Fully automatic summarization of radiology reports using natural language processing with large language models

Nishio, Mizuho ; Matsunaga, Takaaki ; Matsuo, Hidetoshi ; Nogami, Munenobu ; Kurata, Yasuhisa ; Fujimoto, Koji ; Sugiyama, Osamu ;…

(Citation)
Informatics in Medicine Unlocked,46:101465

(Issue Date)
2024

(Resource Type)
journal article

(Version)
Version of Record

Informatics in Medicine Unlocked 46 (2024) 101465

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Fully automatic summarization of radiology reports using natural language processing with large language models

Mizuho Nishio [a,*], Takaaki Matsunaga [a], Hidetoshi Matsuo [a], Munenobu Nogami [a,b], Yasuhisa Kurata [c], Koji Fujimoto [d], Osamu Sugiyama [e], Toshiaki Akashi [f], Shigeki Aoki [f], Takamichi Murakami [a]

[a] *Department of Radiology, Kobe University Graduate School of Medicine, 7-5-2 Kusunoki-cho, Chuo-ku, Kobe, 650-0017, Japan*
[b] *Division of Medical Imaging, Biomedical Imaging Research Center, University of Fukui, 23-3 Matsuokashimoaizuki, Eiheiji, Yoshida, Fukui, 910-1193, Japan*
[c] *Department of Diagnostic Imaging and Nuclear Medicine, Kyoto University Graduate School of Medicine, 54 Shogoin Kawahara-cho, Sakyo-ku, Kyoto, 606-8507, Japan*
[d] *Advanced Imaging in Medical Magnetic Resonance, Kyoto University Graduate School of Medicine, 54 Shogoin Kawahara-cho, Sakyo-ku, Kyoto, 606-8507, Japan*
[e] *Department of Informatics, Kindai University, 3-4-1 Kowakae, Higashiosaka City, 577-8502, Japan*
[f] *Department of Radiology, Juntendo University Graduate School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo, 113-8421, Japan*

ARTICLE INFO

Keywords:
Transformer
Natural language processing
Radiology report
Summarization

ABSTRACT

Purpose: Natural language processing using language models has yielded promising results in various fields. Language models can help improve the workflow of radiologists. This retrospective study aimed to construct and evaluate language models for automatic summarization of radiology reports.
Methods: Two radiology report datasets from the MIMIC Chest X-ray (MIMIC-CXR) database and the Japan Medical Image Database (JMID) were included in this study. The MIMIC-CXR is an open database comprising chest radiograph reports. The JMID is a large database comprising computed tomography and magnetic resonance imaging reports from 10 academic medical centers in Japan. A total of 128,032 and 1,101,271 reports were included in this study from the MIMIC-CXR database and JMID, respectively. Four Text-to-Text Transfer Transformer (T5) models were constructed. Recall-Oriented Understudy for Gisting Evaluation (ROUGE), a quantitative metric, was used to evaluate the quality of the text summarized from 19,205 and 58,043 test sets from the MIMIC-CXR and JMID, respectively. The Wilcoxon signed-rank test was used to evaluate the differences among the ROUGE values of the four T5 models. Moreover, the subsets of automatically summarized text in the test sets were manually evaluated by two radiologists. The best T5 models were selected for automatic summarization using the Wilcoxon signed-rank test.
Results: The quantitative metrics of the best T5 models were as follows: ROUGE-1 = 57.75 ± 30.99, ROUGE-2 = 49.96 ± 35.36, and ROUGE-L = 54.07 ± 32.48 in the MIMIC-CXR; and ROUGE-1 = 50.00 ± 29.24, ROUGE-2 = 39.66 ± 30.21, and ROUGE-L = 47.87 ± 29.44 in the JMID. The radiologists' evaluations revealed 86% and 85% of the texts automatically summarized from the MIMIC-CXR and JMID, respectively, to be clinically useful.
Conclusion: The T5 models constructed in this study were able to perform automatic summarization of the radiology reports. The radiologists' evaluations demonstrated most of the automatically summarized texts to be clinically valuable.

1. Introduction

Radiology reports, which are a valuable source of information, play a crucial role in improving clinical practice and supporting research. A multitude of radiology reports have been written in recent years owing to the advances in the field of radiology. However, manually processing a large number of unstructured reports is challenging as radiology reports are often documented as unstructured data.

Natural language processing (NLP) by computers [1,2] facilitates the extraction of structured information from electronic medical records and radiology reports. Thus, NLP has been used for text classification, summarization, and generation in the field of radiology [3–5]. Recent advances in NLP include the application of deep learning.

* Corresponding author.
E-mail addresses: nishiomizuho@gmail.com, nmizuho@med.kobe-u.ac.jp (M. Nishio).

https://doi.org/10.1016/j.imu.2024.101465
Received 27 December 2023; Received in revised form 21 February 2024; Accepted 1 March 2024
Available online 4 March 2024

Abbreviations

CXR	Chest x-ray
JMID	Japan Medical Image Database
NLP	Natural language processing
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
T5	Text-to-Text Transfer Transformer
BERT	Bidirectional Encoder Representations from Transformers
CT	computed tomography
MRI	magnetic resonance imaging

NLP can potentially reduce the workload of radiologists by extracting structured information from radiology reports, thus aiding clinicians and radiologists in the decision-making process and identifying patients for research. However, unlike the computer vision [6], NLP has not received significant attention in the field of radiology, and reviews concerning the application of NLP are limited [5].

The development of language models is considered a promising advance in NLP; language models are neural networks trained using a large amount of text data, and the number of parameters in the model can be deemed a measure of performance. Several language models, such as Bidirectional Encoder Representations from Transformers (BERT) [7], Text-to-Text Transfer Transformer (T5) [8,9], and Generative Pre-Training-1 [10], Generative Pre-Training-2 [11], Generative Pre-Training-3 [12], and so on [13], have demonstrated state-of-the-art performance in NLP tasks.

Radiology reports are divided into two sections: findings and impression. Automatic summarization of the impression section according to the findings section would reduce the workload of radiologists. Thus, this study aimed to investigate the effectiveness of a language model to summarize radiology reports automatically. The contributions of this study are as follows: (i) The T5 language model was used to summarize radiology reports automatically. (ii) Automatic summarization of radiology reports was performed in two languages: chest radiograph (CXR) reports in English and computed tomography (CT) and magnetic resonance imaging (MRI) reports in Japanese. (iii) Recall-Oriented Understudy for Gisting Evaluation (ROUGE), a quantitative evaluation metric, and a semi-quantitative evaluation performed by radiologists were used to evaluate the automatically summarized sentences, and the relationship between the ROUGE metrics and radiologists' evaluations was investigated. (iv) A dataset with more than one million Japanese radiology reports was used for constructing and evaluating the T5 language model.

2. Material and method

This retrospective study was approved by the Institutional Review Boards of the Japan Medical Image Database (JMID) project and Kobe University Hospital; the requirement for informed consent was waived. This study was conducted in accordance with the Checklist for Artificial Intelligence in Medical Imaging [14].

2.1. Dataset

Two datasets were used from the MIMIC-CXR database and JMID in the study. The MIMIC-CXR database comprises chest radiographs and corresponding reports [15]. The MIMIC-CXR database is a large dataset of 377,110 CXR images for 65,379 patients who visited the Beth Israel Deaconess Medical Center between 2011 and 2016. Most of the CXR images included their radiology reports. The JMID dataset was obtained from the JMID project, in which 10 academic medical centers in Japan (Juntendo University, Kyushu University, Keio University, The University of Tokyo, Okayama University, Kyoto University, Osaka University, Hokkaido University, Ehime University, and Tokushima University) collaborated to create a large radiology database with de-identified patient data. Each center of the JMID project collected images and CT and MRI radiology reports from clinical radiology examinations and sent them to a central server of the JMID project. All the data of the clinical radiology examinations were collected in each center. The JMID project data were used for centralized management of medical resources of the radiology department, development of computer-aided diagnosis systems using artificial intelligence, centralized management of radiology reports, development for appropriate usage in radiology examinations, Japan Quantitative Imaging Biomarkers Alliance, and radiation dose management.

The modality of the MIMIC-CXR dataset was radiography, with the chest being the specified location; the reports were written in English. The modalities in the JMID dataset were CT and MRI, with locations spanning different parts of the body, and the reports were written in Japanese. All reports dated from 8/4/2010 to 3/31/2023 were collected from the JMID database. Fig. 1 presents the flowchart of data inclusion in this study. Reports with missing data in the findings or impression

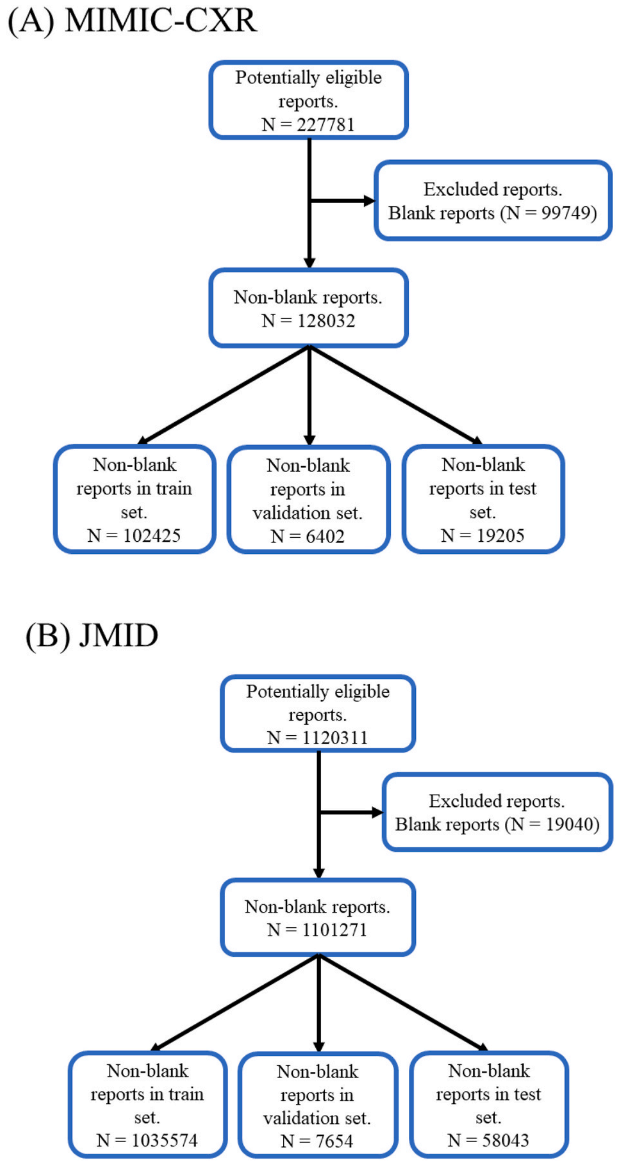

(A) MIMIC-CXR
Potentially eligible reports. N = 227781
Excluded reports. Blank reports (N = 99749)
Non-blank reports. N = 128032
Non-blank reports in train set. N = 102425
Non-blank reports in validation set. N = 6402
Non-blank reports in test set. N = 19205
(B) JMID
Potentially eligible reports. N = 1120311
Excluded reports. Blank reports (N = 19040)
Non-blank reports. N = 1101271
Non-blank reports in train set. N = 1035574
Non-blank reports in validation set. N = 7654
Non-blank reports in test set. N = 58043

Fig. 1. Flowcharts of the included radiology reports. (A) MIMIC-CXR, (B) JMID. Abbreviations: CXR, Chest X-ray; JMID, Japan Medical Image Database.

sections were excluded. Pairs of findings and impression sections were collected from each report in the two datasets. Table 1 presents the characteristics of the two datasets used in this study. Among the 227,781 reports present in the MIMIC-CXR dataset, 128,032 reports had no missing data. The findings or impression sections were often missing in the reports of the MIMIC-CXR dataset. On the other hand, 1,101,271 had no missing data among the 1,120,311 reports in the JMID dataset.

2.2. Dataset partition

The MIMIC-CXR images were randomly divided into training, validation, and test sets in a 16:1:3 ratio. The reports acquired from the JMID were divided into three sets according to the date of the reports: training set, 8/4/2010–11/30/2022 and 12/10/2022–12/31/2022; validation set, 12/1/2022–12/9/2022; and test set, 1/1/2023–3/31/2023. Date-based dataset partitioning was not possible in the MIMIC-CXR dataset as the MIMIC-CXR reports were not dated. After dataset partitioning, the training, validation, and test sets of the MIMIC-CXR dataset comprised 102,425, 6,402, and 19,205 reports, respectively. In the JMID dataset, the training, validation, and test sets comprised 1,035,574, 7,654, and 58,043 reports, respectively.

2.3. Ground truth

The MIMIC-CXR database and JMID include clinical radiology reports; thus, the impression sections of actual reports were used as the ground truth.

2.4. Language model

T5, which is a transformer-based neural network model, comprises an encoder and decoder that uses a text-to-text approach [8,9]. Most NLP tasks, including translation, question answering, and classification, require delivering input sentences to the models and training them for generating target sentences, which facilitates the use of the same model, loss function, and hyperparameters of T5 for a variety of tasks. T5 has reported state-of-the-art results on many NLP benchmarks, while maintaining sufficient flexibility to be fine-tuned for a variety of important downstream tasks. T5 is considered particularly useful in text summarization as the downstream task. A large amount of unlabeled text data and an objective analogous to BERT's "masked language modeling" were used to pre-train the T5 model. Tokens of input text were randomly corrupted with special tokens during pre-training. The pre-trained T5 reconstructed the corrupted tokens of the input text after pre-training. Fine-tuning the pre-trained T5 model can improve the performance of text summarization, question answering, and text classification [16,17]. The present study focused on text summarization (summarization of radiology reports). Thus, text summarization of radiology reports was performed as a downstream task in this study. To our knowledge, few studies have used the fine-tuning of language models (T5, BERT, and so on) for automatic summarization of radiology reports. The input and output of the T5 model comprised the text of the findings and impressions sections, respectively. The T5 model was trained to automatically summarize the findings section via fine-tuning. Pre-trained T5 models were obtained from Hugging Face (https://huggingface.co/models) for fine-tuning. Two pre-trained T5 models ("t5-base" [18] and "google/mt5-base" [19]) were obtained for the MIMIC-CXR database. Two pre-trained T5 models ("megagonlabs/t5-base-japanese-web" [20] and "google/mt5-base" [19]) were obtained for the JMID. "t5-base," "megagonlabs/t5-base-japanese-web," and "google/mt5-base" are pre-trained English, Japanese, and multilingual models, respectively.

Table 1
Characteristics of the datasets.

Item	MIMIC-CXR	JMID
Number of reports	227781	1120311
Number of non-blank reports	128032	1101271
Number of reports in train set	102425	1035574
Number of reports in validation set	6402	7654
Number of reports in test set	19205	58043
Age (year)		
train set	Not available	62.33 ± 18.49
validation set	Not available	62.78 ± 18.16
test set	Not available	62.46 ± 18.82
Sex (male:female)		
train set	Not available	555975:479599
validation set	Not available	4002:3652
test set	Not available	30463:27580
Modality	X-ray	CT, MRI
Location	Chest	Various locations
Language	English	Japanese
Private/Public datasets	Public dataset	Private dataset

Note: Non-blank reports indicate that neither the findings nor impression sections are blank in the radiology reports. Information concerning age was unavailable for 16754, 726, and 6210 reports in the training, validation, and test sets, respectively.
Abbreviations: JMID, Japan Medical Image Database.

2.5. Model training

Fig. 2 summarizes the model development process and prediction using the T5 model. As shown, the following pre-trained T5 models were used to summarize the radiology reports in the MIMIC-CXR and JMID datasets.

- MIMIC-CXR: "t5-base" or "google/mt5-base"
- JMID: "megagonlabs/t5-base-japanese-web" or "google/mt5-base"

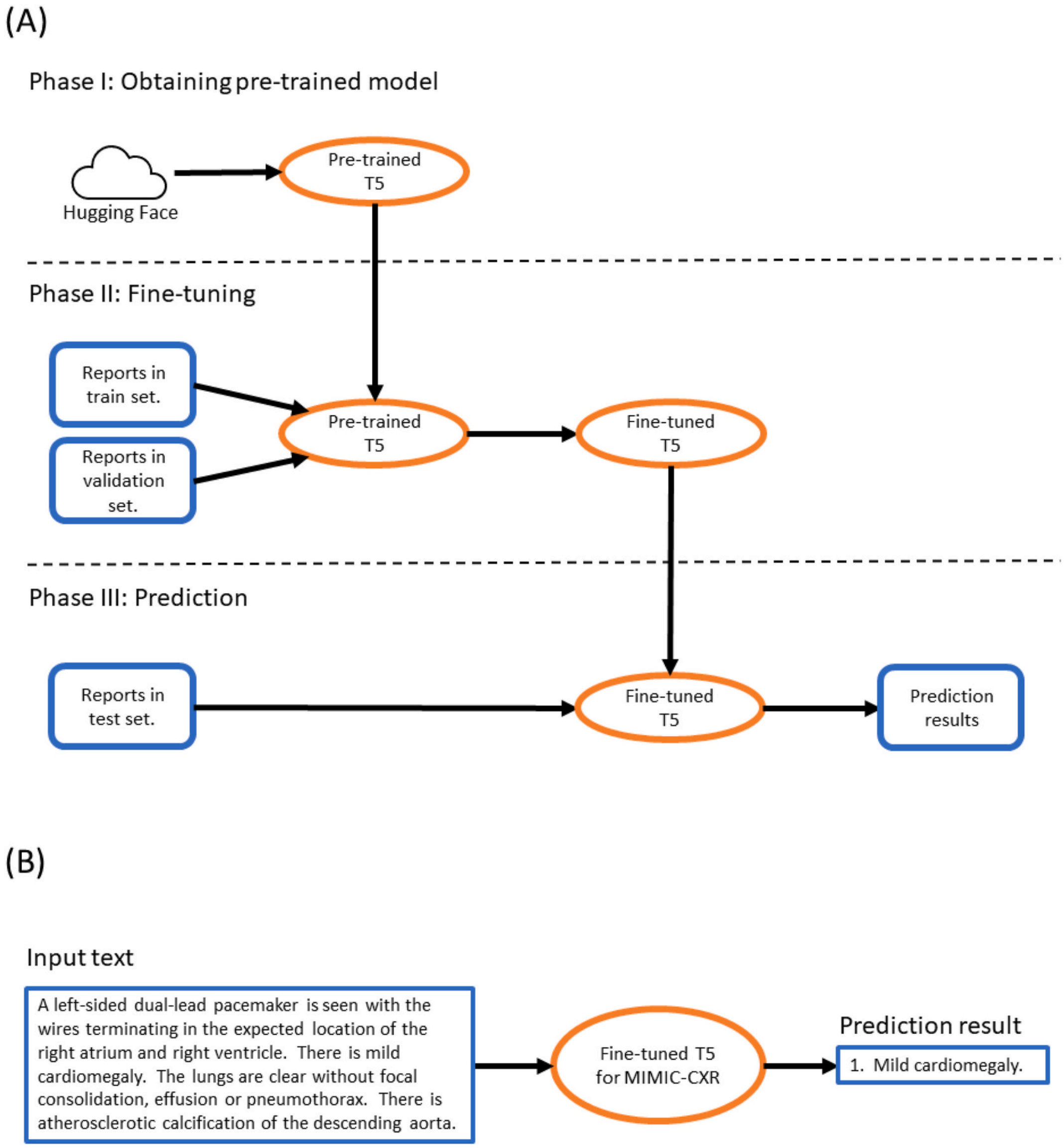
Character encoding conversion to UTF-8 was performed for the Japanese text of JMID as a preprocessing step before fine-tuning the T5 models. With the exception of character encoding conversion in JMID, no data preprocessing or normalization techniques were employed in the present study. Apart from the report selection process shown in Fig. 1, the feature selection process was not used. The following hyperparameters were used to fine-tune the pre-trained T5 models.

- Number of tokens in input: 1024
- Number of tokens in output: 128
- Number of training epochs: 5
- Batch size: 2 or 8
- Number of steps for gradient accumulation: 32
- Learning rate: 5e-5
- Learning scheduler: cosine annealing
- Warmup ratio in cosine annealing: 0.05

As shown, batch sizes of 2 and 8 were used for fine-tuning. Two types of pre-trained models and batch sizes for MIMIC-CXR and JMID were possible, resulting in four combinations. The following parameter was used to generate the predicted impression using the fine-tuned T5 models.

- Number of tokens in beam search: 6

Python (version, 3.8.8), pytorch (version, 1.8.0), and transformers (version, 4.22.2) were used for the development of the model and summary prediction. The source code for the model development was run_summarization.py of transformers (https://github.com/huggingface/transformers/blob/main/examples/pytorch/summarization/run_summarization.py). The T5 models were developed and evaluated on a workstation with NVIDIA(R) RTX(TM) A6000.



A left-sided dual-lead pacemaker is seen with the wires terminating in the expected location of the right atrium and right ventricle. There is mild cardiomegaly. The lungs are clear without focal consolidation, effusion or pneumothorax. There is atherosclerotic calcification of the descending aorta.

Fig. 2. Outline of model development and prediction. (A) Flowchart for obtaining a pre-trained T5 model, fine-tuning the T5 model from the pre-trained model, and predicting the text of the impression section with the fine-tuned model. (B) Examples of the summary text predicted from the findings section. Abbreviations: T5, Text-to-Text Transfer Transformer.

2.6. Evaluation

The summarized text obtained from the fine-tuned T5 models was evaluated quantitatively using ROUGE metrics [21,22] and semi-quantitatively by radiologists. First, the ROUGE metrics were calculated between the impression text of the actual report and the predicted text. ROUGE is a metric commonly used for evaluating text summarization. This metric measures the alignment between human-(reference summaries) and model-generated summaries. ROUGE has several variants; ROUGE-1, ROUGE-2, and ROUGE-L were used in this study. ROUGE-1 and ROUGE-2 are basic metrics that measure the alignment on an n-gram basis. ROUGE-1 and ROUGE-2 use unigrams and bigrams, respectively. The original definition of ROUGE-1 and ROUGE-2 is as follows:

$$ROUGE - n = \frac{\sum\limits_{S\in referencesgram_n\in S}\sum Count_{match}(gram_n)}{\sum\limits_{S\in referencesgram_n\in S}\sum Count(gram_n)}, \tag{1}$$

where n represents the length of the n-gram and $Count_{match}(gram_n)$ represents the maximum number of n-gram co-occurring in a model-generated summary and reference summary. ROUGE-1 and ROUGE-2 can be calculated according to the recall, precision, and F-measure. F-measure-based ROUGE-1 and ROUGE-2 were used in this study. ROUGE-1 and ROUGE-2 use the frequency of n-gram co-occurring; however, ROUGE-L uses common subsequence between human-generated and model-generated summaries. Considering two sequences X and Y (where X is a reference summary sentence, and Y is a model-generated summary sentence), the longest common subsequence (LCS) of X and Y is defined as a common subsequence with maximum length. To estimate the similarity between two summaries X of length m and Y of length n, ROUGE-L (F_{lcs}) is defined as LCS-based F-measure according to the following equations:

$$R_{lcs} = \frac{LCS(X, Y)}{m} \tag{2}$$

$$P_{lcs} = \frac{LCS(X, Y)}{n} \tag{3}$$

$$F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \quad (4)$$

Summary text should be tokenized for evaluating the alignment for calculating ROUGE metrics. Tokenizer of BERT was used to perform tokenization in this study.

Next, 100 reports were randomly selected from the test sets for semi-quantitative evaluation. Two radiologists with 17 and 7 years of experience in clinical radiology independently rated the predicted impression of the 100 reports on a 5-point scale as the semi-quantitative evaluation. The following text was semi-quantitatively evaluated: (i) the pairs of actual and predicted impression sections and (ii) finding sections of the actual reports. The 5-point scores were defined as follows: 1, the predicted impression could not be used clinically without rewriting; 2, most of the predicted impressions requires rewriting to be clinically useful; 3, approximately half of the predicted impression requires rewriting to be clinically useful; 4, the predicted impression is clinically useable with minor modifications; and 5, the predicted impression is clinically useable without modification. Summarized texts with scores of 4 and 5 were considered clinically useful. A consensus was reached through discussion in case of disagreements between the two radiologists.

2.7. Statistical analysis

The differences among the ROUGE metrics of the four T5 models were compared using the Wilcoxon signed-rank test. Quadratic-weighted kappa values were calculated for the scores of the two radiologists. The kappa values were interpreted using the following criteria: 0.00–0.20, none to slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1.00, almost perfect agreement. Spearman's correlation coefficients between the ROUGE-2 values and radiologists' consensus scores were determined for the 100 reports of the MIMIC-CXR and JMID, and the coefficients were statistically evaluated. Statistical significance was set at a p-value of 0.05. Statistical analyses were performed using R (version 4.2.2), Python (version 3.8.8), and Scipy (version 1.10.1).

3. Results

Table 2 presents the ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L) results of the fine-tuned T5 models in the MIMIC-CXR and JMID test sets. The results of the fine-tuned T5 models in the validation sets of the MIMIC-CXR and JMID datasets are presented in Supplemental Material 1. The fine-tuned "t5-base" model with a batch size of 2 demonstrated the highest ROUGE-1, ROUGE-2, and ROUGE-L values (ROUGE-1 = 57.75 ± 30.99, ROUGE-2 = 49.96 ± 35.36, and ROUGE-L = 54.07 ± 32.48) in the MIMIC-CXR dataset. In contrast, the fine-tuned "google/mt5-base" model with a batch size of 8 achieved the lowest values (ROUGE-1 = 52.35 ± 31.36, ROUGE-2 = 43.97 ± 35.48, and ROUGE-L = 49.05 ± 32.53). The fine-tuned "google/mt5-base" model with a batch size of 2 demonstrated the highest ROUGE-1, ROUGE-2, and ROUGE-L values (ROUGE-1 = 50.00 ± 29.24, ROUGE-2 = 39.66 ± 30.21, and ROUGE-L = 47.87 ± 29.44) in the JMID dataset. In contrast, the fine-tuned "megagonlabs/t5-base-japanese-web" model with a batch size of 8 achieved the lowest values (ROUGE-1 = 44.30 ± 28.56, ROUGE-2 = 34.01 ± 28.44, and ROUGE-L = 42.11 ± 28.54). These results underscore the interaction between the type of pre-trained model, batch size, and dataset characteristics.

Table 2
ROUGE values in the test sets of the MIMIC-CXR and JMID.

Dataset	T5 model	ROUGE-1	ROUGE-2	ROUGE-L
MIMIC-CXR	t5-base, B = 2	57.75 ± 30.99	49.96 ± 35.36	54.07 ± 32.48
MIMIC-CXR	t5-base, B = 8	55.27 ± 30.55	46.98 ± 34.79	51.34 ± 31.99
MIMIC-CXR	google/mt5-base, B = 2	55.84 ± 31.26	47.86 ± 35.61	52.54 ± 32.59
MIMIC-CXR	google/mt5-base, B = 8	52.35 ± 31.36	43.97 ± 35.48	49.05 ± 32.53
JMID	megagonlabs/t5-base-japanese-web, B = 2	48.99 ± 29.49	38.80 ± 30.20	46.91 ± 29.64
JMID	megagonlabs/t5-base-japanese-web, B = 8	44.30 ± 28.56	34.01 ± 28.44	42.11 ± 28.54
JMID	google/mt5-base, B = 2	50.00 ± 29.24	39.66 ± 30.21	47.87 ± 29.44
JMID	google/mt5-base, B = 8	46.18 ± 28.39	35.69 ± 28.65	43.91 ± 28.46

Note: Values represent the mean ± standard deviation. The numbers of reports are 19205 and 58043 for the test sets of the MIMIC-CXR and JMID, respectively. Abbreviations: T5, Text-to-Text Transfer Transformer; B, batch size; ROUGE, Recall-Oriented Understudy for Gisting Evaluation, JMID, Japan Medical Image Database.

The differences in the ROUGE-1, ROUGE-2, and ROUGE-L values were statistically evaluated between each pair of the four fine-tuned models. The p-values of ROUGE-1, ROUGE-2, and ROUGE-L were <0.001 in the pairs of the four models in the MIMIC-CXR dataset, except for the p-values of ROUGE-1 between "google/mt5-base" with a batch of 2 and "t5-base" with a batch of 8. The p-values of ROUGE-1 between "google/mt5-base" with a batch of 2 and "t5-base" with a batch of 8 was 0.12. As the number of test set was larger in JMID dataset than that in MIMIC-CXR dataset, the p-values of ROUGE-1, ROUGE-2, and ROUGE-L were <0.001 for each pair of the four models in the JMID. Thus, we focused on the optimal fine-tuned models (the fine-tuned "t5-base" model with a batch size of 2 for the MIMIC-CXR and the fine-tuned "google/mt5-base" model with a batch size of 2 for the JMID) based on the p-values.

Table 3 presents the results of the radiologists' semi-quantitative scores for the predicted summaries of the 100 reports generated by the optimal fine-tuned models of the MIMIC-CXR and JMID datasets. The kappa values of the semi-quantitative scores between the two radiologists were 0.785 (95% confidence interval = 0.669–0.900) and 0.736 (95% confidence interval = 0.590–0.883) for the 100 reports acquired from the MIMIC-CXR and JMID test sets, respectively, indicating substantial agreement between the two radiologists. The number of reports for the consensus scores was as follows: score 1 = 1, score 2 = 2, score 3 = 11, score 4 = 15, and score 5 = 71 in the MIMIC-CXR; and score 1 = 2, score 2 = 3, score 3 = 10, score 4 = 25, and score 5 = 60 in the JMID. These results indicate that 86% (86/100) and 85% (85/100) of the automatically summarized texts were clinically useful in the MIMIC-CXR and JMID datasets, respectively.

Fig. 3 presents the scatter plots illustrating the relationships between the ROUGE-2 values and semi-quantitative scores of the two radiologists. Significant positive correlations were observed between the ROUGE-2 values and semi-quantitative scores in both the datasets. The calculated correlation coefficients were 0.446 (95% confidence interval = 0.274–0.591) and 0.261 (95% confidence interval = 0.0681–0.435) for the MIMIC-CXR and JMID datasets, respectively. The corresponding p-values of the correlation coefficients were <0.001 for both the datasets, indicating that the positive correlations observed between the

Table 3
Results of semi-quantitative evaluation by radiologists in 100 radiology reports of test sets.

Item	MIMIC-CXR	JMID
Number of Score 1	1	2
Number of Score 2	2	3
Number of Score 3	11	10
Number of Score 4	15	25
Number of Score 5	71	60
Number of reports evaluated by radiologists	100	100

Note: Scores were determined by the consensus of the two radiologists. Abbreviations: JMID, Japan Medical Image Database.

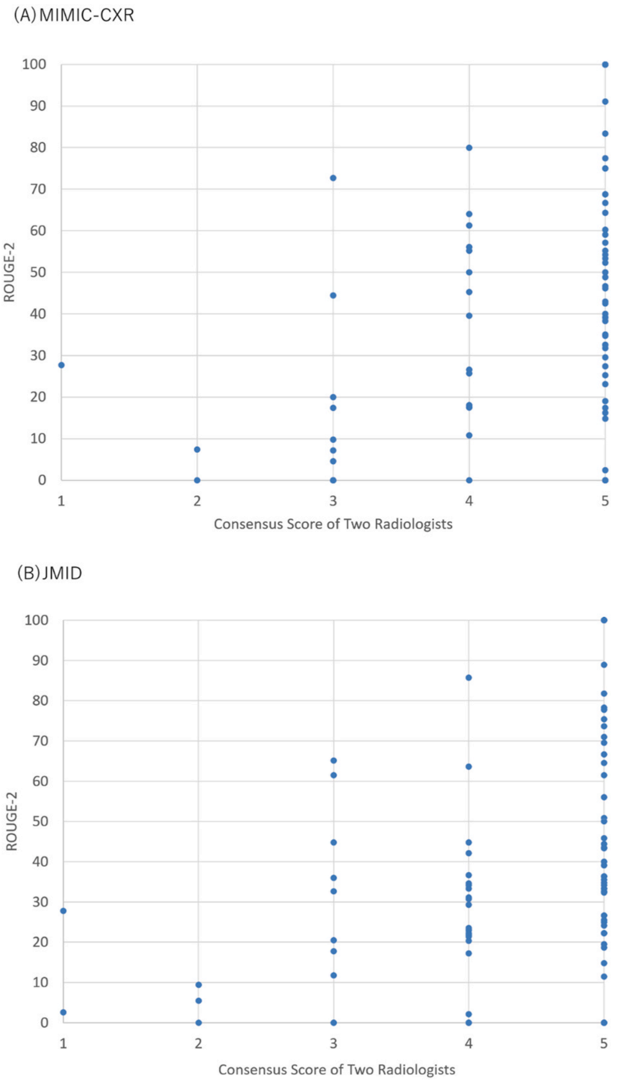


Fig. 3. Scatter plots between the ROUGE-2 values and semi-quantitative consensus scores by two radiologists. (A) Scatter plot for MIMIC-CXR, (B) Scatter plot for JMID. Note for (A): Correlation coefficient and p-value between ROUGE-2 values and consensus scores were 0.446 and < 0.001, respectively. Note for (B): Correlation coefficient and p-value between ROUGE-2 values and consensus scores were 0.261 and < 0.001, respectively. Abbreviations: CXR, Chest X-ray; JMID, Japan Medical Image Database; ROGUE, Recall-Oriented Understudy for Gisting Evaluation.

ROUGE-2 values and semi-quantitative scores were statistically significant. Scatter plots of the ROUGE-2 values and semi-quantitative scores for each of the two radiologists are presented in Supplemental Materials 2 and 3. Fig. 4 illustrates representative examples of radiology reports and the summary text predicted by the fine-tuned T5 models.

4. Discussion

An automatic summarization model of NLP was constructed using the T5 model. The fine-tuned T5 model was capable of summarizing radiology reports automatically. The present study revealed that the fine-tuned “t5-base” model with a batch size of 2 for the MIMIC-CXR and the fine-tuned “google/mt5-base” model with a batch size of 2 for the JMID were the best T5 models. The scores of the radiologists’ semi-quantitative evaluation of the 100 reports were ≥4 for 86% in the MIMIC-CXR test set and 85% in the JMID test set, thereby indicating the clinical usefulness of most of the predicted text in the impression section. These results demonstrate the usefulness of the T5 summarization models in the automatic summarization of radiology reports. Moreover, statistically significant correlations were observed between the radiologists’ semi-quantitative scores and quantitative evaluation using ROUGE metrics for the MIMIC-CXR and JMID datasets.

The MIMIC-CXR is a database comprising CXR reports and images [15]; however, the anatomical locations and diseases included in the MIMIC-CXR are relatively limited. In contrast, the JMID comprises CT and MR images of all locations, and the anatomical locations and diseases involved are broader than those in the MIMIC-CXR. Consequently, variation in the reports was greater in the JMID dataset, and report summarization was more challenging in the JMID than in the MIMIC-CXR. Table 3 shows that the radiologists’ scores were comparable for the two datasets. The JMID contained a greater number of reports, approximately ten times more than the MIMIC-CXR (Table 1). The results presented in Tables 1 and 3 suggest that the number of radiology reports influenced the performance of the automatic summarization models. Thus, the dataset size for constructing language models may be important in NLP, similar to computer vision [23,24].

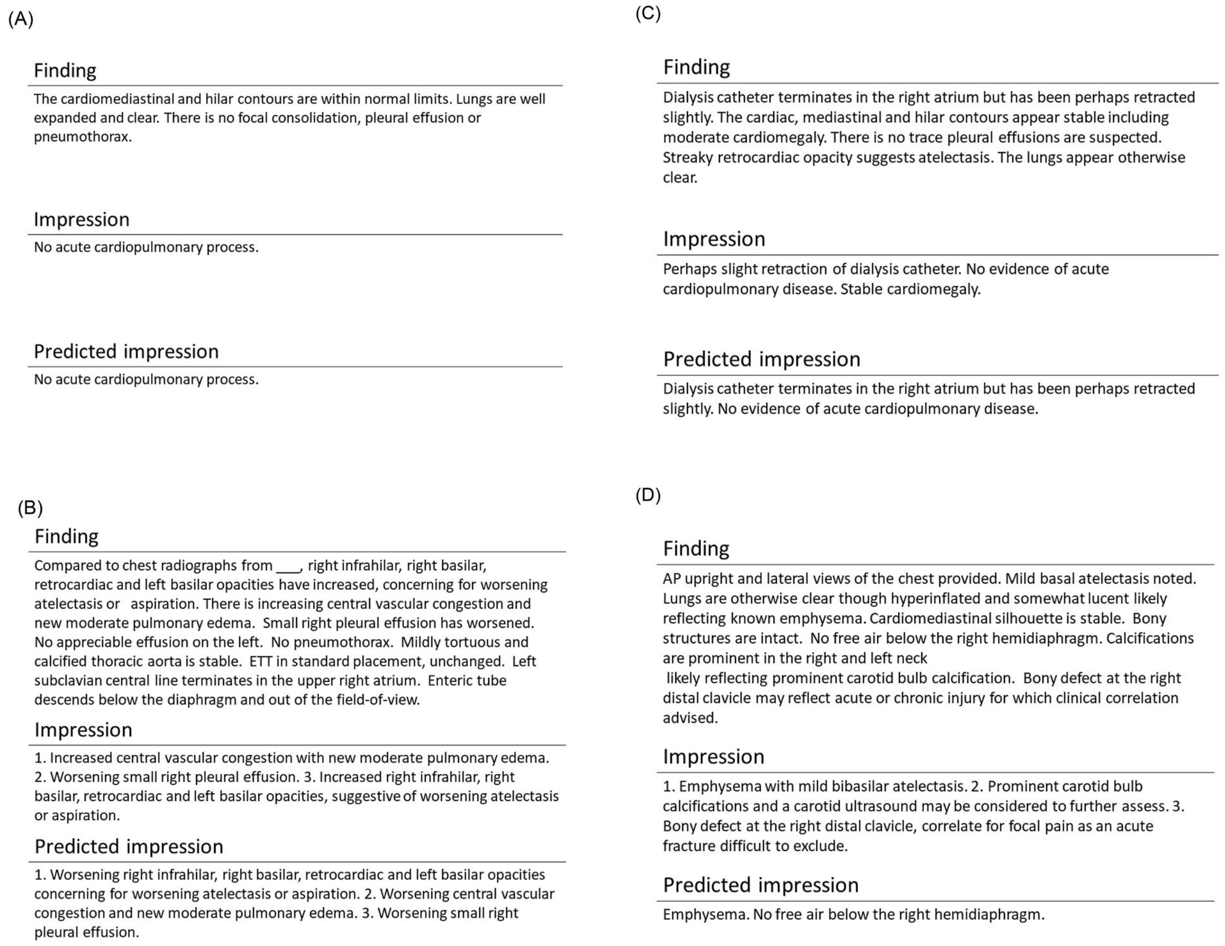
Two different pre-trained T5 models were used for the MIMIC-CXR and JMID in the present study: “t5-base” and “google/mt5-base” for MIMIC-CXR; and “megagonlabs/t5-base-japanese-web” and “google/mt5-base” for JMID. The best ROUGE values were obtained with “t5-base” for the MIMIC-CXR and “google/mt5” for the JMID (Table 2), thereby indicating that the English T5 model outperformed the multilingual T5 model in the English task, whereas the multilingual T5 model outperformed the Japanese T5 model in the Japanese task. In general, considering the variation in the dataset sizes based on the languages, the English dataset was more prominent owing to its larger size. The results of the present study and the size of English datasets suggest greater effectiveness of pre-trained English models for English tasks owing to sufficient generalizability. However, the size of the dataset used for pre-training was smaller for languages other than English. Nevertheless, the multilingual model would be more effective for non-English tasks because the multilingual model was pre-trained using both English and non-English datasets [8].

Table 4 shows the comparison between the present study and previous studies on automatic summarization of radiology reports. Radiology reports written in English were mainly used in the studies presented in Table 4 [25–30]. The present study included radiology reports of CXR in English and those of CT and MRI of various locations in Japanese, which is a major difference between previous studies and the present study. Moreover, the JMID dataset has the largest size among the datasets of studies presented in Table 4 and the JMID dataset was collected from multi centers. Thus, model development and evaluation of our T5 models were more robust compared to those of other studies, which is also a major strength of the present study.

Similar to a previous study [21], statistically significant correlations were observed between the ROUGE-2 values and the radiologists’ semi-quantitative scores for the 100 reports selected from the test sets. The Spearman’s correlation coefficients were 0.446 and 0.261 for the MIMIC-CXR and JMID, respectively, indicating weak correlations. The ROUGE-2 values exhibited significant variability even when the radiologists’ score was 5 (Fig. 3). Therefore, depending solely on the ROUGE metrics for evaluating individual summarized sentences may not be reliable. However, the average ROUGE values could potentially serve as surrogate for the average radiologist scores when the test set (evaluation dataset) is sufficiently large. These two evaluation methods should be used complementary, particularly when the evaluation dataset is limited.

JMID is a large dataset with a size of more than one million; creating a larger dataset than the JMID is challenging. Thus, significant improvements in the summarization model architecture or the pre-trained model are warranted for improving the performance of automatic summarization.

This study had certain limitations. First, we speculate that the usefulness of a multilingual or language-specific pre-trained model depends



(A)

Finding

The cardiomediastinal and hilar contours are within normal limits. Lungs are well expanded and clear. There is no focal consolidation, pleural effusion or pneumothorax.

Impression

No acute cardiopulmonary process.

Predicted impression

No acute cardiopulmonary process.

(B)

Finding

Compared to chest radiographs from ___, right infrahilar, right basilar, retrocardiac and left basilar opacities have increased, concerning for worsening atelectasis or aspiration. There is increasing central vascular congestion and new moderate pulmonary edema. Small right pleural effusion has worsened. No appreciable effusion on the left. No pneumothorax. Mildly tortuous and calcified thoracic aorta is stable. ETT in standard placement, unchanged. Left subclavian central line terminates in the upper right atrium. Enteric tube descends below the diaphragm and out of the field-of-view.

Impression

1. Increased central vascular congestion with new moderate pulmonary edema. 2. Worsening small right pleural effusion. 3. Increased right infrahilar, right basilar, retrocardiac and left basilar opacities, suggestive of worsening atelectasis or aspiration.

Predicted impression

1. Worsening right infrahilar, right basilar, retrocardiac and left basilar opacities concerning for worsening atelectasis or aspiration. 2. Worsening central vascular congestion and new moderate pulmonary edema. 3. Worsening small right pleural effusion.

(C)

Finding

Dialysis catheter terminates in the right atrium but has been perhaps retracted slightly. The cardiac, mediastinal and hilar contours appear stable including moderate cardiomegaly. There is no trace pleural effusions are suspected. Streaky retrocardiac opacity suggests atelectasis. The lungs appear otherwise clear.

Impression

Perhaps slight retraction of dialysis catheter. No evidence of acute cardiopulmonary disease. Stable cardiomegaly.

Predicted impression

Dialysis catheter terminates in the right atrium but has been perhaps retracted slightly. No evidence of acute cardiopulmonary disease.

(D)

Finding

AP upright and lateral views of the chest provided. Mild basal atelectasis noted. Lungs are otherwise clear though hyperinflated and somewhat lucent likely reflecting known emphysema. Cardiomediastinal silhouette is stable. Bony structures are intact. No free air below the right hemidiaphragm. Calcifications are prominent in the right and left neck likely reflecting prominent carotid bulb calcification. Bony defect at the right distal clavicle may reflect acute or chronic injury for which clinical correlation advised.

Impression

1. Emphysema with mild bibasilar atelectasis. 2. Prominent carotid bulb calcifications and a carotid ultrasound may be considered to further assess. 3. Bony defect at the right distal clavicle, correlate for focal pain as an acute fracture difficult to exclude.

Predicted impression

Emphysema. No free air below the right hemidiaphragm.

Fig. 4. Representative example of radiology reports and predicted summary text from MIMIC-CXR. Note: The radiologists' scores are as follows: A = 5, B = 5, C = 4, and D = 2.

Table 4
Comparison of the model for radiology report summarization.

Paper	Dataset	Dataset Size	Source of dataset	Language	Anatomical Location	Model	Metric
[25]	■ Stanford Hospital dataset	■ 87,127	■ Single center	■ English	■ Various Locations	■ Neural Sequence-to-Sequence Model (Bi-LSTM and LSTM) + Pointer-Generator Network	■ ROUGE-L = 47.06
	■ Indiana University Chest X-ray Dataset for cross-organization	■ 2691	■ Public dataset	■ English	■ Chest■		■ ROUGE-L = 34.56 (cross-organization)
[26]	■ MIMIC-CXR	■ 124,577	■ Public dataset	■ English	■ Chest	■ A pre-trained encoder and a randomly initialized Transformer-based decoder + Graph Enhanced Encoder + Contrastive Learning	■ ROUGE-L = 47.12 and 56.13
	■ OPENI	■ 3268	■ Public dataset	■ English	■ Chest		■ ROUGE-L = 64.45
[27]	■ MedStar Georgetown University	■ 41,066	■ Single center	■ English	■ Various Locations	■ RadLex PG	■ ROUGE-L = 37.02
[28]	■ MIMIC-CXR	■ 124,577	■ Public dataset	■ English	■ Chest	■ WGSUM	■ ROUGE-L = 55.32
	■ OPENI	■ 3268	■ Public dataset	■ English	■ Chest		■ ROUGE-L = 63.97
[29]	■ Stanford dataset	■ 130,850	■ Single center	■ English	■ Chest	■ Pointer-generator model + Reinforcement-learning with the ROUGE reward	■ ROUGE-L = 49.5
	■ RIH datasets	■ 139,654	■ Single center	■ English	■ Chest		■ ROUGE-L = 55.7
[30]	■ MIMIC-CXR	■ 123,620	■ Public dataset	■ English	■ Chest	■ ChatGPT + Dynamic Prompt + Iterative Optimization	■ ROUGE-L = 47.93
	■ OPENI	■ 2976	■ Public dataset	■ English	■ Chest		■ ROUGE-L = 65.47
Ours	■ MIMIC-CXR	■ 128,032	■ Public dataset	■ English	■ Chest	■ T5	■ ROUGE-L = 52.54
	■ JMID	■ 1,101,271	■ Multi centers	■ Japanese	■ Various Locations	■ mT5	■ ROUGE-L = 47.87

on the language of the radiology reports. However, this may depend on the dataset characteristics other than the language. As only two datasets were utilized, the usefulness of the pre-trained models could not be adequately evaluated in this study. Second, only two languages were used in this study. Thus, other languages should be investigated in future studies. The summarization performance of mT5 could be improved by studying large datasets with multilingual radiology reports. Third, the sizes of the MIMIC-CXR and JMID datasets were relatively large for medical NLP. Smaller datasets were not used in this study.

In conclusion, this study demonstrated the feasibility of automatic report summarization. According to the semi-quantitative evaluations performed by radiologists, majority of the automatically summarized text were clinically useful. Significant correlations were observed between the semi-quantitative and quantitative evaluations performed by the radiologists and ROUGE metrics, respectively. The use of quantitative assessments provided by the ROUGE metrics in conjunction with the semi-quantitative scores by radiologists could be complementary. These results were confirmed in the large MIMIC-CXR and JMID datasets written in English and Japanese, which could contribute to further advances in NLP radiology research. Based on the results of the present study, we aim to explore the usefulness of our models in clinical radiology workflow in future studies.

Ethical statement

This retrospective study was approved by the Institutional Review Boards of the Japan Medical Image Database (JMID) project and Kobe University Hospital; the requirement for informed consent was waived.

CRediT authorship contribution statement

Mizuho Nishio: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Takaaki Matsunaga:** Writing – review & editing, Investigation, Data curation. **Hidetoshi Matsuo:** Writing – review & editing, Validation, Funding acquisition, Data curation. **Munenobu Nogami:** Writing – review & editing, Resources, Project administration. **Yasuhisa Kurata:** Writing – review & editing, Validation. **Koji Fujimoto:** Writing – review & editing, Project administration, Funding acquisition. **Osamu Sugiyama:** Writing – review & editing, Resources. **Toshiaki Akashi:** Writing – review & editing, Data curation. **Shigeki Aoki:** Writing – review & editing, Data curation. **Takamichi Murakami:** Writing – review & editing, Supervision.

Declaration of competing interest

None Declared

Acknowledgements

We thank JMID project and the following 10 academic medical centers.

Juntendo University.
Kyushu University.
Keio University.
The University of Tokyo.
Okayama University.
Kyoto University.
Osaka University.
Hokkaido University.
Ehime University.
Tokushima University.

This work was supported by JSPS KAKENHI (Grant Numbers: 22K07665, 23K07154, 23K17229, and 23KK0148). There was no role of the funding source in the study design, the collection, and analysis and interpretation of data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2024.101465.

References

[1] Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. Multimed Tool Appl 2023;82:3713–44. https://doi.org/10.1007/s11042-022-13428-4.

[2] Chowdhary KR. Natural Language processing. Fundamentals of artificial intelligence. New Delhi: Springer India; 2020. p. 603–49. https://doi.org/10.1007/978-81-322-3972-7_19.

[3] Pons E, Braun LM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016;279:329–43. https://doi.org/10.1148/radiol.16142770.

[4] Linna N, Kahn Jr CE. Applications of natural language processing in radiology: a systematic review. Int J Med Inf 2022;163:104779. https://doi.org/10.1016/j.ijmedinf.2022.104779.

[5] Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, Grover C, Suárez-Paniagua V, Tobin R, Whiteley W, Wu H, Alex B. A systematic review of natural language processing applied to radiology reports. BMC Med Inf Decis Making 2021;21:179. https://doi.org/10.1186/S12911-021-01533-7.

[6] Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018;9:611–29. https://doi.org/10.1007/s13244-018-0639-9.

[7] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 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). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–86. https://doi.org/10.18653/v1/N19-1423.

[8] Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C. 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. Online: Association for Computational Linguistics; 2021. p. 483–98. https://doi.org/10.18653/v1/2021.naacl-main.41.

[9] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020;21:5485–551.

[10] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf; 2018.

[11] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf; 2019.

[12] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language Models are Few-Shot Learners. ArXiv 2020;abs/2005.14165. https://arxiv.org/abs/2005.14165..

[13] He P, Liu X, Gao J, Chen W. DeBERTa: decoding-enhanced BERT with disentangled attention. In: Proceedings in ICLR2021; 2021.

[14] Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2:e200029. https://doi.org/10.1148/RYAI.2020200029.

[15] Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, Mark RG, Horng S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6:317. https://doi.org/10.1038/s41597-019-0322-0.

[16] Mastropaolo A, Scalabrino S, Cooper N, Palacio DN, Poshyvanyk D, Oliveto R, Bavota G. Studying the usage of text-to-text transfer transformer to support code-related tasks. In: 2021 IEEE/ACM 43rd international conference on software engineering (ICSE); 2021. p. 336–47. https://doi.org/10.1109/ICSE43902.2021.00041.

[17] Abacha AB, Yim WW, Adams G, Snider N, Yetisgen-Yildiz M. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In: Proceedings of the 5th clinical Natural language processing workshop. Toronto, Canada: Association for Computational Linguistics; 2023. p. 503–13. https://doi.org/10.18653/v1/2023.clinicalnlp-1.52.

[18] t5-base · Hugging Face. https://huggingface.co/t5-base. Accessed September 22, 2023..

[19] google/mt5-base · Hugging Face. https://huggingface.co/google/mt5-base. Accessed September 22, 2023..

[20] megagonlabs/t5-base-japanese-web · Hugging Face. https://huggingface.co/megagonlabs/t5-base-japanese-web. Accessed September 22, 2023..

[21] Lin CY. ROUGE: a package for automatic evaluation of summaries. Text summarization branches out. Barcelona, Spain: Association for Computational Linguistics; 2004. p. 74–81. https://aclanthology.org/W04-1013.

[22] Lin CY, Hovy E. 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; 2003. 150–7, https://aclanthology.org/N03-1020.

[23] Luo C, Li X, Wang L, He J, Li D, Zhou J. How does the data set affect CNN-based image classification performance? 2018 5th international conference on systems and informatics. ICSAI); 2018. p. 361–6. https://doi.org/10.1109/ICSAI.2018.8599448.

[24] Dawson HL, Dubrule O, John CM. Impact of dataset size and convolutional neural network architecture on transfer learning for carbonate rock classification. Comput Geosci 2023;171:105284. https://doi.org/10.1016/j.cageo.2022.105284.

[25] Zhang Y, Ding DY, Qian T, Manning CD, Langlotz CP. Learning to summarize radiology findings. In: Lavelli A, Minard AL, Rinaldi F, editors. Proceedings of the ninth international workshop on health text mining and information analysis. Association for Computational Linguistics; 2018. p. 204–13. https://doi.org/10.18653/v1/W18-5623.

[26] Hu J, Li Z, Chen Z, Li Z, Wan X, Chang TH. Graph enhanced contrastive learning for radiology findings summarization. In: Muresan S, Nakov P, Villavicencio A, editors. Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics; 2022. p. 4677–88. https://doi.org/10.18653/v1/2022.acl-long.320.

[27] MacAvaney S, Sotudeh S, Cohan A, Goharian N, Talati I, Filice RW. Ontology-aware clinical abstractive summarization. Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. SIGIR'19, vols. 1013–6. Association for Computing Machinery; 2019. https://doi.org/10.1145/3331184.3331319.

[28] Hu J, Li J, Chen Z, et al. Word graph guided summarization for radiology findings. In: Zong C, Xia F, Li W, Navigli R, editors. Findings of the association for computational linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics; 2021. p. 4980–90. https://doi.org/10.18653/v1/2021.findings-acl.441.

[29] Zhang Y, Merck D, Tsai E, Manning CD, Langlotz C. Optimizing the factual correctness of a summary: a study of summarizing radiology reports. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics; 2020. p. 5108–20. https://doi.org/10.18653/v1/2020.acl-main.458.

[30] Ma C, Wu Z, Wang J, et al. ImpressionGPT: an iterative optimizing framework for radiology report summarization with ChatGPT. Published online. 2023. arXiv: 2304.08448, https://arxiv.org/abs/2304.08448.