MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220073

Clinical Comparable Corpus Describing the Same Subjects with Different Expressions

Yuta Nakamura[a], Shouhei Hanaoka[a, b], Yukihiro Nomura[c], Naoto Hayashi[c], Osamu Abe[a, b], Shunrato Yada[d], Shoko Wakamiya[d], Eiji Aramaki[d]

[a] Division of Radiology and Biomedical Engineering, Graduate School of Medicine, The University of Tokyo, Bunkyo, Tokyo, Japan
[b] The Department of Radiology, The University of Tokyo Hospital, Bunkyo, Tokyo, Japan
[c] The Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, Bunkyo, Tokyo, Japan
[d] Nara Institute of Science and Technology, Ikoma, Nara, Japan

Abstract

Medical artificial intelligence (AI) systems need to learn to recognize synonyms or paraphrases describing the same anatomy, disease, treatment, etc. to better understand real-world clinical documents. Existing linguistic resources focus on variants at the word or sentence level. To handle linguistic variations on a broader scale, we proposed the Medical Text Radiology Report section Japanese version (MedTxt-RR-JA), the first clinical comparable corpus. MedTxt-RR-JA was built by recruiting nine radiologists to diagnose the same 15 lung cancer cases in Radiopaedia, an open-access radiological repository. The 135 radiology reports in MedTxt-RR-JA were shown to contain word-, sentence- and document-level variations maintaining similarity of contents. MedTxt-RR-JA is also the first publicly available Japanese radiology report corpus that would help to overcome poor data availability for Japanese medical AI systems. Moreover, our methodology can be applied widely to building clinical corpora without privacy concerns.

Keywords:

Radiology Report, Natural Language Processing, Artificial Intelligence

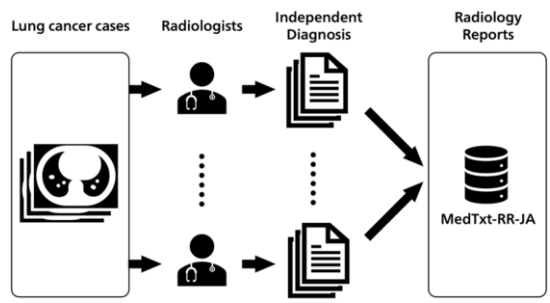
Introduction

Recent progress in deep learning techniques has enabled medical artificial intelligence (AI) systems to collect information quickly from clinical documents [1]. Real-world clinical documents are full of synonyms and paraphrases [2]. There are often substitutes to describe the same anatomy, disease, treatment, etc., such as "epidermal cyst" and "atheroma." For a better understanding of clinical documents, AI systems must be able to integrate such variants correctly without being confused by the superficial differences; however, they are still under development.

There have been helpful resources to handle a variety of expressions. For example, the unified medical language system (UMLS) [3] is an ontology that maps synonyms onto a canonical form (e.g., normalization of "lung cancer" and "pulmonary cancer" into "malignant neoplasm of lung"). MedNLI [4] is a dataset for training AI systems for natural language inference (NLI), which is a task to respond to yes-no questions as to whether a hypothesis can be inferred from a premise (e.g., to answer yes given the premise "the patient presented with seizures and hypotension" and the hypothesis "the patient had neurological symptoms"). UMLS focuses on word-level diversity, whereas MedNLI focuses on sentence-level diversity by handling entirely different sentences that describe the same phenomenon.

Additionally, a "comparable corpus" can be another valuable resource by serving as a box of assorted expressions with word-, sentence-, and document-level diversity. A "comparable corpus" refers to more than two sets of documents sampled from different populations but in the same strategy [5]. In machine translation, for instance, collections of English newspapers and French newspapers can constitute a comparable corpus across languages. Similarly, sampling clinical documents written in the same language but by different *clinicians* can also provide a comparable corpus since clinicians have different preferences of expressions to describe the same content in clinical documents. Building such a comparable corpus across clinicians can be helpful in tackling the difficulty in managing the diversity of expressions.

Figure 1 – Overview of MedTxt-RR-JA, which is built by multiple radiologists diagnosing the same lung cancer cases independently.



This paper accompanies the release of the Medical Text Radiology Report section Japanese version (MedTxt-RR-JA), the first clinical comparable corpus. As in Figure 1, MedTxt-RR-JA was built by recruiting radiologists, assigning them the same task to diagnose the same lung cancers independently, and collecting the produced radiology reports. MedTxt-RR-JA thus provides a variety of expressions for the same medical concepts by multiple clinicians. Because we used lung cancer cases in Radiopaedia, an open-access radiological reference, MedTxt-RR-JA can be made public without privacy concerns.

Our contributions are threefold. First, we proposed the first clinical comparable corpus that can be useful in investigating how the same medical subject is described with various synonyms and paraphrases among clinicians. Second, we built the first publicly available corpus of Japanese radiology reports, which can relieve the poor accessibility of non-English clinical documents. Third, we provided a procedure to create a public radiology report corpus without suffering from privacy concerns. MedTxt-RR-JA is available on the website of the Social Computing Laboratory, Nara Institute of Science and Technology (https://sociocom.naist.jp/medtxt-en/rr).

Methods

Radiopaedia.org

We used lung cancer cases in Radiopaedia, an open-access radiology reference with articles and case presentations. Radiopaedia was considered suitable for four reasons: (i) It provides a wide variety of radiological cases with the same disease but different details. (ii) Radiological images in Radiopaedia can be easily shared. Each case can be accessed by simply entering a URL to a web browser with no sign-up or application. (iii) We were free from privacy concerns. Since all cases in Radiopaedia are already de-identified, radiology reports for Radiopaedia cases can be published without privacy breaches. (iv) The diagnostic experience on Radiopaedia is close to the real-world workflow. Radiopaedia offers a lightweight interface for browsing radiological images and scrolling slices through a series. It is like a picture archiving and communication system (PACS) viewer, which is familiar to radiologists.

Case selection

Lung cancer cases were selected to cover most of the imaging findings involved in lung cancer staging according to the 8th edition of the Union for International Cancer Control (UICC) TNM Classification [6]. First, we performed text-based searches in Radiopaedia cases with six queries: "lung cancer," "lung carcinoma," "pulmonary cancer," "pulmonary carcinoma," "bronchogenic cancer," and "bronchogenic carcinoma." Meanwhile, three mandatory conditions were set: the lung cancer was pathologically proven (Diagnosis = "Certain"), at least one CT study is included (Modality = "CT"), and the chest is covered in the field-of-view (Systems = "Chest"). Second, all of the matched cases were manually reviewed, and cases were excluded if (i) all of the CT studies were a part of PET or SPECT studies, (ii) neither the sex nor the age of the patient was provided, or (iii) no CT images were presented as a stacked series but as a single key slice. Finally, we selected 15 lung cancer cases covering 26 of the 36 image findings used in lung cancer staging. Table 1 lists the details.

Table 1– 15 lung cancer cases selected for MedTxt-RR-JA.

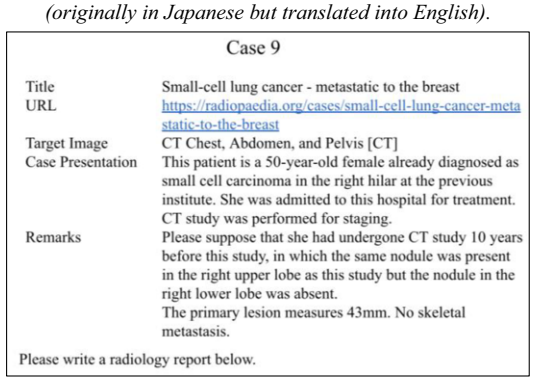
Case No.	Stage	Case Title
1	T1N0M0	Adenocarcinoma in situ of lung
2	T1N0M0	Minimally invasive adenocarcinoma of the lung
3	T1N0M0	Lepidic predominant adenocarcinoma of the lung
4	T2aN0M0	T2a lung cancer
5	T2bN0M0	Adenocarcinoma of the lung
6	T2N3M0	Squamous cell carcinoma of the lung
7	T3N1M0	Squamous cell carcinoma of the lung
8	T3N3M0	Pancoast tumor
9	T3N3M1c	Small-cell lung cancer - metastatic to the breast
10	T4N0M0	Bronchogenic carcinoma with upper lobe collapse
11	T4N0M1a	Pancoast tumor with cystic cerebral metastasis
12	T4N0M1c	Bronchoalveolar carcinoma
13	T4N2M0	Cavitating lung cancer
14	T4N3M1a	Bronchogenic carcinoma with left atrial large deposit - T4N3M1a
15	T2N2M1c	Metastatic pulmonary small cell carcinoma (SCC) with left upper lobe collapse

Data collection

Next, some modifications were made to the age, sex, and clinical background information of the cases to avoid inter-observer disagreement of diagnosis as much as possible. This is because the absence of clinical information can sometimes hinder radiologists from narrowing down the differential diagnoses. For instance, when advanced lung cancer is accompanied by a well-marginated small nodule, it is sometimes difficult to determine whether the nodule is benign (e.g., a granuloma or an intrapulmonary lymph node) or malignant (e.g., a metastasis) without prior CT studies for comparison. This ambiguity can lead to unwanted variations in lung cancer staging. For each case, all the images and case presentations were carefully reviewed. We then provided remarks to complement clinical information, such as "please suppose that there are prior CT studies of this patient and the lung nodule in the same lobe as the primary lesion has been unchanged for five years." We also added size information to the cases with missing tumor diameters. We thus indirectly guided radiologists to make specific decisions by adding or changing clinical information, not by explicitly showing a list of findings and impressions to write, because we wanted them to report naturally as they do in the actual radiological practice.

We prepared a reporting form in a docx format for each case. The reporting forms had the case title, the URL, and the modified case presentation accompanied by the remarks. The sample is shown in Figure 2. We ordered a teleradiology service in Japan to recruit radiologists. We then sent the reporting forms to the radiologists and asked them to access the URL, browse the radiological images, and write radiology reports. We imposed no restrictions but to use clinical information in the reporting forms. We obtained consent to publish the radiology reports from all of the radiologists.

Figure 2 – Reporting forms used in the MedTxt-RR-JA project (originally in Japanese but translated into English).


Case 9

Title	Small-cell lung cancer - metastatic to the breast
URL	https://radiopaedia.org/cases/small-cell-lung-cancer-metastatic-to-the-breast
Target Image	CT Chest, Abdomen, and Pelvis [CT]
Case Presentation	This patient is a 50-year-old female already diagnosed as small cell carcinoma in the right hilar at the previous institute. She was admitted to this hospital for treatment. CT study was performed for staging.
Remarks	Please suppose that she had undergone CT study 10 years before this study, in which the same nodule was present in the right upper lobe as this study but the nodule in the right lower lobe was absent. The primary lesion measures 43mm. No skeletal metastasis.

Please write a radiology report below.

Post-processing

Cleaning

We manually reviewed the radiology reports and made minimal necessary modifications to ensure clinical validity. We only

corrected suspected typos or contradictory stagings, such as "a tumor with a diameter of 60mm compatible with *T2* lung cancer is present," where *T3* is correct. All of the errors were fixed after confirmation by the involved radiologist. We made no further changes to the radiology reports.

Tagging of medical concepts

We tagged the following biomedical entities in each radiology report using the tag set proposed by Yada et al. [7]

- <D>: Diseases and symptoms, accompanied by existence "positive", "suspicious", or "negative"
- <A>: Anatomical terms
- <F>: Features and measurements
- <C>: Temporal changes
- <TIMEX3>: Temporal expressions
- <T-test/key/val>: Test names, value names, and values of clinical examinations
- <M-key/val>: Names and doses of medications
- <R>: Treatments other than medications (the abbreviation of "Remedy")
- <CC>: Clinical Context

Analysis

We analyzed how well the corpus satisfied the two objectives: inter-observer agreement of diagnosis and linguistic diversity.

Inter-observer agreement of diagnosis

We checked the clinical staging (i.e., stage IA to IVB) that was assigned to each case by each radiologist. Subsequently, we calculated the Krippendorff's alpha coefficient [8] for ordinary scales to measure the inter-observer agreement. We also assessed fine-grained level inter-observer agreement for each of the T, N, and M staging. For this analysis, we treated both T2a and T2b as T2 because our case collection contained lung cancers staged T2 for findings other than the primary lesion diameter, which do not fit the subdivision. We did not differentiate Tis, T1a, T1b, or T1c because the subdivision was sometimes impossible when radiology reports for a T1 lung cancer case only mentioned the size of the whole sub-solid primary lesion, not the size of its solid component. We used the fast-krippendorff v0.4.0 Python library for the calculations.

Document-level diversity

We evaluated the diversity of radiology reports with the following three measures:

- Variance of lengths
- Variance of the number of positive, suspicious, and negative findings
- Automated metric for diversity measurement

Radiologists may have various preferences regarding how intensively they describe imaging findings with little clinical impact. This may result in variances of (i) the word counts of radiology reports and (ii) the number of positive, suspicious, and negative findings appearing in radiology reports. To count words, we split each radiology report into words using the MeCab tokenizer [9] and Manbyo Dictionary as of July 2019 [10]. Note that Japanese word boundaries cannot be uniquely determined and depend on the choice of tokenization tools, because no punctuation is placed between words in Japanese orthography. To count positive, suspicious, and negative findings mentioned by each radiologist, we calculated the total frequency of <D> tags with certainty "positive," "suspicious," and "negative" for each radiologist, respectively. We then performed a chi-square test of independence with a significance level of 0.05 to examine the diversity among radiologists.

We also calculated the Self-BLEU metric [11] for each lung cancer case to measure document-level diversity. Self-BLEU is the application of BLEU [12] for the measurement of document diversity. BLEU evaluates the similarity between two documents based on the word-level overlap. BLEU ranges between 0 and 1, and a high BLEU score indicates a high similarity. Self-BLEU shows how *less* diverse a group of documents is, and is an average of $n \times (n\text{-}1)/2$ BLEU scores calculated for all the pairs among the targeted n documents. A low Self-BLEU score indicates higher diversity. We calculated the Self-BLEU scores for each lung cancer case using nltk v3.5 and scipy v1.5.4 Python libraries.

Sentence-level to word-level diversity

We manually reviewed synonyms or paraphrases in MedTxt-RR-JA and observed how well sentence- and word-level diversity of expressions were represented in MedTxt-RR-JA.

Results

Radiologists and Radiology Reports

Nine board-certified radiologists participated, who had 13.8-year experience on average in diagnostic radiology. A total of 135 radiology reports were collected (15 cases × 9 radiologists). Below are the samples translated into English:

- (Case 6, Radiologist 4) Atelectasis is seen in the right lower lobe. The right inferior lobar bronchus is occluded, and a mass suspected to be lung cancer is seen in the proximal part of the bronchus. The mass seems to involve enlarged mediastinal lymph nodes around the bronchus. It is difficult to determine the boundary between atelectasis and the tumor, and accurate tumor size cannot be evaluated. No findings suggestive of mediastinal infiltration. No lesions suspected to be accessory tumor nodules can be noted in the lung field. It seems to be equivalent to T2 in T classification. #1L, #2R, #2L, #4R, #4L, #5, #6, and #7 lymph nodes are swollen in a round shape and metastasis is suspected. It seems to be equivalent to N3 in N classification. No pleural effusion is observed. No significant abnormality can be identified in the upper abdominal organs.
- (Case 6, Radiologist 9) A 15 mm large nodule is seen in the lower lobe of the right lung, obstructing the trachea of the lower right lobe. There is atelectasis of the lower lobe of the right lung. In addition, the subcarinal lymph nodes are swollen to 22 mm in size, and lymphadenopathy is also seen in the bilateral mediastinal lymph nodes and the left supraclavicular lymph node. Based on the above, lung cancer and multiple lymph node metastasis are suspected. No tumors are seen in the liver, adrenal glands, or bone. There is a decrease in the Th7 vertebral body height, and compression fracture is suspected.

Inter-observer agreement of diagnosis

As shown in Table 2, the Krippendorff's alpha coefficient for the staging by nine radiologists exceeded 0.9, implying that the lung cancer staging agreed well. Besides, at a more fine-grained level, the alpha coefficients for T and M stagings were also above 0.9. The, the alpha coefficient for N staging was 0.7559. The comparatively low agreement may be attributed to two reasons. First, some lung cancer cases with multiple lymph node metastases had insufficient information on the size of the lymph nodes, and size measurement function was not available in the Radiopaedia image viewer. This may have split the judgments on equivocally enlarged lymph nodes. Second, we

treated all the radiology reports as staging N2 if they mentioned multiple mediastinal lymph node metastases without clarifying the laterality, resulting in disagreements between N2 (unilateral mediastinal lymph node metastases) and N3 (bilateral metastases) stagings among radiologists in some cases.

Table 2– Inter-observer agreement of the staging of lung cancer cases in MedTxt-RR-JA.

Target	Krippendorff's alpha
Staging	0.9193
T staging	0.9165
N staging	0.7559
M staging	0.9372

Corpus Diversity

Document-level diversity

Each radiologist wrote radiology reports of different lengths, even with the same lung cancer cases. Table 3 shows that the average lengths of radiology reports ranged from 83.7 to 203.1 words. Table 4 shows the total of positive, suspicious, and negative findings mentioned by each radiologist. The ratios among the three findings were significantly different among the radiologists (p<0.0001). Table 5 shows the Self-BLEU scores, which would serve as a baseline to assess the diversity of outputs of automated reporting AI systems in Japanese [11].

Table 3– Average lengths of radiology reports for 15 lung cancer cases by each radiologist in MedTxt-RR-JA.

Radiologist	Average length (words)
1	113.1
2	197.2
3	203.1
4	167.1
5	117.1
6	135.1
7	83.7
8	141.9
9	127.1
Mean±SD	142.8±37.3

Table 4– Total of positive, suspicious, and negative findings mentioned by nine radiologists diagnosing the same 15 lung cancers in MedTxt-RR-JA.

	Total findings		
Radiologist	Positive	Suspicious	Negative
1	89	61	28
2	127	32	59
3	155	75	23
4	100	48	37
5	77	31	42
6	91	48	37
7	106	9	10
8	103	38	38
9	105	27	23
Mean±SD	105.9±23.0	41.0±19.6	33.0±14.0

Table 5– Self-BLEU of nine radiology reports for each lung cancer case in MedTxt-RR-JA.

Case No.	Self-BLEU	Case No.	Self-BLEU
1	0.3630	9	0.4094
2	0.3547	10	0.4424
3	0.3683	11	0.4105
4	0.3915	12	0.3896
5	0.3762	13	0.3895
6	0.3932	14	0.3422
7	0.3753	15	0.4209
8	0.3923	Mean±SD	0.3879±0.025

Sentence-level to word-level diversity

Figure 3 shows samples of the first sentence mentioning the primary lesion of one lung cancer case (case No.1) with a translation into English. The samples are rich in paraphrases and indicate sentence-level diversity. Word-level diversity is also observed. For instance, various synonyms referring to the sub-solid nodule are found: “SSN”, “ground-glass opacity”, “GGN”, and “ground-glass nodule.”

Figure 3 – The first sentences of description of the same lung cancer primary lesion by different radiologists.

- 左上葉に径18mm 大のSSN を認めます。(An SSN with a diameter of 18 mm is seen in the left upper lobe.)
- 左肺上葉にやや分葉状を呈する、境界明瞭なすりガラス状陰影を認めます。(A well-marginated ground-glass opacity is observed in the upper lobe of the left lung, which is slightly lobulated.)
- 左肺上葉の外側に主座を置く、約18 mm の境界明瞭な類円形のすりガラス状陰影 (lepidic pattern) を認めます。(A well-defined round ground-glass opacity (lepidic pattern) of approximately 18 mm is seen, located at the periphery of the upper lobe of the left lung.)
- 左肺上葉S1+2 に長径18mm 大の境界明瞭なGGN が認められます。(A well-defined GGN with a diameter of 18 mm is seen in the upper lobe S1 + 2 of the left lung.)
- 左肺上葉S1+2には長径18mm大のすりガラス結節がみられ，充実性成分および気管支拡張を伴っています。(The left upper lung lobe S1 + 2 has a large ground-glass nodule with a length of 18 mm, accompanied by solid components and tracheal dilatation.)

Discussion

We built MedTxt-RR-JA, the first clinical comparable corpus, with 135 radiology reports for the same 15 lung cancer cases diagnosed by nine radiologists. It was suggested that MedTxt-RR-JA contained various synonyms and paraphrases maintaining conformity of diagnosis among radiologists.

Our method for building MedTxt-RR-JA has several advantages. First, collecting multiple radiology reports for the same cases resulted in high comparability. MedTxt-RR-JA is a comparable corpus but is close to a parallel corpus, a superior type of corpus in its comparability [5]. In MedTxt-RR-JA, words, phrases, and sentences in a radiology report have their correlates in another radiology report by another radiologist for the same case, as long as the diagnosis is exactly the same. Such high comparability realizes an observation of lexical diversity by comparing words, phrases, or sentences one by one. A comparable corpus may also be obtained by simply retrieving a bulk of radiology reports in a PACS database in a hospital, where all radiological studies are diagnosed only once. However, our methodology ensures higher comparability by preparing multiple radiology reports for the same case. The high comparability of MedTxt-RR-JA would contribute to the development pf AI systems that can correctly recognize of synonyms or paraphrases. This ability plays an important role in various tasks such as query-based case retrieval and

automated conversion of free-text radiology reports into structured reports [13].

Another advantage of MedTxt-RR-JA is that it provides multiple gold standards to the same image, which is common in non-biomedical AI image captioning datasets [14]. All existing radiology report datasets, such as IU X-Ray [15], MIMIC-CXR [16], and PadChest [17], have one radiology report per image because they are sampled from a hospital database. It is valuable to highly appreciate automated radiology reporting AI systems capable of diverse outputs as they can perform well on unseen images [18]. Providing multiple gold standards to automated radiology reporting AI systems would reinforce the ability to output diverse passages by allowing multiple alternatives to the same image.

Furthermore, MedTxt-RR-JA covers a wide range of imaging findings involved in lung cancer staging. This feature can also be valuable in developing automated lung cancer staging systems using information in unstructured free-text radiology reports [19].

One limitation of this study is that a slight unconformity of diagnosis among radiologists was inevitable because we did not directly instruct radiologists to write down specific findings. We could have provided a set of positive and negative findings to include and requested the radiologists to follow them. Instead, we decided on an indirect strategy to carefully review each case and to add clinical information that would help to narrow down the differential diagnoses. We prioritized having the radiologists write radiology reports in a natural and relaxed manner. Table 2 suggests that our preparation was successful in limiting diagnostic disagreement to an acceptable level.

Another limitation is the considerably small size of MedTxt-RR-JA compared to existing radiology report datasets. This may limit the contribution of MedTxt-RR-JA in this era of data-hungry deep learning methods. We are working on expanding MedTxt-RR-JA by extending target diseases and by recruiting more radiologists.

As aforementioned, MedTxt-RR-JA has the potential for wide applications including case retrieval, automated free-text or structured radiology reporting, and cancer auto-staging. Our future work will involve the development of various medical AI systems for Japanese radiology reports using MedTxt-RR-JA.

Conclusions

We proposed MedTxt-RR-JA, which is the first clinical comparable corpus and the first publicly available Japanese radiology report corpus. We believe that MedTxt-RR-JA will considerably stimulate future research on AI systems that can better understand, compare, search, and generate clinical documents.

Acknowledgements

We are grateful to have received much help from Y's Reading, Inc. to build MedTxt-RR-JA.

References

[1] S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, and H. Xu, Deep learning in clinical natural language processing: a methodical review, *Journal of the American Medical Informatics Association*. **27** (2020) 457–470. doi:10.1093/jamia/ocz200.

[2] K.J. Cios, and G. William Moore, Uniqueness of medical data mining, *Artif. Intell. Med.* **26** (9/2002) 1–24.

[3] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Res.* **32** (2004) D267–70.

[4] A. Romanov, and C. Shivade, Lessons from Natural Language Inference in the Clinical Domain, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. (2018). doi:10.18653/v1/d18-1187.

[5] G.M. Anderman, and M. Rogers, Incorporating Corpora: The Linguist and the Translator, Multilingual Matters, 2008.

[6] M.K. Gospodarowicz, J.D. Brierley, and C. Wittekind, TNM Classification of Malignant Tumours, John Wiley & Sons, 2017.

[7] S. Yada, A. Joh, R. Tanaka, F. Cheng, E. Aramaki, and S. Kurohashi, Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge: Starting From Critical Lung Diseases, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020: pp. 4565–4572.

[8] K. Krippendorff, Reliability in Content Analysis, *Human Communication Research*. **30** (2004) 411–433. doi:10.1111/j.1468-2958.2004.tb00738.x.

[9] T. Kudo, MeCab: Yet Another Japanese Dependency Structure Analyzer, https://taku910.github.io/mecab/format.html (accessed March 26, 2021).

[10] Manbyo Dictionary, https://sociocom.naist.jp/manbyo-dic-en/ (accessed March 26, 2021).

[11] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu, Texygen: A benchmarking platform for text generation models, in: 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, Association for Computing Machinery, Inc, 2018: pp. 1097–1100.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, BLEU, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. (2001). doi:10.3115/1073083.1073135.

[13] E. Pons, L.M.M. Braun, M.G.M. Hunink, and J.A. Kors, Natural Language Processing in Radiology: A Systematic Review, Radiology. 279 (05/2016) 329–343.

[14] M.D.Z. Hossain, F. Sohel, M.F. Shiratuddin, and H. Laga, A Comprehensive Survey of Deep Learning for Image Captioning, ACM Computing Surveys. 51 (2019) 1–36. doi:10.1145/3295748.

[15] D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani, G.R. Thoma, and C.J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, J. Am. Med. Inform. Assoc. (2016). doi:10.1093/jamia/ocv080.

[16] A.E.W. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, C.-Y. Deng, R.G. Mark, and S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci Data. 6 (2019) 317.

[17] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, PadChest: A large chest x-ray image dataset with multi-label annotated reports, Med. Image Anal. 66 (2020) 101797.

[18] S. Venugopalan, L.A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, Captioning Images with Diverse Objects, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017). doi:10.1109/cvpr.2017.130.

[19] E.K. Gupta, R. Thammasudjarit, and A. Thakkinstian, NLP automation to read radiological reports to detect the stage of cancer among lung cancer patients, in: Workshop on Widening NLP, 2019. http://www.winlp.org/wp-content/uploads/2019/final_papers/189_Paper.pdf.

Address for correspondence

If you have any questions, please contact Yuta Nakamura, the corresponding author (yutanakamura-tky@umin.ac.jp).