data Mining*.

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

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. *Association for Computational Linguistics*.

Xubai Xuan Zhihui Shen and Wang Ruimin. 2019. The standards for pet/ct diagnostic reports: Setting and exploring. *Labeled Immunoassays and Clinical Medicine*, pages 1614–1617.