Finding-Centric Structuring of Japanese Radiology Reports and Analysis of Performance Gaps for Multiple Facilities

Yuki Tagawa[1] **Yohei Momoki**[1] **Norihisa Nakano**[1] **Ryota Ozaki**[1]
Motoki Taniguchi[1] **Masatoshi Hori**[2] **Noriyuki Tomiyama**[2]
[1]FUJIFILM Corporation [2]Osaka University Graduate School of Medicine
yuki.tagawa@fujifilm.com

Abstract

This study addresses two key challenges in structuring radiology reports: the lack of a practical structuring schema and datasets to evaluate model generalizability. To address these challenges, we propose a "Finding-Centric Structuring," which organizes reports around individual findings, facilitating secondary use. We also construct JRadFCS, a large-scale dataset with annotated named entities (NEs) and relations, comprising 8,428 Japanese Computed Tomography (CT) reports from seven facilities, providing a comprehensive resource for evaluating model generalizability. Our experiments reveal performance gaps when applying models trained on single-facility reports to those from other facilities. We further analyze factors contributing to these gaps and demonstrate that augmenting the training set based on these performance-correlated factors can efficiently enhance model generalizability.

1 Introduction

A radiology report documents abnormal findings and suspected diseases observed in medical images. Radiology reports contain expert insights; however, they are often recorded in free-text format, limiting their secondary application. Structuring these reports through information extraction (IE) can support a wide range of applications, such as report generation (Delbrouck et al., 2022; Zhang et al., 2020) and multimedia reports (Folio et al., 2018).

Despite advancements in IE from radiology reports (Yada et al., 2020; Cheng et al., 2022; Delbrouck et al., 2024), two critical challenges hinder the practical application of structured reports: the lack of a well-designed structuring schema for practical use and datasets suitable for evaluating the generalizability of structuring models.

We propose **Finding-Centric Structuring (FCS)**, which organizes reports around individual findings to address the first challenge. Figure 1 shows an overview of FCS. Our approach structures reports into individual findings along with related attributes such as characteristics and diagnoses. Structured data created by FCS can be useful for a variety of applications. For example, FCS can be applied to Medical Visual Grounding (Zhang et al., 2022), which aligns sentences in reports with corresponding objects in images. By decomposing these reports into finding-centric data, fine-grained Medical Visual Grounding for individual findings is promoted. Furthermore, FCS allows radiologists to efficiently track changes in the size of each finding and monitor the effectiveness of treatments. FCS enables us to go beyond existing secondary uses such as report retrieval, supporting applications focused on individual findings.

The second challenge involves assessing the generalizability of structuring models. Nakamura et al. (2022) reports that radiologists use diverse terminologies. For example, they may describe subsolid nodules using synonyms such as "GGN." This variability raises concerns about the ability of the model to accurately structure reports with varied writing styles and across facilities. Most existing studies on structuring reports (Sugimoto et al., 2023; Lau et al., 2023; Park et al., 2024) use reports from a single facility or focus on specific diseases to validate their models, limiting the evaluation of model generalizability.

We construct JRadFCS, a large-scale dataset annotated with NEs and relations based on our schema, comprising 8,428 Japanese CT reports from seven facilities, to address second challenge. JRadFCS includes a wide variety of reports covering different organs and diseases by collecting all reports written during a specific period. This diversity makes JRadFCS suited for evaluating the generalizability of models across various reports.

In developing a model for practical use, it is difficult to use data from multiple facilities as a training set due to contractual and cost constraints. There-

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 70–85

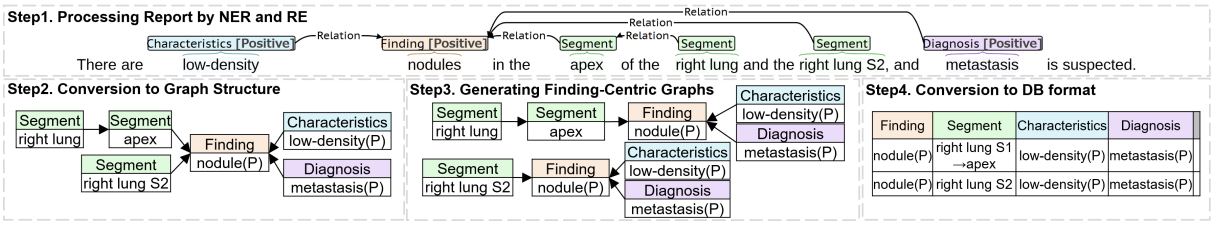

Figure 1: An overview of our proposed FCS. In step 1, our approach structures the report through Named Entity Recognition (NER) and Relation Extraction (RE). In step 3, our approach transforms the output graphs of NER and RE into Finding-Centric Graphs to structure reports into each finding. Structuring in this manner allows us to build a Finding-Centric Structured Database. This DB can serve as a foundation for various applications.

fore, as a more practical setting, we evaluate the performance of a model trained on single-facility reports when applied to reports from other facilities. We evaluate various BERT (Devlin et al., 2019) models, including our BERT for the radiology domain and a Large Language Model (LLM), revealing large performance gaps between facilities. Additionally, to identify the factors contributing to these performance gaps, we analyze the relationship between metrics indicating the complexity of reports, such as the length of the report, and model performance. Furthermore, we demonstrate that training set augmentation based on the identified complexity metrics can efficiently improve performance on reports from other facilities.

2 Related Work

Various annotation schemes for radiology reports have been proposed. Yada et al. (2020) propose a schema for NEs, which has been applied in various studies (Yada et al., 2022; Cheng et al., 2022; Nakamura et al., 2022). This scheme treats multiple findings such as "結節と網状影 (nodule and reticular shadows)" as a single NE. Sugimoto et al. (2023) and Lau et al. (2023) annotate multiple segments such as "左第7、8肋骨 (left 7th, 8th ribs)" as a single NE. These schemas define coarse-grained NEs, which hinder FCS and limit applications requiring precise statistics.

RadGraph (Jain et al., 2021) and its extension, RadGraph-XL (Delbrouck et al., 2024), focus on structuring chest X-ray and CT/MR reports, respectively. Unlike X-rays, CT scans provide 3D imaging, which enables radiologists to observe detailed characteristics such as the shape and condition of findings. However, RadGraph-XL lacks specific labels for characteristics and temporal changes, instead labels them as findings ("observations" in their schema). Our approach extracts relevant attributes, such as characteristics, as distinct labels from findings, ensuring FCS and a finer granularity suited for the complexity of CT scans.

Other efforts include report labeler (Irvin et al., 2019; Johnson et al., 2019), NE and/or RE schemas (Patel et al., 2018; Bustos et al., 2020; Datta et al., 2020; Park et al., 2024) have been proposed. Contrary to prior studies, we uniquely focus on FCS.

NE Label	Definition
Finding (*F*)	Abnormalities or abnormal conditions.
Diagnosis (*D*)	Diseases inferred from the findings.
Characteristics (*C*)	Features of findings, such as state, nature, or degree of brightness.
Temporal change (*T*)	Changes compared to past tests.
Segment (*S*)	Regions based on anatomical definitions, organs or parts of organs.
Measurement result (*R*)	Measured values or qualitative size expressions.
Measurement item (*I*)	Items for measured values.
Quantity (*Q*)	The number of findings.

Table 1: NE labels and their definitions. The symbols in parentheses are abbreviations.

3 Finding-Centric Structuring

Following discussions with three board-certified radiologists, we developed a set of entities and relations to capture critical information.

3.1 NEs and Relations

Table 1 shows the NE labels and their respective definitions. For the labels *F*, *D*, *C*, and *T*, we assign a factuality attribute: Positive if the concept is observed, and Negative if it is not.

We define relations from NE labels *D*, *C*, *T*, *S*, *R*, and *Q* to label *F* to capture the relevant attributes of each finding. Furthermore, we define hierarchical anatomical relations from higher anatomical label *S* to lower anatomical label *S*, and relations from label *I* to *R* to associate measured items with their values (e.g., "diameter → 3cm"). In Figure 1, the

label F assigned to "nodules" is connected to "low-density" and "metastasis," capturing attributes of "nodules." The relations "right lung → apex → nodule" represents the detailed position of "nodules" along with the hierarchical anatomical relations.

3.2 Generating Finding-Centric Graphs

Radiology reports often describe multiple findings within a single sentence, necessitating additional processing to separate each finding. For example, the report in Figure 1 states that nodules are in two distinct segments. Relying on NER and RE is insufficient to accurately determine the number of findings described in the report. Therefore, we introduce rule-based processing that transform the output of NER and RE into finding-centric graphs (step 3 in Figure 1). The following is an example of the rules. Details are provided in Appendix A.

- **Segment-Path Rule**

  For the graphs containing multiple *Segments*, finding-centric graphs are generated based on the paths from each terminal segment to the findings. For example, in Figure 1, two paths are identified:, "right lung → apex → nodule" and "right lung S2 → nodule"; thus, two finding-centric graphs are generated by adding each segment path.

3.3 Evaluating Finding-Centric Structuring

We introduce the Finding-centric Graph Score (FGS) to evaluate FCS. A predicted graph is considered correct if it exactly matches the gold graph. This implies that all NEs must have the correct labels, factuality, and spans, and that all relations must correctly connect the NEs. The FGS F1 Score F_{FGS} is the harmonic mean of FGS Precision P_{FGS} and FGS Recall R_{FGS}. P_{FGS} is the ratio of correctly predicted finding graphs N_{tp} to the total predicted graphs N_{pred}: $P_{\text{FGS}} = \frac{N_{\text{tp}}}{N_{\text{pred}}}$, and R_{FGS} is the ratio of N_{tp} to the total gold graphs N_{gold}: $R_{\text{FGS}} = \frac{N_{\text{tp}}}{N_{\text{gold}}}$.

FGS evaluates the comprehensiveness of relevant attributes for individual findings and the correctness of the number of generated finding-centric graphs. This is critical for practical applications that rely on the integrity of structured data.

RadGraphF1 (Yu et al., 2023) is an evaluation metric based on RadGraph for report generation models. RadGraphF1 calculates the F1 score based on the matching of NEs (nodes) and their relations (edges) in the RadGraph outputs, which interprets it a metric for the local correctness of the generated report. In contrast, FGS measures the exact matching of graphs, allowing for a comprehensive evaluation of findings and their relevant attributes. Especially for CT scans, which provide 3D imaging, many kinds of findings and their attributes can be described in the report. Thus, it is also important to evaluate generative or structuring models in terms of the comprehensiveness of attributes and the correctness of the number of findings. Overall, FGS offers a more holistic evaluation compared to RadGraphF1.

Facility	#Training	#Validation	#Test	Collection Period
OUH	1,344	200	1,536	Jun. 2-15, 2021 (14 days)
A	0	200	781	Jun. 1-7, 2021 (7 days)
B	0	200	583	Oct. 1-7, 2020 (7 days)
C	0	200	420	Jun. 1-7, 2021 (7 days)
D	0	200	1,141	Jun. 1-7, 2021 (7 days)
E	0	200	624	Dec. 1-7, 2020 (7 days)
F	0	200	599	Jun. 1-7, 2021 (7 days)

Table 2: The number of reports in the JRadFCS dataset. The facility name "OUH" refers to Osaka University Hospital, while the other A to F are placeholders for different hospitals. In the training set, we randomly sampled reports regardless of the period.

Research	Anatomy	#Facilities	#Reports
Hassanpour and Langlotz (2016)	Chest	3	150
Yada et al. (2020)	Lung	2	1,498
Cheng et al. (2022)	Lung	Not mentioned	1,000
Nakamura et al. (2022)	Lung	1 (Radiopaedia)	135
Sugimoto et al. (2023)	Chest, abdomen	1	1,040
Lau et al. (2023)	Chest	1	500
Park et al. (2024)	Whole body	1	203
Delbrouck et al. (2024)	Chest, abdomen/pelvis	2	1,200
Zhao et al. (2024)	Whole body	1 (MIMIC-IV)	1,816
JRadFCS (Ours)	Whole body	**7**	**8,428**

Table 3: Comparison of CT report datasets, manually annotated NEs and/or relations. **Anatomy** denotes the imaging part of the reports. **#Facilities** and **#Reports** denote the number of source facilities and reports.

4 JRadFCS

We constructed JRadFCS, a dataset of Japanese CT reports annotated by our schema. Two annotators, each with over 10 years of experience in annotation for medical NLP tasks, were employed to annotate NEs and their relations. Each report was annotated by a single annotator.

We collected all CT reports written during a specific period from each facility. Table 2 shows the statistics for the reports included in JRadFCS. This sampling approach allows us to simulate the performance of a structuring model when deployed over

a defined period, which is crucial for assessing its real-world applicability. Moreover, this approach ensures that JRadFCS includes reports covering a wide range of organs and diseases.

Table 3 compares JRadFCS with existing datasets. JRadFCS contains the largest number of CT reports and multi-facilities reports. The diversity in facility sources, coupled with the variety of organs and diseases represented, provides a key advantage for developing models that can be generalized across various clinical scenarios.

The training set consists only of the OUH reports to evaluate the performance for other-facility reports (Table 2). Note that the validation sets for facilities A to F are only used for later analyses and are not utilized for model training, nor even for checkpoint selection. Further details of JRadFCS are provided in Appendix B.

5 Experiments

In this section, we evaluate the performance of the structuring model trained on OUH reports when applied to those from other facilities. Specifically, we compare the performance of different BERT-based models, including UTH-BERT (Kawazoe et al., 2021), Tohoku-BERT (2024) and our BERT trained on radiology reports. Additionally, we analyze the performance gaps among the facilities and explore potential reasons for these gaps.

5.1 Experimental Settings

We utilized a pipeline for NER and RE based on BERT (Devlin et al., 2019). Fine-tuned BERT models have demonstrated strong results in various IE tasks (Cheng et al., 2022; Shibata et al., 2024).

For the NER model, we trained BERT-CRF (Souza et al., 2020) with labels that combine NE labels with factuality labels (e.g., Finding-Positive), allowing it to handle the NER and factuality prediction simultaneously.

For the RE model, we trained a binary classification model to predict the relations between NEs. We used BERT embeddings for the subject, object, and the span between them, computed through average pooling of the token embeddings. These embeddings were concatenated and fed into a softmax classifier to predict the probability of relation existence. We fine-tuned the model using cross-entropy loss. During inference, the model predicted relations for all subject and object pairs.

In domain-specific tasks, pre-trained language models (PLMs) trained on domain-specific texts typically outperform those trained on general-domain data (Gu et al., 2021; Ghosh et al., 2023). From this perspective, we constructed JRadBERT, a PLM with a character-level tokenizer, trained on approximately 758K Japanese radiology reports (over 10.6M sentences and 103.3M words) from OUH. Importantly, the pre-training dataset for JRadBERT does not overlap with the reports or patients included in JRadFCS. JRadBERT is a BERT-base model trained on Masked-LM, where 15% of the words in the text are masked. The vocabulary size is 3,930. Details on the training of NER, RE, and JRadBERT are presented in Appendix D.

We compared JRadBERT with UTH-BERT and Tohoku-BERT. UTH-BERT is a BERT-base model trained on approximately 120M lines of Japanese clinical text and uses J-Medic (Ito et al., 2018) to treats medical terms as one token. This model outperforms general-BERT in some clinical tasks (Nishigaki et al., 2023). Tohoku-BERT is a BERT-base model trained on 79.2GB of general-domain Japanese text, and achieves high performance in some NLP tasks (Tsukagoshi et al., 2023).

5.2 Experimental Results

Table 4a shows the F1 scores for NER, RE, and FGS using models fine-tuned on reports from OUH. Tohoku-BERT achieved the highest scores at several facilities; however, our JRadBERT demonstrated superior performance in both Macro and Micro-F1 scores, with lower SD, despite its smaller pre-training text of 0.32GB, which is approximately 1/250 of the size of that of Tohoku-BERT. These results suggest that domain-specific PLM enhances performance and robustness across facilities. The performance of NER and RE for each label is provided in Appendix E.

One reason for the lower performance of UTH-BERT is its use of J-Medic, which treats medical terms as one token. For instance, it tokenizes "腹水なし (no ascites)" as one token, whereas our schema requires it to be extracted as "腹水 (ascites)." This difference in token granularity leads to NER errors. Conversely, our JRadBERT uses a character-level tokenizer to mitigate these errors.

LLMs have been proven effective in various NLP tasks (Liu et al., 2023). Table 4b shows the F_{FGS} of JRadBERT and GPT-4o with 20-shots on the validation set. The F_{FGS} of GPT-4o at the best-performing facility was 57.36, significantly lower than JRadBERT. We observed that GPT-4o tends to

	UTH-BERT			Tohoku-BERT			JRadBERT		
	F_{NER}	F_{RE}	F_{FGS}	F_{NER}	F_{RE}	F_{FGS}	F_{NER}	F_{RE}	F_{FGS}
OUH	84.92	90.04	64.88	95.76	95.18	85.47	**96.01**	**95.30**	**85.84**
A	74.27	81.41	45.59	93.82	**94.33**	**83.89**	**94.01**	94.29	83.83
B	77.91	86.19	47.31	93.21	94.82	**81.25**	**93.28**	**94.92**	81.12
C	71.09	84.63	43.54	89.60	92.83	74.28	**91.90**	**94.15**	**80.08**
D	68.84	84.57	39.71	**91.42**	93.61	**78.91**	91.13	**94.13**	78.44
E	73.96	83.59	38.00	91.65	91.68	69.23	**92.23**	**94.53**	**77.20**
F	68.90	84.84	39.14	**90.89**	91.59	74.52	90.74	**93.32**	**76.33**
Micro w/o OUH	72.59	84.27	42.51	92.00	93.46	78.33	**92.28**†	**94.31**†	**79.92**†
Macro w/o OUH	72.49	84.21	42.21	91.76	93.14	77.01	**92.21**	**94.22**	**79.50**
SD w/o OUH	3.54	1.60	3.81	1.54	1.35	5.35	**1.25**	**0.53**	**2.76**

(a) F1 scores on the test set.

JRadBERT	GPT-4o (20-shots)
F_{FGS}	F_{FGS}
83.31	57.36
83.94	44.19
81.51	46.51
82.48	50.85
79.09	40.40
74.96	38.90
74.18	40.26
80.15	45.81
79.36	45.26
4.04	5.77

(b) F1 scores on the validation set.

Table 4: F1 scores of NER (F_{NER}), RE (F_{RE}) and FGS (F_{FGS}). SD represents the standard deviation. **Bold** indicates the best performance.† indicates a significant difference with the other models (McNemar's test, $p < 0.01$).

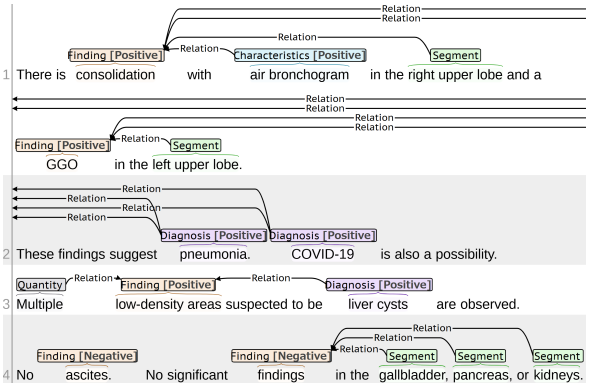

Figure 2: An example of an annotated report. Multiple graphs are generated from the line 1 and 2, each centered on the "consolidation" and "GGO." Therefore, these are counted as MG. Besides, since the factuality is Positive, these are also counted as PG and PMG. In the last sentence, three graphs are generated according to **Segment-Path Rule**, and these are counted as MG, however, not as PG and PMG because their factuality is Negative. In this report, there are seven graphs in total, resulting in RG being 7, PG being 3/7 $\approx$ 0.43, and PMG being 2/7 $\approx$ 0.29.

make errors in the spans of NE that not appeared in the few-shot samples. Details of comparison with GPT-4o are provided in Appendix F.

Our domain-specific model achieves the highest performance; however, performance gaps remain across facilities. Surprisingly, there is a significant gap of nearly 9.5 pt in FGS between OUH and facility F. These results indicate that evaluating models using reports from only a few facilities might not adequately reflect their generalizability.

5.3 Performance Degradation Factor Analysis

We defined metrics indicating the complexity of a report to examine factors contributing to performance degradation on reports from other facilities. If the F1 scores decrease as the complexity metric values increase, the correlation indicates a negative value. Therefore, metrics with high negative correlation can be considered as factors contributing to performance degradation.

Table 5 shows the defined metrics, their definitions, and Pearson's correlation coefficients on validation sets from facilities A to F. We defined the metrics from three perspectives: entity-level, report-level, and graph-level. An example of an annotated report and the values of the complexity metrics for this report are shown in Figure 2. Detailed observations are listed as follows:

- **Entity-level metrics have an influence on NER and FGS.**

  OOE exhibits the highest negative correlation of all the metrics in NER and FGS. This indicates that reports with a higher proportion of unknown NEs tend to exhibit lower performance.

  Similarly, EL exhibits a negative correlation with NER, indicating that reports with longer NE tend to have lower performance. For instance, complex *Diagnosis* NEs include noun phrases such as "薬剤性肺炎の再燃 (Recurrence of drug-induced pneumonia)." Such expressions make it challenging for the model to accurately determine the boundaries.

- **Graph-level metrics have a greater impact than report-level metrics.**

  Report-level metrics, indicating the complexity of the overall report, exhibit a lower correlation. Conversely, graph-level metrics, indicating the complexity of individual findings, exhibit a higher negative correlation. Sentences describing abnormal findings such as the first and second lines in Figure 2, tend to be linguistically

Complexity Metric	Definition of Metric	r_{NER}	r_{RE}	r_{FGS}
Out of Entity (OOE)	The percentage of entities not included in the training set.	-39.9†	-20.5†	-40.7†
Entity Length (EL)	The average number of characters per entity.	-28.5†	-10.4†	-26.2†
Report Length (RL)	The number of characters in the report.	-6.4	-17.2†	-16.7†
Report Relations (RR)	The number of relations in the report.	-0.2	-15.5†	-17.1†
Report Graphs (RG)	The number of graphs in the report.	8.5†	-6.6	4.0†
Graph Relations (GR)	The average number of relations per graph.	-4.7	-18.7†	-29.7†
Positive Graphs (PG)	The percentage of graphs where the factuality of the *Finding* is positive.	-22.1†	-17.3†	-37.7†
Positive Graph Length (PGL)	The average number of characters per sentences containing positive graph (PG).	-18.1†	-34.9†	-40.3†
Multiple-Finding Graphs (MG)	The percentage of graphs generated from sentences containing multiple graphs.	-5.5	-22.2†	-18.8†
Positive Multiple-Finding Graphs (PMG)	The percentage of graphs that are both positive graphs (PG) and multiple graphs (MG).	-19.7†	-30.9†	-39.8†

Table 5: Pearson's correlation coefficients r between F1 scores of NER, RE and FGS and complexity metrics in the validation set from facilities A to F. † denotes $p < 0.01$ in a significance test of the correlation.

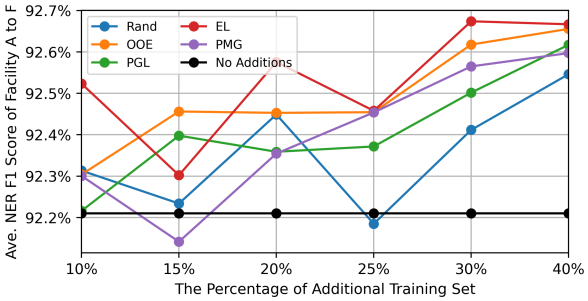

Figure 3: Average NER F1 scores of augmented models on reports from facilities A to F. "No Additions" represents the performance without any augmentation, as shown in Table 4a. The x-axis shows the percentage of the additional set relative to the original training set.

complex, as they need to convey the relevant attributes such as *Characteristics* for differential diagnosis. Consequently, the graphs generated from these complex sentences tend to be complex structures. The negative correlation observed in PGL and PMG suggests that the model struggles to accurately structure these complex sentences.

5.4 The Effect of Metric-Based Augmentation

In this section, we explore strategies to reducing performance degradation by augmenting the training set based on correlated metrics. A straightforward approach is to add reports from each facility to the training set. However, it is difficult to use data from multiple facilities as a training set due to contractual and cost constraints. Thus, we focused on adding only OUH reports to improve performance on reports from other facilities. This setting addresses a more challenging scenario and practical issues with a limited available training set.

We aim to achieve more efficient training by sampling additional OUH reports based on the key metrics identified in the previous section. Specifically, we examined whether this strategy improves performance on reports from other facilities more efficiently than random sampling. The performance gap is greater for F_{NER} than for F_{RE} (Table 4a). Therefore, we focused on NER in this experiment.

Facilities	No additions	Rand	OOE	EL	PGL	PMG
A to F	79.50	80.07	80.01	**80.27**†	80.11	79.59
E	77.20	77.82	78.29†	78.22†	78.17	**78.46**†
F	76.33	76.95	77.20	**77.38**	**77.38**†	77.18

Table 6: FGS F1 scores across facilities A to F, using 40% augmented NER model, whereas the RE model remained unchanged. † indicates a significant difference compared with **Rand**. (McNemar's test, $p < 0.01$)

5.4.1 Experimental Settings

We added a portion of the OUH test set to the training set and examined performance for facilities A to F. The sampling process is as follows: First, we made predictions on the OUH test set using a model trained on the training set. Next, we calculated each metric for each report from the prediction results. Finally, we selected reports with high values of the metrics preferentially and add them to the training set along with their gold annotations.

5.4.2 Experimental Results

Figure 3 shows the Macro-F1 scores on augmented NER models. The metrics-based augmentation tends to result in higher performance compared to random sampling. The augmented models using OOE and EL, which exhibited the highest negative correlation in the NER task (Table 5), achieved the best performance.

Table 6 shows the FGS scores when using the NER model with 40% augmented data, whereas the RE model remained unchanged. Similar to NER, performance improvements in FGS were observed. Facilities E and F, which initially had lower FGS F1 scores compared to others, demonstrated greater performance improvement.

We observed significant improvement in facility E with the OOE-based augmentation, but smaller improvement in facility F. Since only OUH reports were augmented, the increased diversity of NEs in OUH reports may not translate to other facilities. Therefore, for reports containing many facility-specific terms, the performance improvement from OOE-based augmentation may be limited. This is a limitation of using only single-facility reports to improve the performance of reports from other facilities. Additionally, PMG-based augmentation showed a lower score than random sampling across facilities A to F. As shown in Table 5, RE showed a higher correlation with PMG compared to NER. Thus, although the performance gap in F_{RE} is smaller than F_{NER}, incorporating this augmentation in RE could potentially improve F_{FGS}.

6 Conclusion

We addressed two key challenges in structuring radiology reports: the lack of a practical schema and datasets to evaluate model generalizability. To address these challenges, we proposed a FCS that structures radiology reports by each finding and constructed JRadFCS, a large-scale dataset containing 8,428 Japanese CT reports from seven facilities. We evaluated the performance of a model trained on single-facility reports applied to reports from other facilities, revealing performance gaps. We identified factors causing performance gaps and confirmed improvements of F1 scores on NER and FGS through augmentation based on these factors. Moreover, we observed that the improvement is larger for facilities with lower initial performance.

Our future work is to extend the JRadFCS dataset to include reports from other imaging modalities such as magnetic resonance and ultrasound. Additionally, we plan to demonstrate whether the FCS schema actually improves any downstream tasks.

Limitations

The JRadFCS dataset comprises only Japanese CT reports, raising uncertainty about how well the proposed FCS and the experimental observations generalize to reports in other languages or from other imaging modalities, such as magnetic resonance and ultrasound. In future work, we plan to expand the dataset to include reports in other languages and from these modalities. This direction could enable a more comprehensive evaluation of the FCS and its model generalizability.

Additionally, the JRadFCS dataset cannot be made publicly available due to ethical and privacy constraints, as it is derived from sensitive medical data. While this ensures compliance with data governance policies and the protection of patient confidentiality, it limits the broader adoption and reproducibility of our study.

Ethical Consideration

This study adheres to the Association for Computing Machinery (ACM) Code of Ethics and Professional Conduct[1], which has been adopted by the Association for Computational Linguistics (ACL).

All reports used in this study were de-identified; patient names, doctor names, contact information, and other identifiers were removed to protect patient privacy. Additionally, we did not use any accompanying information such as patient sex, age, purpose of the request, or diagnosis fields in this study. Radiology reports were collected with consent from the patients or their representatives, and the Institutional Review Board has approved this study.

References

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.

Fei Cheng, Shuntaro Yada, Ribeka Tanaka, Eiji Aramaki, and Sadao Kurohashi. 2022. JaMIE: A Pipeline Japanese Medical Information Extraction System with Novel Relation Annotation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3724–3731, Marseille, France. European Language Resources Association.

Surabhi Datta, Morgan Ulinski, Jordan Godfrey-Stovall, Shekhar Khanpara, Roy F. Riascos-Castaneda, and Kirk Roberts. 2020. Rad-SpatialNet: A Frame-based Resource for Fine-Grained Spatial Relations in Radiology Reports. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2251–2260, Marseille, France. European Language Resources Association.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360,

[1] https://www.acm.org/code-of-ethics

Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. 2024. RadGraph-XL: A Large-Scale Expert-Annotated Dataset for Entity and Relation Extraction from Radiology Reports. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12902–12915, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Les R. Folio, Laura B. Machado, and Andrew J. Dwyer. 2018. Multimedia-enhanced Radiology Reports: Concept, Components, and Challenges. *Radiographics*, 38(2):462.

Rikhiya Ghosh, Oladimeji Farri, Sanjeev Kumar Karn, Manuela Danu, Ramya Vunikili, and Larisa Micu. 2023. RadLing: Towards Efficient Radiology Report Understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 640–651, Toronto, Canada. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1):2:1–2:23.

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

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597. Number: 01.

Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, and Eiji Aramaki. 2018. J-MeDic: A Japanese Disease Name Dictionary based on Real Clinical Usage. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong N. Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, and Pranav Rajpurkar. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

Alistair E. W. 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):317. Publisher: Nature Publishing Group.

Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. 2021. A clinical specific BERT developed using a huge Japanese clinical text corpus. *PLOS ONE*, 16(11):e0259763. Publisher: Public Library of Science.

Wilson Lau, Kevin Lybarger, Martin L. Gunn, and Meliha Yetisgen. 2023. Event-Based Clinical Finding Extraction from Radiology Reports with Pre-trained Language Model. *Journal of Digital Imaging*, 36(1):91–104.

Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel Castro, Maria Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, Pranav Rajpurkar, Sameer Khanna, Hoifung Poon, Naoto Usuyama, Anja Thieme, Aditya Nori, Matthew Lungren, Ozan Oktay, and Javier Alvarez-Valle. 2023. Exploring the Boundaries of GPT-4 in Radiology. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14414–14445, Singapore. Association for Computational Linguistics.

Youmi Ma, Tatsuya Hiraoka, and Naoaki Okazaki. 2022. Joint Entity and Relation Extraction Based on Table Labeling Using Convolutional Neural Networks. In *Proceedings of the Sixth Workshop on Structured Prediction for NLP*, pages 11–21, Dublin, Ireland. Association for Computational Linguistics.

Yuta Nakamura, Shouhei Hanaoka, Yukihiro Nomura, Naoto Hayashi, Osamu Abe, Shunrato Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. Clinical Comparable Corpus Describing the Same Subjects with Different Expressions. In *MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation*, pages 253–257. IOS Press.

Daiki Nishigaki, Yuki Suzuki, Tomohiro Wataya, Kosuke Kita, Kazuki Yamagata, Junya Sato, Shoji Kido,

and Noriyuki Tomiyama. 2023. BERT-based Transfer Learning in Sentence-level Anatomic Classification of Free-Text Radiology Reports. *Radiology: Artificial Intelligence*, 5(2):e220097. Publisher: Radiological Society of North America.

Namu Park, Kevin Lybarger, Giridhar Kaushik Ramachandran, Spencer Lewis, Aashka Damani, Özlem Uzuner, Martin Gunn, and Meliha Yetisgen. 2024. A Novel Corpus of Annotated Medical Imaging Reports and Information Extraction Results Using BERT-based Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1280–1292, Torino, Italia. ELRA and ICCL.

Pinal Patel, Disha Davey, Vishal Panchal, and Parth Pathak. 2018. Annotation of a Large Clinical Entity Corpus. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2033–2042, Brussels, Belgium. Association for Computational Linguistics.

Daisaku Shibata, Emiko Shinohara, Kiminori Shimamoto, and Yoshimasa Kawazoe. 2024. Towards Structuring Clinical Texts: Joint Entity and Relation Extraction from Japanese Case Report Corpus. In *MEDINFO 2023 — The Future Is Accessible*, pages 559–563. IOS Press.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Portuguese Named Entity Recognition using BERT-CRF. *arXiv preprint*. ArXiv:1909.10649.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Kento Sugimoto, Shoya Wada, Shozo Konishi, Katsuki Okada, Shirou Manabe, Yasushi Matsumura, and Toshihiro Takeda. 2023. Extracting Clinical Information From Japanese Radiology Reports Using a 2-Stage Deep Learning Approach: Algorithm Development and Validation. *JMIR Medical Informatics*, 11(1):e49041. Number: 1 Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

Tohoku-BERT. 2024. tohoku-nlp/bert-base-japanese-v3. Accessed on 2024-12-2.

Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. 2023. Japanese SimCSE Technical Report. *arXiv preprint*. ArXiv:2310.19349.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, Relation, and Event Extraction with Contextualized Span Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Shuntaro Yada, Ayami Joh, Ribeka Tanaka, Fei Cheng, Eiji Aramaki, and Sadao Kurohashi. 2020. Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge: Starting From Critical Lung Diseases. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4565–4572, Marseille, France. European Language Resources Association.

Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. 2022. Real-mednlp: Overview of real document-based medical natural language processing task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, pages 285–296.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed Levitated Marker for Entity and Relation Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns (New York, N.Y.)*, 4(9):100802.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2022. Contrastive Learning of Medical Visual Representations from Paired Images and Text. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, pages 2–25. PMLR. ISSN: 2640-3498.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. RaTEScore: A Metric for Radiology Report Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, Miami, Florida, USA. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction.

In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

A Details of Rule-Based Processing to Generate Finding-Centric Graphs

To transform the output of NER and RE into finding-centric graphs, we applied the following two rules:

- **Segment-Path Rule** For the graphs containing multiple *segments*, finding-centric graphs are generated based on the paths from each terminal segment to the findings. For example, in Figure 1, two paths are identified:, "right lung → apex → nodule" and "right lung S2 → nodule"; thus, two finding-centric graphs are generated by adding each segment path.
- **Size-Path Rule** For the graphs containing multiple *Measurement results* indicating size of findings, finding-centric graphs are generated based on the edges from each size expression labeled *Measurement result* to the finding. Size expressions are combinations of numbers (e.g., "1.0," "1.0×1.5×2.0") and units ("mm" and "cm"), and we determine whether they are size expressions using regular expressions applied to the NEs labeled as *Measurement results*. For example, in the report "Nodules of 2cm and 3cm are seen," two finding-centric graphs are generated: one is the "2cm → nodule" and another is "3cm → nodule."

When multiple segments and size expressions appear within a single graph, we create pairs of segments and sizes according to their order of appearance and generate finding-centric graphs for each pair. For example, for the sentence "Nodules of 1cm in the right lung, and 2cm and 3cm in the left lung are seen," we create the graphs "right lung → nodule ← 1cm," "left lung → nodule ← 2cm," and "left lung → nodule ← 3cm" based on the order of appearance.

The aforementioned rules are simple, but there were no erroneous reports on the validation set. We concluded that radiologists avoid using complex structures that would make it difficult for readers to understand the size and location of abnormalities; therefore, no reports required more complex processing.

B JRadFCS Dataset

B.1 Named Entities and Factuality

Table 7 shows statistics of NE labels on the JRadFCS dataset. Unique expressions assigned the

NE Label	#NEs	#Unique NEs	Example of NEs
Finding	75,619	5,295	結節 (nodule), 腫瘤 (mass), すりガラス影 (ground-glass opacity), 嚢胞 (cyst)
Finding (Positive)	43,801	4,613	
Finding (Negative)	31,818	1,222	
Diagnosis	14,675	3,078	転移 (metastasis), 肺癌 (lung cancer), 良性病変 (benign lesion), 活動性病変 (active lesion)
Diagnosis (Positive)	11,882	2,893	
Diagnosis (Negative)	2,793	364	
Characteristics	5,908	1,170	低吸収 (low absorption), 不整 (irregular), 限局性 (localized), 石灰化 (calcification)
Characteristics (Positive)	6,219	1,074	
Characteristics (Negative)	857	219	
Temporal change	14,301	149	変化 (change), 増大 (increase)
Temporal change (Positive)	5,056	121	
Temporal change (Negative)	9,245	57	
Segment	56,191	6,954	肺 (lung), 主膵管 (main pancreatic duct), 頭部 (head), 大腿骨 (thigh bone)
Measurement result	5,185	872	大きい (large), 高い (high), 縮小 (reduction), 10mm, 1.2×2.5cm
Measurement item	3,290	194	長径 (major axis), CT値 (CT value)
Quantity	2,283	55	複数 (several), 多数 (many), 2個 (two)

Table 7: Statistics of NE labels on JRadFCS dataset. **#NEs** and **#Unique NEs** denote the number of NEs and unique NEs, respectively.

Factuality	Example of Frequency Clue Expression
Positive	認められる (is seen), 疑われる(is suspected), 出現 (appear), (+)
Negative	明らかでない (is not clear), 消失 (disappear), (-)

Table 8: Examples of clue expressions for annotating factuality labels.

Quantity and *Temporal change* labels are limited, however, the *Finding* , *Diagnosis*, *Characteristics* labels have diverse expressions.

We assigned a factuality attribute to *Finding*, *Characteristics*, *Temporal change*, and *Diagnosis*: Positive if the entity is observed, and Negative if it is not. The factuality can be assigned based on clue expressions. Examples of these frequently occurring clue expressions are presented in Table 8.

B.2 Relations

Table 9 shows the statistics of relations in the JRadFCS dataset. As stated in the examples of Table 7 and Table 9, the JRadFCS dataset includes segment and disease terms for various organs. This indicates that JRadFCS encompasses radiology reports addressing the anatomy of the entire body and a broad spectrum of diseases.

C Annotation Process

We employed two annotators with over 10 years of experience in medical domain NLP tasks to annotate NEs and relations. We used Brat (Stenetorp et al., 2012) for annotation.

We randomly sampled 5 reports from each facility, resulting in a total of 35 reports, to calculate the Inter-Annotator Agreement between the two annotators. Since this task involved annotating both NEs and relations, we calculated the F1 score based on perfect matches in the span, label, and factuality of both the subject and object NEs, as well as the relations between NEs. The precision, recall, and F1 score are 0.88, 0.87, and 0.88, respectively.

C.1 Statistics

Table 10 shows the statistics of reports in the JRadFCS dataset. It can be observed that the statistics of reports vary by facility. This variation suggests that different facilities and radiologists have different styles of reporting, such as whether multiple findings are summarized in one sentence or listed individually. Similar analysis were reported by Nakamura et al. (2022). This statistics and diversity emphasize the importance of evaluating model performance across diverse reports.

D Details of Training

D.1 JRadBERT

We trained a BERT-based model using Japanese radiology reports to construct a PLM specialized for radiology. The details of JRadBERT are described below.

Subject	Object	#Relations	Example of Relations
Segment	*Finding*	48,446	脾→異常 (spleen→abnormality), 両腎→嚢胞 (bilateral kidneys→cyst)
Diagnosis	*Finding*	18,011	肺転移→結節 (lung metastasis→nodule), 嚢胞→低吸収域 (cyst→low absorption area)
Characteristics	*Finding*	6,936	石灰化→腫瘤 (calcification→mass), 病的→液体貯留 (pathological→fluid accumulation)
Temporal change	*Finding*	16,157	変化→結節 (change→nodule), 増大→腫瘤 (increase→mass)
Measurement result	*Finding*	4,981	粗大→出血 (coarse→hemorrhage), 少量→腹水 (small amount→ascites)
Quantity	*Finding*	2,361	多発→嚢胞 (multiple→cyst), 散見→低吸収域 (scattered→low absorption area)
Measurement item	*Measurement result*	1,924	径→1cm (diameter→1cm), サイズ→小さく (size→small)
Segment	*Segment*	2,628	縦隔→リンパ節 (mediastinum→lymph nodes), 甲状腺→両葉 (thyroid→bilateral lobes)

Table 9: Statistics of relations in the JRadFCS dataset. **#Relations** denotes the number of relations.

Facility	$\overline{Sents}$	$\overline{Words}$	$\overline{NEs}$	$\overline{Relations}$	$\overline{Graphs}$
OUH	12.6 / 13.1 / 9.9	128.7 / 132.7 / 92.1	26.3 / 27.6 / 18.9	14.3 / 15.1 / 9.8	10.3 / 10.0 / 8.7
A	0 / 9.4 / 9.3	0 / 96.9 / 97.6	0 / 19.5 / 19.9	0 / 11.0 / 11.6	0 / 11.6 / 11.8
B	0 / 13.3 / 13.1	0 / 148.5 / 147.1	0 / 29.4 / 29.0	0 / 18.9 / 18.7	0 / 15.3 / 15.0
C	0 / 11.6 / 12.0	0 / 103.7 / 109.7	0 / 20.8 / 21.5	0 / 11.3 / 11.6	0 / 11.4 / 11.4
D	0 / 9.9 / 9.7	0 / 102.8 / 102.9	0 / 20.5 / 20.5	0 / 11.5 / 11.7	0 / 10.9 / 10.7
E	0 / 11.1 / 10.3	0 / 107.9 / 98.2	0 / 19.3 / 17.8	0 / 11.7 / 10.8	0 / 9.0 / 8.5
F	0 / 7.8 / 8.1	0 / 75.3 / 77.8	0 / 13.3 / 13.9	0 / 7.6 / 8.2	0 / 6.5 / 6.8

Table 10: Statistics of reports in the JRadFC dataset and their distribution into training, validation, and test sets. $\overline{Sents}$, $\overline{Words}$, $\overline{NEs}$, $\overline{Relations}$, and $\overline{Graphs}$ represent the average number of sentences, words, NEs, relations, and finding-centric graphs, respectively.

	NER	RE
Batch size	8	32
Epoch size	10	10
Learning rate	Linear warmup for the first 10% of train steps to 5e-5, then linear decay to 0	
Dropout rate	0.1	0.1
Optimizer	AdamW	AdamW

Table 11: The hyperparameters of NER and RE.

Dataset We used approximately 15 years of radiology reports from OUH for training. This dataset consists of 758,017 Japanese radiology reports (over 10.6M sentences and 103.3M words). Additionally, no overlapping reports or patients between this pre-training dataset and the reports were included in JRadFCS.

Pre-processing As pre-processing steps for the input reports, we sequentially applied NFKC normalization, converted text to lowercasing, and replaced spaces with underscores.

Tokenizer We constructed a character-level tokenizer with a vocabulary of 3,930 tokens. The pre-processed input reports are first tokenized by MeCab with the IPA dictionary and then split into characters.

Training JRadBERT was trained using a masked language model with a Whole-Word-Masking strategy, where 15% of the words in the input report were masked. This model was trained for 30 epochs. The batch size was set to 256 and the max token length to 512.

D.2 NER and RE

We fine-tuned JRadBERT using OUH training set to train the NER and RE models. We did not use the validation sets for facilities A to F for training or selecting the best model. The hyperparameters of NER and RE are defined in Table 11. These parameters were determined by a Grid search, evaluating the performance against the OUH validation set across several variations.

E Performance of Each Label on NER and RE

E.1 NER

Table 12 shows F1 scores for each label on the test set using the JRadBERT model fine-tuned on the train set. It can be observed that the performance for the *Characteristics* is low compared to other lables, across all facilities. From Table 12, it is evident that *Characteristics* has a high number of unique NEs despite its low frequency compared to other labels. This result suggests that to correctly

NE Label	OUH	A	B	C	D	E	F	Average
Finding (Positive)	92.49	91.45	88.82	87.36	87.18	89.10	87.46	89.26
Finding (Negative)	97.74	95.26	96.17	95.24	94.38	92.21	94.90	95.33
Diagnosis (Positive)	93.64	82.87	90.06	88.26	87.18	88.65	85.69	88.80
Diagnosis (Negative)	95.22	92.70	87.88	94.99	92.91	88.77	90.45	91.06
Characteristics (Positive)	80.98	75.22	75.51	73.68	76.22	79.03	67.99	76.50
Characteristics (Negative)	72.62	68.18	54.29	51.85	57.67	59.15	52.94	62.37
Temporal change (Positive)	96.68	94.67	92.17	94.85	90.23	94.41	94.41	94.01
Temporal change (Negative)	98.56	97.58	96.00	97.93	93.08	98.57	96.08	97.02
Segment	98.01	96.46	96.35	94.18	93.96	94.89	92.69	95.48
Measurement result	98.05	94.37	94.28	93.27	91.94	94.30	92.60	94.43
Measurement item	87.17	83.93	80.00	67.03	70.73	79.01	70.59	78.72
Quantity	98.21	96.36	98.87	97.74	97.94	97.85	96.00	97.50

Table 12: F1 scores for each label on the test set using the JRadBERT model fine-tuned on the train set.

Subject	Object	OUH	A	B	C	D	E	F	Average
Segment	*Finding*	96.29	96.19	96.24	95.02	95.20	95.47	94.69	95.59
Diagnosis	*Finding*	93.32	90.41	93.59	93.59	92.97	93.06	91.91	92.69
Characteristics	*Finding*	89.90	86.78	87.94	87.97	89.91	89.46	86.97	88.42
Temporal change	*Finding*	96.72	94.99	94.70	95.57	94.61	95.26	95.73	95.37
Measurement result	*Finding*	97.46	94.82	97.82	95.45	96.10	98.20	93.63	96.21
Quantity	*Finding*	98.55	90.66	98.03	95.61	96.37	96.44	97.50	96.16
Measurement item	*Measurement result*	99.08	96.40	96.30	98.31	93.12	98.95	81.48	94.80
Segment	*Segment*	86.77	84.36	81.64	83.74	85.78	86.69	84.08	84.72

Table 13: F1 scores for each relation on the test set using the JRadBERT model fine-tuned on the train set.

	JRadBERT	GPT-4o		
		1-shot	10-shots	20-shots
OUH	83.31	43.79	53.77	57.36
A	83.94	32.54	41.04	44.19
B	81.51	37.79	44.94	46.51
C	82.48	34.60	46.82	50.85
D	79.09	33.59	37.96	40.40
E	74.96	22.53	36.28	38.90
F	74.18	30.79	39.15	40.26

Table 14: Comparison of FGS F1 scores between GPT-4o and JRadBERT, on validation set. To evaluate GPT4o, we append few examples of reports and their gold outputs as a few-shot setting.

predict *Characteristics*, the model needs to rely not only on the surface form of the words but also on the contextual information.

E.2 RE

Table 13 shows F1 scores for each relation on the test set using the JRadBERT model fine-tuned on the train set. It can be observed that the performance for the relations between *Characteristics* and *Finding* is particularly low among the relations targeting *Finding*. Predicting the relation from *Diagnosis* to *Finding* is relatively easy compared to predicting the relation from *Characteristics* to *Finding*. This is because diagnoses are determined by synthesizing information from all findings. Consequently, in cases where both finding and diagnosis appear in a sentence, a relation is usually linked between them. On the other hand, characteristics differ for each finding, the model only needs to link related characteristics and findings. This difficulty is causing performance degradation.

Additionally, our RE model can not takes the NE label information. Therefore, to utilize NE label information in the RE model, we could improve performance to change the model into a NE marker model (Zhong and Chen, 2021; Ye et al., 2022) or a multi-task model for NER and RE (Wadden et al., 2019; Ma et al., 2022).

F GPT-4o Evaluations

We benchmarked the performance of GPT-4o on the JRadFCS validation set. Given an input radiology report, we used GPT-4o to extract the entire finding-centric graphs. Table 15 shows the prompt used for GPT-4o evaluations. Table 16 shows an English translation of the Japanese prompt.

Table 14 shows the FGS F1 scores of GPT-4o and JRadBERT on the validation set. GPT-4o performed significantly lower than JRadBERT. Our error analysis revealed that GPT-4o fails to extract NEs according to our schema. For example, in the sentence "気道病変を思わせる粒状影あり。(There are granular shadows suggestive of airway disease.)," GPT-4o incorrectly extracted "気道病変を思わせる (suggestive of airway disease)" as a *Diagnosis*. The term "思わせる (suggestive of)" is a clue of positive factuality and signifies a relation between "気道病変 (airway disease)" and "粒状影 (granular shadows)," but it does not need to be extracted as a separate entity. We qualitatively confirmed that GPT-4o is particularly prone to making such mistakes with expressions that are not included in the few-shot samples.

```
# 質問

## タスク
- あなたのタスクは入力される読影レポートを所見毎に関連する情報と共に構造化することです。下記の指示に従って構造化処理を行って下さい。

## 指示
- Segment、Finding、Diagnosis、Characteristics、Temporal change、Measurement result、Measurement item、Quantityに該当する用語を抽出する。
- 用語クラスの定義は以下に定める。
    - Segment：臓器または臓器を解剖学定義に基づいて区画した領域
    - Finding：画像上で医師が指摘した異常（正常ではない状態・変化）を指す用語
    - Diagnosis：findingから推定・判断される情報を指す用語。標準病名マスタの用語とその同義語
    - Characteristics：findingの状態や性質などの特徴や撮影画像上での明暗や染まりの度合を示す用語
    - Temporal change：findingの経時的な変化表現
    - Measurement result：findingの計測された値や定性的なサイズを示す用語
    - Measurement item：findingの計測した項目を示す用語
    - Quantity：findingの数を示す用語
    - 複合名詞に対して、重複したスパンで用語を抽出することはなく、1つのクラスを割り当てる。
    - 抽出した用語のクラスがFinding、Diagnosis、Characteristics、Temporal changeの場合は、factualityとして0か1で判定する。
    - factualityは「認めない、ない」など対象の用語が存在しない場合は0、「認める、疑う」など存在している場合は1とする。
    - factualityを判断するための手がかりとなる表現は抽出しない。

- 抽出した用語に対して、findingを中心とした用語間の関係性を抽出する。
    - Segment→Finding：抽出した所見とその所見が確認された区域との関係
    - Diagnosis→Finding：抽出した所見から疑われる診断情報との関係
    - Characteristics→Finding：抽出した所見とその所見の性状との関係
    - Temporal change→Finding：抽出した所見とその所見の経時変化との関係
    - Measurement result→Finding：抽出した所見とその所見の計測項目との関係
    - Quantity→Finding：抽出した所見とその所見の個数との関係
    - Measurement item→Measurement result：抽出した計測項目に対応する計測結果との関係
    - Segment→Segment：解剖学的に上位の解剖区域から下位の解剖区域への関係
- 抽出した用語と関係性から読影レポートを所見毎に構造化する。

## 入力レポートと出力の例
{"input": "肝臓に嚢胞あり。 ...", "output": ["Segment": ["word": "肝臓"],"finding":"word": "嚢胞", "factuality": 1, ...],...}

## 出力形式
- 出力形式はjsonである。
- キーの"output"に対する値はlist型とし、そのlistの各要素はdict型とする。このdictにある1つのFindingとそのFindingに関連する情報が格納される。
- "word"には入力レポートに含まれる用語クラスに概要する表現を格納する。
- キーの"finding"は必ずdict型とする。その他は複数の要素が存在する可能性があるため、全てlist型とする。
- 入力レポートにFindingに該当する用語がなく、Diagnosisに概要する用語がある場合はwordとfactualityをFindingとして抽出し、Diagnosisとしては抽出しない。
- 「肝臓のS1」というようにSegmentに該当する用語が階層関係にある場合は、同一のリストに上位階層の区域から順に格納する。
- 入力レポート中に含まれるFindingの数だけdictを作成し、格納する。
    - 同一のFindingが異なる複数のSegmentで確認されているレポートの場合
    - Findingと関係するSegmentの数と同数の構造化結果を作成する。
        - 同一のFindingが異なる複数のサイズを示すMeasurement result(3cm 等)と関係をもつ存在する場合
        - Findingと関係するサイズを示すmeasurement resultの数と同数の構造化結果を作成する。

上述の指示通りに質問に答えてください。
繰り返しになりますが、この会話内で、構造化するとは、出力形式に従った構造化を指し、必ずjsonで出力して下さい。
```

Table 15: Japanese input prompt used by GPT-4o in order to extract finding-cetric graphs. For few-shot prompting, we append example reports and its ideal outputs to the end of this prompt.

```
# Question

## Task
- Your task is to structure the incoming radiology report with related information for each finding as instructed below.

## Instructions
- Extract terms that correspond to Segment, Finding, Diagnosis, Characteristics, Temporal change, Measurement result, Measurement item, and Quantity.
- The definitions of term classes are specified as follows:
    - Segment: Terms indicating regions based on anatomical definitions, such as organs or parts of organs.
    - Finding: Terms indicating abnormalities or abnormal conditions.
    - Diagnosis: Terms indicating diseases inferred from the findings.
    - Characteristics: Terms indicating features of findings, such as state, nature, or degree of brightness.
    - Temporal change: Terms indicating changes compared to past tests.
    - Measurement result: Terms indicating measured values or qualitative size expressions.
    - Measurement item: Terms indicating items for measured values.
    - Quantity: Terms indicating the number of findings.
    - For compound nouns, do not extract terms in duplicate spans but assign a single class.
    - If the extracted term class is Finding, Diagnosis, Characteristics, or Temporal change, determine factuality as 0 or 1.
    - Factuality should be 0 if terms like "not observed" or "absent" indicate the term does not exist, and 1 if terms like "recognized" or "suspected" indicate it exists.
    - Do not extract expressions that provide clues for determining factuality.

- For the extracted terms, extract the relationships between terms centered on the finding.
    - Segment→Finding: Indicates where the finding is located with in the anatomical structure.
    - Diagnosis→Finding: Represents the suspected diagnosis from the finding.
    - Characteristics→Finding: Represents the characteristics of the finding.
    - Temporal change→Finding: Represents the temporal changes of the finding.
    - Measurement result→Finding: Represents the measurement results of the finding.
    - Quantity→Finding: Represents the number or amount of the finding.
    - Measurement item→Measurement result: Links the items of measurement to its result.
    - Segment→Segment: Shows the spatial relationship between two segments. Links from higher-level to lower-level segments.

## Input Report and Output Example
{"input": "There is a cyst in the liver. ...", "output": ["Segment": ["word": "liver"], "finding": "word": "cyst", "factuality": 1, ...], …}

## Output Format
- The output format should be JSON.
- The value corresponding to the key "output" should be a list, and each element of this list should be a dictionary. This dictionary will contain one Finding and related information for that Finding.
- The "word" will store the expression corresponding to the term class found in the input report.
- The key "finding" should always be a dictionary, and other keys should be lists as they may contain multiple elements.
- If there is no term corresponding to Finding in the input report but there is a term corresponding to Diagnosis, extract it as "word" and "factuality" for Finding, and do not extract it as Diagnosis.
- If terms corresponding to Segment have hierarchical relationships such as "S1 of the liver", store them in the list in order from the higher-level region to the lower-level region.
- Create and store a dictionary for each finding present in the input report.
    - In the case of reports where the same Finding is confirmed in different Segments:
    - Create as many structuring results as the number of Segments relating to the Finding.
        - If the same Finding has multiple related Measurement results indicating different sizes (e.g., "3cm"):
        - Create as many structuring results as the number of size-indicating Measurement results relating to the Finding.

Answer the question according to the instructions above.
Once again, in this conversation, structuring refers to structuring as per the output format, and always output in JSON.
```

Table 16: An English translation of the Japanese prompt.