Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task

Shuntaro Yada
Nara Institute of Science and Technology
Japan
s-yada@is.naist.jp

Yuta Nakamura
The University of Tokyo
Japan
yutanakamura-tky@umin.ac.jp

Shoko Wakamiya
Nara Institute of Science and Technology
Japan
wakamiya@is.naist.jp

Eiji Aramaki
Nara Institute of Science and Technology
Japan
aramaki@is.naist.jp

ABSTRACT

A standard dataset collection is essential for the development of information science. Particularly in the medical field, in which privacy protection is a critical issue, the importance of the dataset is significant. To discuss the validness of various methods, we build the clinical text dataset, Real-MedNLP, for multiple medical tasks. The goal of Real-MedNLP is threefold: (1) Real datasets: Previous medical shared tasks, MedNLP, MedNLP2, and MedNLPDoc, were based on the pseudo dataset, which was built from medical textbooks or dummy clinical texts. This task prepares real radiology and case reports. (2) Bilingual capability: Both English and Japanese data are handled. (3) Practicality: Both fundamental (named entity recognition) and applied practical tasks are handled. This study introduces the task setting of Real-MedNLP and submitted systems. The methods mostly share the common paradigm, which is based on a fundamental language model, such as BERT, aiming to separate the resource problems. Based on their results, this study discusses the feasibility of their approaches to bring us the future direction of medical NLP. Note that the Real-MedNLP is a shared task that handles real Japanese medical texts.

KEYWORDS

Medical Natural Language Processing, Named Entity Recognition, Case Reports, Radiography Reports, Adverse Drug Event, Case Identification

SUBTASKS

Subtask1-CR-EN, Subtask1-CR-JA, Subtask1-RR-EN, Subtask1-RR-JA, Subtask2-CR-EN, Subtask2-CR-JA, Subtask2-RR-EN, Subtask2-RR-JA, Subtask3-CR-EN (ADE), Subtask3-CR-JA (ADE), Subtask3-RR-EN (CI), Subtask3-RR-JA (CI)

1 INTRODUCTION

Recently, an increasing number of medical records have been written in electronic format instead of paper, which leads to a higher importance of natural language processing (NLP) techniques in medical fields [2]. Because NLP is a hot topic in computer science, the number of medical NLP studies is increasing dramatically each year. Despite the large number of studies, the amount of privacy-free medical text data is still small in non-English languages such as Japanese and Chinese.

To resolve this no-resource situation, we conducted a series of four previous medical natural language processing (MedNLP) tasks: MedNLP-1 [13], MedNLP-2 [11], MedNLPDoc [12], and MedWeb [17]. In MedNLP-1, an NTCIR-10 pilot task, we proposed a fundamental task, named entity recognition (NER), using dummy medical records created by medical doctors. In MedNLP-2, as an NTCIR-11 core task, we performed a term normalization task using dummy medical records created by medical doctors. In MedNLP-Doc, one of the NTCIR-12 core tasks, we designed a complete task starting from a medical record from a medical textbook to provide a proper disease name represented by the ICD code. In MedWeb, as an NTCIR-13 core task, a disease tweet classification task was designed to simulate the use of social media data in medical and healthcare domains, and dummy Twitter data were created in Japanese and translated into English and Chinese. The MedNLP task successfully produced valuable datasets. However, two problems were identified. (1) The data were not real clinical texts, but were dummy records or sample texts from medical textbooks. (2) The dataset was limited to Japanese, which makes it difficult to compare the results with other English-based shared task results[1].

The pilot task, called **real-MedNLP**, first introduces real clinical text. Our data consisted of two core resources: (1) Case Report corpus (**MedTxt-CR**) and (2) the Radiology Report corpus (**MedTxt-RR**).

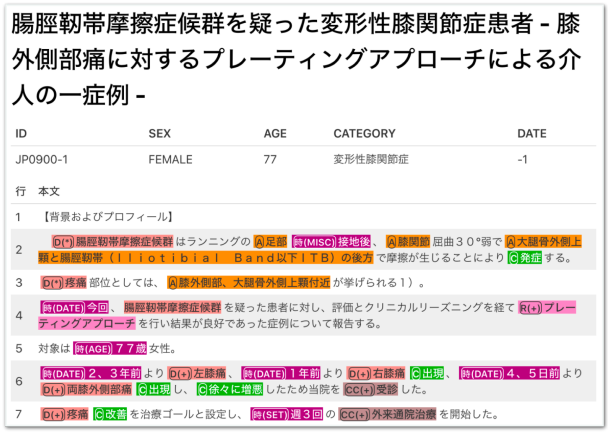
In addition, we prepared real data in Japanese and translated the original reports into English, enabling us to develop the first benchmark for multilanguage medical NLP.

Considering the data, we also redesigned the task scheme for our ultimate goal, which promotes systems applicable to hospitals. The challenges of this task are twofold.

Few-resource Named Entity Recognition (Subtasks 1 & 2) : Participants extract important information from real medical texts. This challenge is categorized into two tasks: exactly 100 training (Subtask 1) and guideline learning (Subtask 2). The task is classified according to the amount of training data (small or no data).

Applications (Subtask 3) : This challenge was designed from a practical viewpoint. For case reports, we designed an information extraction task for adverse drug events (ADE). The ADE task has been challenged through workshops (e.g.,

[1] https://n2c2.dbmi.hms.harvard.edu/2022-challenge



腸脛靭帯摩擦症候群を疑った変形性膝関節症患者 - 膝外側部痛に対するプレーティングアプローチによる介入の一症例 -

ID	SEX	AGE	CATEGORY	DATE
JP0900-1	FEMALE	77	変形性膝関節症	-1

行	本文
1	【背景およびプロフィール】
2	D(+)腸脛靭帯摩擦症候群 はランニングの A足部 時(MISC)接地後、 A膝関節 屈曲３０°弱で A大腿骨外側上顆と腸脛靭帯（ＩＴＢ）の後方 で摩擦が生じることにより C発症 する。
3	D(+)疼痛 部位としては、 A膝外側部、大腿骨外側上顆付近 が挙げられる１）。
4	時(DATE)今回、 腸脛靭帯摩擦症候群 を疑った患者に対し、評価とクリニカルリーズニングを経て R(+)プレーティングアプローチ を行い結果が良好であった症例について報告する。
5	対象は 時(AGE)７７歳 女性。
6	時(DATE)２、３年前 より D(+)左膝痛、 時(DATE)１年前 より D(+)右膝痛 C出現、 時(DATE)４、５日前 より D(+)両膝外側部痛 C出現 し、 C徐々に増悪 したため当院を CC(+)受診 した。
7	D(+)疼痛 C改善 を治療ゴールと設定し、 時(SET)週３回 の CC(+)外来通院治療 を開始した。

Figure 1: A sample case report corpus in Japanese (MedTxt-CR-JA)

n2c2, 2009). For radiographic reports, we designed a case identification (CI) task to detect reports originating from the same patient. This task was proposed for the unique corpus.

These challenges yield promising technologies for the development of practical systems to support a wide range of medical services.

2 MATERIALS

Two corpora covering multiple languages were used (see Table 11). Note that this paper notates the corpus as a combination of the name and its language, such as MedTxt-CR-JA.

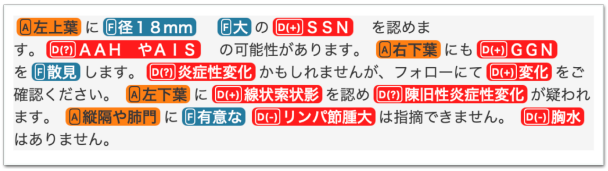
2.1 Corpus

Case report. A case report is a medical research paper written for the patient. Case report analysis has two potential advantages: the case report covers most of the disease timeline or a history of the target disease, and the number of case reports is greater than the other papers because each medical society usually has a submission truck for case reports. Considering these advantages, case reports could be a rich source of information. The format of a case report is similar to that of a discharge summary, which is a type of medical report. Therefore, techniques for case report analysis can be expanded to analyze discharge summaries.

MedTxt-CR-JA comprises approximately 200 open-access case reports available at CiNii[2] (in Japanese). Figure 1 shows its sample. Because the number of medical societies that produce open-access publications is limited, the types of patients and diseases reported in open-access case reports are highly biased. To reduce the bias caused by the publication policy of each medical society, we selected approximately 200 case reports based on the actual frequencies of patients and diseases. Approximately 200 reports from the J-Stage article database have been translated from Japanese (MedTxt-CR-JA) to English (MedTxt-CR-EN).

Radiology Report. A radiology report is a type of clinical document (also called a *report* in this paper) written by a radiologist. Each radiology report discusses a single radiological examination such as an

[2]https://ci.nii.ac.jp/



A左上葉 に F径１８mm F大 の D(+)ＳＳＮ を認めます。 D(?)ＡＡＨ やＡＩＳ の可能性があります。 A右下葉 にも D(+)ＧＧＮ を F散見 します。 D(?)炎症性変化 かもしれませんが、フォローにて D(+)変化 をご確認ください。 A左下葉 に D(+)線状索状影 を認め D(?)陳旧性炎症性変化 が疑われます。 A縦隔や肺門 に F有意な D(-)リンパ節腫大 は指摘できません。 D(-)胸水 はありません。

Figure 2: Sample of the radiology report corpus in Japanese (MedTxt-RR-JA)

X-ray, CT, or MRI scan. A radiology report contains (i) descriptions of all normal and abnormal findings and (ii) interpretations of the findings, including disease diagnosis and recommendations for the next clinical test or treatment. Although most of the radiology AI research tends to focus only on images because image-based AI draws much attention, NLP on radiology reports also has potential for a wide variety of clinical application [14].

MedTxt-RR was created by Nakamura et al. and composed of 135 radiology reports. Figure 2 shows its sample. MedTxt-RR aims to provide information about the diversity of expressions used by different radiologists to describe the same diagnosis. One of the difficulties in analyzing radiology reports is the variety in writing styles, but simply collecting radiology reports from medical institutions cannot provide enough information, because only one report is issued for a single radiological examination in usual clinical practice. MedTxt-RR-JA was created to overcome this problem by crowdsourcing, for which nine doctors independently wrote radiology reports for the same series of fifteen lung cancer cases. A total of 135 radiology reports are available in MedTxt-RR-JA. MedTxt-RR-EN is an English translation of MedTxt-RR-JA by nine translators in which different translators translated radiology reports by different radiologists.

2.2 Subtasks and Annotations

Few-resource NER challenge for MedTxt-CR and for MedTxt-RR (Subtask1/2-CR/RR)

Because NER is the most fundamental information extraction for MedNLP, we designed challenges regarding NER for our real clinical reports, which have only 100-200 reports. This corpus size scale tends to be regarded as “few-resource machine learning,” which is the de facto standard among any kind of MedNLP in general.

1) **Just 100 Training (Subtask1)** Use the training split of 100 reports to build a model. This corresponds to the standard type of few-resource supervised learning.
2) **Guideline Learning (Subtask2)** For each tag, we give only a handful of sentence examples. This simulates the training of human annotators, who often learn from annotation guidelines provided by researchers.

Task challengers can use any other resources outside this project if they find them useful for their methods.

We adopted the following types of medical entities [19]:

- **Diseases and symptoms** <d>
- **Anatomical entities** <a>
- **Features and measurements** <f>
- **Change** <c>

- **Time** <timex3>
- **Test** <t-test/key/val>
- **Medicine** <m-key/val>
- **Remedy** <r>
- **Clinical Context** <cc>

Detailed definitions and information on modality are provided in [19]. The two corpora share the same NER tag set. The detailed corpus statistics are presented in the Appendix. We evaluated the following tag sets: <a>, <d>, <t-test>, <timex3>, <m-key>, <m-val>, <t-key>, and <t-val> for MedTxt-CR, and <a, <d>, <t-test>, and <timex3> for MedTxt-RR, because the others are rare in the corpora.

ADE challenge for MedTxt-CR (Subtask3-CR)

This subtask was specifically designed for MedTxt-CR. Given an input report, the system extracts ADE information from the report. The ADE information consists of two tables: <d>-table for disease and symptom names and <m-key>-table for medication (drug) names as shown in Figure 3. For each entity in these tables, four levels of ADE certainty (ADEval)[3] are to be given:

3 - Definitely
2 - Probably
1 - Unlikely
0 - Unrelated (no ADE)

For disease names, these values are interpreted as the likelihood of *being an ADE*, whereas the likelihood of *causing an ADE* is the interpretation of the medication names. Note that these values are annotated regardless of the relationship between the ADE and its trigger. This is because we could not confidently prove explicit causal relationships among multiple medications and symptoms, although some case reports might explicitly argue some causal effects. Eagerly detecting ADEs, including potential ones, would contribute to the ideal handling of ADEs (i.e., the recall-oriented reporting strategy, which is adopted in some countries, such as Japan and EU nations).

To annotate these labels, we let annotators follow the author's perspective on whether drugs and symptoms are related to ADE or not (i.e., writer's perception). In other words, the annotators respect the description by the author if the author argues that a symptom is likely to be an ADE of a certain drug. However, if the annotators noticed other possibilities of ADEs that were not explicitly pointed out in the report, we allowed them to label ADEval ≥ 1 as well (i.e., reader's perception). In a real-world situation, even when the author of a report does not state a phenomenon as an ADE, readers of the report may find that it is actually an ADE if they can infer such possibilities.

CI challenge for MedTxt-RR (Subtask3-RR)

The case identification (CI) challenge is an application task designed for MedTxt-RR. In the CI challenge, a large number of radiology reports are provided, and participants are required to find and

[3]We adapted the four levels in ADE reporting proposed in [8], which was based on the FDA's toxicity grading scale guidance from vaccine trials (September 2007, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/toxicity-grading-scale-healthy-adult-and-adolescent-volunteers-enrolled-preventive-vaccine-clinical). We modified the original *possibly* level to *unlikely* by considering NIA adverse event and serious adverse event guidelines (https://www.nia.nih.gov/sites/default/files/2018-09/nia-ae-and-sae-guidelines-2018.pdf).

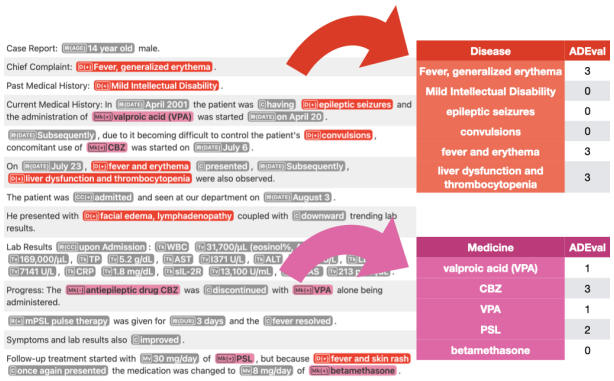


Figure 3: ADE challenge

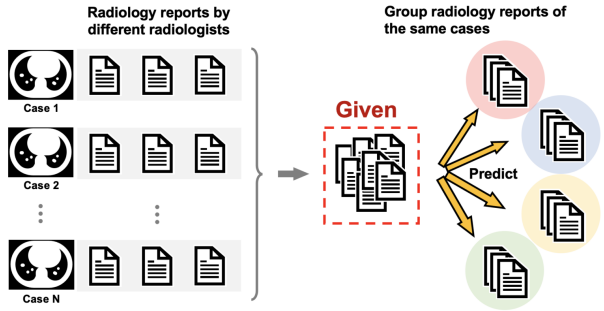


Figure 4: CI challenge

group the radiology reports to diagnose the same image (Figure 4). MedTxt-RR was originally created by collecting radiology reports from multiple radiologists who independently diagnosed the same CT images, and the correspondence between the radiology reports and CT images was used as the gold standard label. Thus, there were no additional annotations for the CI challenge.

A part of MedTxt-RR-JA was already publicly available as a sample before launching Real-MedNLP. Therefore, in the CI challenge, this sample was used as a training set because anyone can look at the gold standard for the correspondence of the radiology report and the CT image. Some of the rest of the radiology reports were used as a test set.

Participants in the CI challenge were asked to submit a prediction as a CSV file consisting of two columns, "id" and "case". The column "id" contains a unique number assigned to each radiology report in advance. Predictions were made by filling the "case" column with numerical values so that only the rows for the radiology reports diagnosing the same CT image had the same numbers. During the task evaluation, only the correspondence between the values in the "case" column and the CT image was evaluated, and the values were ignored. Any numerical value can be stored in the "case" column provided the correspondence is correct.

In real-world clinical institutions, only one reading report is created for each CT image, and a situation in which multiple reading reports are created for the same image does not occur. Therefore, the CI challenge is not a task that directly solves a problem that

Table 1: Number of systems developed by each team

	A	C	D	E	F	G	H	I	J	Total
Subtask1-CR-JA	2			1	4	1			4	12
Subtask1-CR-EN		2			4		5		4	15
Subtask1-RR-JA	2			1		1			4	8
Subtask1-RR-EN							3		4	7
Subtask2-CR-JA	1			1						2
Subtask2-CR-EN										0
Subtask2-RR-JA	1			1						2
Subtask2-RR-EN								1		1
Subtask3-CR-JA (ADE)				1	2					3
Subtask3-CR-EN (ADE)		10			2		6	1		19
Subtask3-RR-JA (CI)			1	1	1				1	4
Subtask3-RR-EN (CI)		10			1			1		12

occurs in a clinical institution. The CI challenge, however, can help AI systems to understand the clinical content of documents accurately without being confused by synonyms or paraphrases, as MedTxt-RR contains radiology reports with almost the same clinical content but various expressions.

3 METHODS

This section briefly introduces each team and the approach to each system. For more information, refer to the system papers for NTCIR-16 Real-MedNLP.

Teams and Systems. In total, nine teams formally submitted their results, which were anonymized by one capital letter. Distinct systems proposed by a team 'X', for example, are denoted in combination with numbers, such as 'X1' and 'X2'. Table 1 lists the number of systems submitted by each team.

- **A** *Subtask1-CR-JA, Subtask1-RR-JA, Subtask2-CR-JA, Subtask2-RR-JA* ; They used Japanese Medical Domain Specific BERT (UTH BERT) [18]. For Subtask 2, Dictionary Manbyo dictionary [7], Hyakuyaku[4], and comeJisyo[5] + bootstrap method are used.
- **B** *Subtask1-CR-JA* ; However, this team finally withdrew, so this study did not include this team's results.
- **C** *Subtask1-CR-EN, Subtask3-CR-EN (ADE), Subtask3-RR-EN (CI)* ; For the NER challenge, they used BioBERT and data augmentation (label-wise token replacement, synonym replacement, mention replacement, and shuffle within segments).
  For the CI challenge, the core technology is as follows: (1) Key feature clustering and (2) document embedding using sentence BERT [15] and K-means clustering.
  For the ADE challenge, Vocabulary adapted the BERT model (VART) with a multi-learning mechanism, where we transformed the ADE challenge into a classification task.
- **D** *Subtask3-RR-JA (CI)* ; They performed parsing-based named entity recognition without a dictionary and calculated case similarity using the distances of the bag-of-entity vectors.
- **E** *Subtask1-CR-JA, Subtask1-RR-JA, Subtask2-CR-JA, Subtask2-RR-JA, Subtask3-CR-JA (ADE), Subtask3-RR-JA (CI)* ; For NER challenges (Subtasks 1 and 2), the systems of this team are the baseline systems for reference to other teams. This approach is based on a simple method without any special techniques. The model was based on general BERT [4]. For Subtask 2, they trained the NER model only from sentences that were available in the annotation guidelines.
- **F** *Subtask1-CR-EN, Subtask1-CR-JA, Subtask3-CR-EN (ADE), Subtask3-CR-JA (ADE), Subtask3-RR-EN (CI), Subtask3-RR-JA (CI)* ; For Subtask 1-CR-EN and JA, they employed two types of close-multilingual approaches: multilingual BERT (mBERT)[6] and XLM-RoBERTa (XLM-R)[7] to compare the effectiveness of the multilingual pre-trained models. Subtask 3 (ADE) regarded ADEval as an additional attribute to named entities and solved the challenge as NER using mBERT and XLM-R.
- **G** *Subtask1-CR-JA, Subtask1-RR-JA* ; BERT [4] + data augmentation (synonym replacement and shuffling within segments).
- **H** *Subtask1-CR-EN, Subtask1-RR-EN, Subtask3-CR-EN (ADE)* ; For Subtask 1, BERT [4], BioBERT [9], clinical BERT [1], PubMed BERT [6], and entity BERT [10]. They utilized span-based NER + data augmentation (back translation (via the Chinese language) and random feature dropout).
  For Subtask 3 (ADE), their approach is a combination of multiclass classification and prompt learning. In addition, they attempted ensembles and data augmentation.
- **I** *Subtask1-RR-EN, Subtask2-RR-JA, Subtask3-CR-EN (ADE), Subtask3-RR-EN (CI)* ; They proposed a pipeline approach for multiple NLP modules (MetaMap [3], BERT [4], ScispaCy[8], and MedLinker (NE linker)).
- **J** *Subtask1-CR-JA, Subtask1-RR-JA, Subtask1-CR-EN, Subtask1-RR-EN, Subtask3-RR-JA (CI)* ; They applied both languages (JA and EN), whose methodologies are completely different for each language. Subtask 1 in JA is based on general BERT and data augmentation, plus ensemble. For Subtask 3 in JA, they used TNM classification estimation[9], which is an international standard cancer staging system. Note that they did not rely on the NE tags.
  For English NER, they utilized domain-specific BERTs (BioBERT [4] and ClinicalBERT [1]), and RoBERTa (general domain) [20].

4 EVALUATION METRICS

4.1 Few-resource NER challenge for MedTxt-CR

We employed a set of standard NER metrics (F-measure) and their variations designed by considering the following three factors:

Joint factor : We use three different levels of NE matching.

(1) span (only recognition) : It considers matching only the recognized span (this level of the NE category). This level disregards the NE labels.

(2) +label (a span and a label joint) : It considers the matching of both the recognized span and its NE category. This is the standard NER evaluation method utilized in most NE-shared tasks (such as CoNLL2003).

[4] https://sociocom.naist.jp/hyakuyaku-dic/
[5] https://ja.osdn.net/projects/comedic/
[6] https://github.com/google-research/bert/blob/master/multilingual.md
[7] https://huggingface.co/xlm-roberta-base
[8] https://allenai.github.io/scispacy/
[9] https://www.uicc.org/resources/tnm

(3) +label+mod (a span and a label+mod joint) : It considers the matching of the recognized span, its NE category and attribute, that is, "scheduled," "executed," "negated," or "other" for `t-test`.

Note that the scores satisfy the following relationship: **span** > **span+label joint** > **span+label+mod joint**).

Matching policy factor (exact/partial) : Usually, NE performance is based on the exact span match (exact match). We consider not only the exact match but also a partial match. If the system-predicted entities partially match (overlap) the corresponding gold-standard annotations, they are regarded as "partially correct". This is because our corpora included many complex compound medical terms and phrase-level entities. For downstream information extraction, partially identified entities would still be helpful. To calculate the partial match score, we considered the proportion of common sub-characters between the gold standards and predicted entities. We regarded the number of common sub-characters divided by the character-length of the predicted entity as the entity-level partial match precision, whereas we regarded the number of common sub-characters divided by the character-length of the gold standard entity as the entity-level partial match recall. Then we obtained the system-level partial match precision as the division of the sum of entity-level partial match precision values for all predicted entities by the number of all predicted entities; the system-level partial match recall is calculated in the same manner.

Frequency factor : We assumed that the rare NEs were difficult to recognize. In particular, Subtask 2 guideline learning contains many rare NEs. To focus on the performance of such rare NEs, we designed a novel f-measure that is weighted by the frequencies of the NE The idea is to penalize the correct guesses of the system if the predicted entity appears frequently in the training dataset. For each gold-standard entity i in the test set, we multiply the entity-level precision and recall scores by the weight w_i based on the term frequency f_i of the same entity in the training set, $w_i = 1/(\log_e(f_i+1)+1)$. The weight w_i gives 1.0 in $f_i = 0$, and it decreases to 0.59 in $f_i = 1$, 0.48 in $f_i = 2$, and so on. This weighting allowed us to observe the extent to which the system relied on high-frequency entities in the training phase.

The following tag sets were evaluated for MedTxt-CR: `<d>`, `<a>`, `<timex3>`, `<t-test>`, `<t-key>`, `<t-val>`, `<m-key>`, and `<m-val>`.

4.2 Few-resource NER challenge for MedTxt-RR

The few-resource NER for MedTxt-RR is the same as that for MedTxt-CR. However, the entity distribution is significantly different from MedTxt-CR, in which several entity categories, such as medication and tests, rarely appear. Therefore, for all tags, the following tag sets were evaluated for MedTxt-RR: `<d>`, `<a>`, `<timex3>`, and `<t-test>`.

4.3 ADE challenge for MedTxt-CR

The ADE challenge is an information extraction task. We employed two levels of evaluation: **entity level** and **report level**.

Entity level for each entity, Precision, Recall, and F1-score of each ADEval (= 0, 1, 2, 3) are calculated.

Report level binary classification for each report. For each report, it was judged whether the report contained ADE information (POSITIVE REPORT) or not (NEGATIVE REPORT). We regarded a report that contained at least one entity with ADEval ≥ 1 as a POSITIVE-REPORT; otherwise, it was a NEGATIVE-REPORT. In this binary classification scheme, the report-wise precision, recall, and F1-score of the POSITIVE-REPORT were used in the evaluation.

4.4 CI challenge for MedTxt-RR

Because the CI challenge is a clustering task with gold-standard labels, we used normalized mutual information (NMI).

Let $\widehat{\mathbf{y}} = (\widehat{y_1}, ..., \widehat{y_N})$ and $\mathbf{y} = (y_1, ..., y_N)$ be the prediction and gold labels for N radiology reports in the test set. First, we collect unique labels in $\widehat{\mathbf{y}}$ and $\mathbf{y}$ to compose two sets, $\widehat{U}$ and U:

$$\widehat{U} = \{\widehat{u_1}, ..., \widehat{u_{|\widehat{U}|}} \mid \widehat{u_i} \in \widehat{\mathbf{y}},\ i \neq k \iff \widehat{u_i} \neq \widehat{u_k}\}$$
$$U = \{u_1, ..., u_{|U|} \mid u_j \in \mathbf{y},\ j \neq k \iff v_j \neq v_k\}$$

$|U|$ is the true number of test cases and $|\widehat{U}|$ is the number of test cases assumed by the participant (i.e., the number of clusters created by the participant). Because we do not provide the value of $|U|$, $|\widehat{U}|$ may differ among participants. The NMI is then calculated as the mutual information (MI) of $\widehat{U}$ and U normalized by the arithmetic average of the entropy of $\widehat{U}$ and U:

$$|\widehat{U}_i| = |\{k \mid 1 \leq k \leq N,\ \widehat{y_k} = \widehat{u}_i\}|,$$
$$|U_j| = |\{k \mid 1 \leq k \leq N,\ y_k = u_j\}|,$$
$$|\widehat{U}_i \cap U_j| = |\{k \mid 1 \leq k \leq N,\ \widehat{y_k} = \widehat{u}_i,\ y_k = u_j\}|,$$
$$H(\widehat{U}) = -\sum_{i=1}^{|\widehat{U}|} \frac{|\widehat{U}_i|}{N} \log \frac{|\widehat{U}_i|}{N},\ H(U) = -\sum_{j=1}^{|U|} \frac{|U_j|}{N} \log \frac{|U_j|}{N},$$
$$\mathrm{MI}(\widehat{U}, U) = \sum_{i=1}^{|\widehat{U}|} \sum_{j=1}^{|U|} \frac{|\widehat{U}_i \cap U_j|}{N} \log \frac{N|\widehat{U}_i \cap U_j|}{|\widehat{U}_i||U_j|},$$
$$\mathrm{NMI}(\widehat{U}, U) = \frac{2\mathrm{MI}(\widehat{U}, U)}{H(\widehat{U}) + H(U)}$$

5 RESULTS AND DISCUSSIONS

5.1 Subtasks 1 and 2

Tables 2, 3, 4, and 5 show results of Subtasks 1 and 2 for MedTxt-CR and MedTxt-RR, respectively.

CR V.S. RR. Overall, the systems for the RR corpus performed better than those for the CR corpus. This result follows the fact that radiology reports mainly consist of frequent patterns and case reports tend to be linguistically diverse [19]. In contrast, the CR corpus has a large vocabulary that covers most medical fields.

JA V.S. EN:. The performances of two languages (JA and EN) are almost the same. In the CR track (exact; +label; normal), the best JA and EN systems were 0.6525 (E1) and 0.6337 (H2), respectively

Table 2: Results of Subtask 1 for MedTxt-CR. Bold font indicates the best score for each evaluation metric.

(a) Results of Subtask1-CR-JA.

	Exact match						Partial match					
	span		+label		+label+mod		span		+label		+label+mod	
System ID	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted
A1	0.6388	0.5433	0.6133	0.5195	-	-	0.8524	0.7462	0.7841	0.6812	-	-
A2	0.6378	0.5425	0.6124	0.5188	-	-	0.8530	0.7469	**0.7846**	**0.6819**	-	-
E1	**0.6988**	**0.5995**	**0.6525**	**0.5550**	**0.5921**	**0.4993**	**0.8617**	**0.7540**	0.7727	0.6689	**0.6977**	**0.5993**
F1	0.6095	0.5112	0.5696	0.4737	0.5249	0.4333	0.8297	0.7212	0.7267	0.6230	0.6552	0.5574
F2	0.6497	0.5445	0.6076	0.5048	0.5602	0.4621	0.8127	0.6992	0.7257	0.6164	0.6596	0.5562
F3	0.5897	0.4987	0.5550	0.4650	0.5171	0.4315	0.8422	0.7364	0.7522	0.6489	0.6828	0.5850
F4	0.6179	0.5218	0.5813	0.4863	0.5420	0.4515	0.8221	0.7115	0.7464	0.6381	0.6821	0.5796
G1	0.6766	0.5754	0.6189	0.5198	-	-	0.8386	0.7281	0.7361	0.6293	-	-
J1	0.3361	0.2627	0.3088	0.2383	0.2591	0.1963	0.7462	0.6572	0.5514	0.4712	0.4488	0.3777
J2	0.3676	0.3057	0.3585	0.2968	0.3013	0.2459	0.7149	0.6363	0.6395	0.5633	0.5307	0.4625
J3	0.2745	0.2279	0.2656	0.2195	0.2247	0.1836	0.6856	0.6159	0.5865	0.5204	0.4820	0.4232
J4	0.2841	0.2399	0.2773	0.2334	0.2308	0.1910	0.6704	0.6062	0.5967	0.5346	0.4963	0.4403

(b) Results of Subtask1-CR-EN.

	Exact match						Partial match					
	span		+label		+label+mod		span		+label		+label+mod	
System ID	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted
C1	0.4601	0.4117	0.4321	0.3850	-	-	0.6357	0.5802	0.5648	0.5124	-	-
C2	0.4697	0.4198	0.4371	0.3890	-	-	0.6401	0.5831	0.5655	0.5122	-	-
F1	0.5104	0.4501	0.4683	0.4092	0.4245	0.3701	0.8094	0.7349	0.6999	0.6280	0.6242	0.5583
F2	0.5292	0.4667	0.4860	0.4247	0.4406	0.3843	0.8002	0.7239	0.6990	0.6252	0.6295	0.5616
F3	0.5240	0.4634	0.4918	0.4326	0.4480	0.3938	0.8128	0.7390	0.7239	0.6528	0.6486	0.5840
F4	0.5473	0.4839	0.5145	0.4525	0.4696	0.4127	0.7936	0.7171	0.7142	0.6404	0.6481	0.5808
H1	0.6246	0.5513	0.5980	0.5255	0.5484	0.4809	0.7938	0.7081	0.7372	0.6535	0.6769	0.5994
H2	**0.6540**	**0.5813**	**0.6337**	**0.5616**	**0.5853**	**0.5181**	0.8389	0.7533	**0.7880**	**0.7042**	**0.7269**	**0.6488**
H3	0.6438	0.5719	0.6231	0.5515	0.5749	0.5080	0.8300	0.7438	0.7790	0.6947	0.7181	0.6394
H4	0.6190	0.5516	0.5933	0.5265	0.5452	0.4831	0.8423	0.7623	0.7784	0.7005	0.7156	0.6435
H5	0.6299	0.5620	0.6033	0.5364	0.5540	0.4917	**0.8453**	**0.7637**	0.7825	0.7034	0.7180	0.6444
J1	0.4882	0.4274	0.4556	0.3965	0.2957	0.2589	0.7992	0.7234	0.7032	0.6303	0.4479	0.4005
J2	0.5551	0.4925	0.5197	0.4589	0.3335	0.2950	0.8308	0.7535	0.7376	0.6638	0.4711	0.4228
J3	0.5503	0.4846	0.5116	0.4478	0.3263	0.2867	0.8261	0.7456	0.7220	0.6453	0.4609	0.4111
J4	0.5270	0.4652	0.4918	0.4317	0.3077	0.2705	0.8159	0.7390	0.7191	0.6455	0.4526	0.4046

(Table 2). In the RR track (exact; +label; normal), the best JA system was 0.8926 (A2) and the best EN system was 0.8266 (H2) (Table 3). Although slightly better results are shown in the JA systems, the difficulties of this task are not language-independent.

Span V.S. Joint: Label-joint NER is more difficult than span recognition. For example, in the CR track (exact; normal), we can see a 10 point difference between span = 0.6988, +label = 0.6525, and +label+mod = 0.5921 (see Table 2). Between Span and Joint, the performance difference is approximately 10 point gaps.

Partial V.S. Exact: The partial scores were at least approximately 10 points larger than the exact scores in general, regardless of the corpus (track), language, and subtask. In particular, the span-only scores that disregard the match of entity classes or modality values achieve approximately a 0.8 F1-score in the CR corpus and an approximately 0.9 F1-score in the RR corpus; almost all systems captured medically important phrases at least partially, even though the training data were very small.

Zero/few-shot: The general effect of zero/few-shot weighting was to decrease scores. However, the performance was robust, even after weighting. The amount of decrease in weighting was larger in the RR corpus, which means that the RR-solving systems relied substantially on high-frequency entities in the training dataset. This would be due to the fact that frequently used patterns often constitute radiology reports. The decrease in the weighting was smaller in the EN corpora, even though they are parallel to the JA corpora, probably because identical Japanese entities were sometimes translated into different surface phrases in English according to the local context.

Subtask 1 V.S. Subtask 2: Subtask 1 is a standard NER with a small corpus. Subtask 2 is our new challenge, relying on only a few examples. Because the number of participants was small in Subtask

Table 3: Results of Subtask 1 for MedTxt-RR. Bold font indicates the best score for each evaluation metric.

(a) Results of Subtask1-RR-JA.

	Exact match						Partial match					
	span		+label		+label+mod		span		+label		+label+mod	
System ID	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted
A1	0.1528	0.1185	0.1505	0.1165	-	-	**0.9807**	0.5823	**0.9639**	0.5668	-	-
A2	**0.9019**	**0.5264**	**0.8926**	**0.5181**	-	-	0.9755	**0.5900**	0.9614	**0.5769**	-	-
E1	0.8704	0.5052	0.8488	0.4871	**0.8079**	**0.4674**	0.9603	0.5824	0.9269	0.5536	**0.8778**	**0.5281**
G1	0.8932	0.5207	0.8703	0.4992	-	-	0.9735	0.5889	0.9385	0.5580	-	-
J1	0.5862	0.3232	0.5811	0.3191	0.4259	0.2550	0.8943	0.5563	0.8201	0.4971	0.5727	0.3693
J2	0.6055	0.3306	0.6022	0.3278	0.4363	0.2572	0.9042	0.5623	0.8370	0.5078	0.5894	0.3781
J3	0.5805	0.3151	0.5779	0.3127	0.4224	0.2480	0.8996	0.5633	0.8213	0.5003	0.5857	0.3764
J4	0.5715	0.3120	0.5674	0.3096	0.4216	0.2477	0.8812	0.5530	0.8201	0.5024	0.5884	0.3803

(b) Results of Subtask1-RR-EN.

	Exact match						Partial match					
	span		+label		+label+mod		span		+label		+label+mod	
System ID	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted
H1	0.8296	0.5532	0.8260	0.5496	**0.7919**	**0.5262**	0.9567	0.6670	0.9286	0.6402	**0.8862**	**0.6095**
H2	**0.8302**	**0.5536**	**0.8266**	**0.5500**	0.7874	0.5231	0.9569	0.6674	**0.9293**	**0.6406**	0.8805	0.6059
H3	0.8140	0.5430	0.8061	0.5358	0.7719	0.5105	**0.9588**	0.6726	0.9224	0.6388	0.8787	0.6050
J1	0.7696	0.5049	0.7592	0.4957	0.6350	0.4107	0.9513	0.6719	0.9085	0.6316	0.7410	0.5062
J2	0.8068	0.5360	0.7997	0.5299	0.6707	0.4400	0.9532	0.6707	0.9132	0.6325	0.7551	0.5163
J3	0.7962	0.5265	0.7877	0.5192	0.6532	0.4264	0.9533	0.6703	0.9156	0.6346	0.7469	0.5104
J4	0.8000	0.5332	0.7895	0.5245	0.6545	0.4309	0.9567	**0.6751**	0.9170	0.6381	0.7513	0.5165

Table 4: Results of Subtask 2 for MedTxt-CR (Subtask2-CR-JA). Bold font indicates the best score for each evaluation metric.

	Exact match						Partial match					
	span		+label		+label+mod		span		+label		+label+mod	
System ID	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted
A1	**0.4212**	**0.4146**	**0.3710**	**0.3644**	**0.3710**	**0.3644**	**0.7458**	**0.7379**	**0.6163**	**0.6091**	**0.6163**	**0.6091**
E1	0.3366	0.3326	0.2512	0.2474	0.1949	0.1912	0.6797	0.6738	0.4589	0.4547	0.3464	0.3424

Table 5: Results of Subtask 2 for MedTxt-RR. Bold font indicates the best score for each evaluation metric.

(a) Results of Subtask2-RR-JA.

	Exact match						Partial match					
	span		+label		+label+mod		span		+label		+label+mod	
System ID	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted
A1	**0.6638**	**0.6370**	**0.6485**	**0.6217**	**0.5133**	**0.4958**	**0.9106**	**0.8834**	**0.8843**	**0.8571**	**0.6864**	**0.6685**
E1	0.6557	0.6315	0.6255	0.6013	0.4668	0.4462	0.8961	0.8711	0.8289	0.8039	0.6094	0.5880

(b) Results of Subtask2-RR-EN.

	Exact match						Partial match					
	span		+label		+label+mod		span		+label		+label+mod	
System ID	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted
I1	0.5628	0.5546	0.5496	0.5422	0.5037	0.4968	0.8843	0.8726	0.8289	0.8179	0.7599	0.7495

2, precise discussion was difficult. However, we can say that several systems surprisingly demonstrated high performance, such as 0.6485 in the RR track (exact; +label, normal) for system A1 (Table 5). In addition, system A1 achieved a partial match of 0.8843

Table 6: Results of Subtask 3 for MedTxt-CR. Bold font indicates the best score for each evaluation metric.

(a) Results of Subtask3-CR-JA (ADE).

	ADEval=0			ADEval=1			ADEval=3			Report-level		
System ID	P	R	F	P	R	F	P	R	F	P	R	F
E1	95.21	76.04	84.55	0.00	0.00	0.00	6.98	**52.94**	12.33	12.73	**77.78**	21.88
F1	95.76	**97.67**	**96.71**	0.00	0.00	0.00	12.50	11.76	12.12	**37.50**	66.67	**48.00**
F2	**96.05**	97.00	96.52	0.00	0.00	0.00	**27.59**	47.06	**34.78**	25.00	44.44	32.00

(b) Results of Subtask3-CR-EN (ADE).

	ADEval=0			ADEval=1			ADEval=3			Report-level		
System ID	P	R	F	P	R	F	P	R	F	P	R	F
C1	95.70	94.94	95.32	20.00	5.26	8.33	62.50	26.32	37.04	22.22	66.67	33.33
C2	95.79	97.00	96.39	14.29	5.26	7.69	43.75	36.84	40.00	29.41	55.56	38.46
C3	95.95	93.52	94.72	6.25	5.26	5.71	28.57	21.05	24.24	19.35	66.67	30.00
C4	96.05	92.10	94.03	25.00	5.26	8.70	22.22	42.11	29.09	18.92	77.78	30.43
C5	95.87	95.26	95.56	0.00	0.00	0.00	56.25	47.37	51.43	25.93	77.78	38.89
C6	96.14	94.47	95.30	25.00	10.53	14.81	50.00	21.05	29.63	21.21	77.78	33.33
C7	95.67	94.31	94.99	0.00	0.00	0.00	33.33	26.32	29.41	19.35	66.67	30.00
C8	96.42	97.79	97.10	20.00	5.26	8.33	47.62	52.63	50.00	50.00	77.78	60.87
C9	96.35	91.79	94.01	0.00	0.00	0.00	23.81	52.63	32.79	18.92	77.78	30.43
C10	95.87	95.26	95.56	7.14	5.26	6.06	26.92	36.84	31.11	23.08	66.67	34.29
F1	96.53	96.68	96.61	0.00	**96.68**	0.00	31.25	52.63	39.22	25.00	55.56	34.48
F2	95.39	98.10	96.73	0.00	0.00	0.00	40.00	42.11	41.03	40.00	44.44	42.11
H1	96.57	97.95	97.25	14.29	5.26	7.69	60.00	63.16	61.54	50.00	66.67	57.14
H2	96.57	97.95	97.25	0.00	0.00	0.00	59.09	**68.42**	63.41	50.00	66.67	57.14
H3	96.28	98.10	97.18	0.00	0.00	0.00	60.00	63.16	61.54	50.00	55.56	52.63
H4	96.41	97.63	97.02	0.00	0.00	0.00	57.14	63.16	60.00	50.00	66.67	57.14
H5	95.88	**99.37**	**97.60**	0.00	0.00	0.00	78.57	57.89	**66.67**	**60.00**	33.33	42.86
H6	95.99	98.26	97.11	**33.33**	5.26	9.09	55.56	52.63	54.05	50.00	44.44	47.06
I1	**97.02**	97.63	97.32	30.00	31.58	**30.77**	**100.00**	26.32	41.67	50.00	**88.89**	**64.00**

(partial; +label, normal), which is good enough for several practical applications, such as case report search. This challenge reveals the potential feasibility of NER without using big training data. We believe that there is much room for future research.

Approach: The best systems vary according to weighted scores. However, the differences between the second systems are small. Among them, we can see three major trends among the best (or second system).

(1) Language Model Approach: Almost all systems employ a BERT (or BERT family) language model. These BERTs are the driving force of the current NLP, and BERT without any special techniques is strong enough like the E1 system, which shows good performance in Subtask1-CR-JA. The next question is what kind of BERT is suitable for medical NLP. The EN trend is a domain-specific BERT, such as BioBERT and Clinical BERT, which is better than the general BERT shown in team-H systems for Subtask1-CR/RR-EN. However, in JA, a general BERT shows a sufficiently good performance, as shown in the team-E systems. In the near future, we will need a more accurate discussion of this issue.
(2) Data augmentation: Many systems utilize data augmentation techniques. The remaining issue is the suitable type of augmentation. One of the best Subtask1-CR-EN systems is the H2 system, which is based on back-translation. Considering that machine translation performance is already feasible, the back-translation approach is promising.

For Subtask1-CR-JA, the best performing system was E1, which applied a general domain BERT. G1 and J1-4 also used it but adopted data augmentation as well. Because their performance was lower, simple rule-based data augmentation may not be suitable for medical-domain NER. The domain-specific BERT used in A1 and A2 did not contribute significantly to performance. UTH-BERT was trained on health records, the writing style of which could be fragmented and compressed owing to the nature of personal notes. Wikipedia articles, on which the general Japanese BERT model was trained, could be more akin to case reports.

However, for the other JA tasks, UTH-BERT performed the best. Radiology reports (RR), another type of in-hospital note, are written in a similar manner as health records, which may have resulted in performance gain. For an extremely few-resource setting, that is, Subtask 2, domain knowledge seems to matter more.

In the English corpora (EN), H1-3 performed best. These systems characteristically apply span-based NER methods. This emerging trend in NER appears promising in a few-resource medical setting. Interestingly, naive data augmentation reduced performance (C <

Table 7: Results of Subtask 3 for MedTxt-RR (CI). Bold font indicates the best score for each evaluation metric.

(a) Results of Subtask3-RR-JA.

System ID	Score (Normalized Mutual Info)
D1	0.3569
E1	**0.5415**
F1	0.1744
J1	0.4161
J1*	0.4622

Revised results were submitted after the deadline of the formal run.

(b) Results of Subtask3-RR-EN.

System ID	Score (Normalized Mutual Info)
C1	**0.8724**
C2	0.8463
C3	0.8468
C4	0.8468
C5	0.8468
C6	0.8468
C7	0.8581
C8	0.8468
C9	0.8468
C10	0.8576
F1	0.2172
I1	0.7879

F and J). We might require medical domain-specific data augmentation policies.

5.2 Subtask 3

5.2.1 ADE challenge for MedTxt-CR. Table 6 lists the results of the ADE challenge for MedTxt-CR. For entity-level ADE, most systems achieved high performance with ADEval = 0, which is reasonable because the vast majority of entities had zero values in both the training and test datasets. However, most systems struggled to identify ADEval = 1 and 3 entities. Especially for ADEval = 1 entities, most systems were almost unable to find the entities. This result is understandable because neither ADEval = 1 nor ADEval = 3 entities appeared frequently in the training dataset. However, all the systems identified many more entities with ADEval = 3 than with ADEval = 1. In fact, the average F1 performance is 10 times larger in ADEval = 3 (i.e., 41.05) than in ADEval = 1 (i.e., 4.87), probably because of the "strength" of the signal; under our annotation scheme, ADEval = 3 is the marker for an ADE or an ADE trigger whereas ADEval = 1 is a weak sign of a negative ADE or an ADE trigger. Linguistic clues may appear more clearly in ADEval = 3 than in ADEval = 1. This result is also preferable because detecting ADE and ADE trigger signals is more important for real-world applications such as automated ADE reporting. Note that the test dataset did not contain any ADEval = 2 entities. Thus, we did not evaluate ADEval scores.

The report-level ADE performance is inconsistent with the entitiy-level performance of some systems. Thus, a better entity-level system is not necessarily a better report-level system. This result implies that such systems generate entity-level false positives, even in non-ADE reports. In other words, the systems with smaller gaps between the entity-level ADEval = 3 scores and report-level scores behave coherently in the micro-and macroscopic identification of ADE.

Although a score gap was found between the EN and JA corpora in both entity- and report-level evaluations, we cannot infer any innate linguistic difficulty in this task because the JA track participants were much fewer than the EN track participants.

For the Japanese corpus (JA), the F systems performed better in ADE signal detection, that is, the F-scores of ADEval=3 and report level. This could be because the F1 and F2 systems incorporated more contextual information than E1 by solving the task as NER.

For the English corpus (EN), I1 performed well at the report-level score. The approach jointly solved this ADE with Subtask 1 RR (NER), which is similar to that of the F1 and F2 systems, although they did not perform much in the English corpus, unlike in the Japanese corpus. We can note that H1-6 performs nicely in general. They adopted an interesting approach based on prompt-based learning, which trained on automatically generated snippets to explicitly explain which entity in a report was related to an ADE. A global remark could be how to enhance context information that specifically matters in this task.

5.2.2 CI challenge for MedTxt-RR. Table 7 lists the NMI scores of each system for the MedTxt-RR CI challenge. The C1 system, which uses heuristics for cancer size matching and Sentence-BERT encoding [15], achieved the highest performance of all systems. As shown in Table 8, the C1 system succeeded in grouping radiology reports of cases 4 and 5 into a single cluster, suggesting that matching lesion size is helpful in case distinction.

A large discrepancy is observed between the scores of the D1 and I1 systems, although they both used the NER-based approach. This may reflect the difference in the availability of biomedical knowledge bases between Japanese and English. While system I1 could use UMLS to normalize biomedical entities, system D1 had to create bag-of-entity vectors from only the training set, which probably had difficulty dealing with unseen entities in the test set.

As shown in Table 9, most systems grouped the test cases into the same number of clusters as the gold standard, although we did not clarify the true cluster number. In this task, the true cluster number was easily determined by the test sample size as used by the C1 system.

The baseline E1 system split the test cases into far more clusters than the true case numbers, although it achieved the highest NMI in RR-JN. This is problematic considering real-world applications, because NMI has failed to penalize the inability to recognize radiology report similarity. The bottom row of Table 10 shows the result of an extreme prediction where all test samples are split into different clusters of size one, which suggests an improper reward from the NMI.

Thus, we re-evaluated each system by adding adjusted normalized mutual information (AMI) [16] and Fowlkes-Mallows (FM) scores [5]. Table 10 shows that AMI and FM penalize the splitting

Table 8: The number of clusters into which each case was split by each system in Subtask 3 for MedTxt-RR (CI).

Case ID	TNM cancer staging	Cluster number																
		RR-JA					RR-EN											Gold standard
		D1	E1	F1	J1	J1*	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	F1	I1	
4	T2aN0M0	5	6	6	3	4	1	1	1	1	1	1	1	1	1	1	5	2	1
5	T2bN0M0	6	7	5	4	5	1	1	1	1	1	1	1	1	1	1	5	2	1
7	T3N1M0	3	5	5	4	3	2	2	2	2	2	2	2	2	2	2	6	1	1
8	T3N3M0	3	8	6	5	4	2	3	3	3	3	3	2	2	2	2	7	1	1
10	T4N0M0	3	6	5	4	4	1	1	1	1	1	1	1	2	2	1	6	1	1
14	T4N3M1a	5	5	5	4	3	2	2	2	2	2	2	2	2	2	2	5	3	1
15	T2N2M1c	2	5	5	2	2	3	3	3	3	3	3	3	3	3	3	5	4	1

The revised results were submitted after the deadline of the formal run.

Table 9: Cluster sizes created by each system in Subtask 3 for MedTxt-RR (CI).

System ID		Cluster number	Cluster size
Gold standard		7	9, 9, 9, 9, 9, 9, 9, 9, 9
RR-JA	D1	8	18, 17, 9, 8, 4, 3, 2, 2
	E1	33	19, 5, 4, 3, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
	F1	7	11, 11, 10, 9, 8, 7, 7
	J1	7	16, 10, 10, 7, 7, 7, 6
	J1*	7	13, 11, 11, 8, 7, 7, 6
RR-EN	C1	7	10, 9, 9, 9, 9, 9, 8
	C2	7	11, 9, 9, 9, 9, 9, 7
	C3	7	10, 10, 9, 9, 9, 9, 7
	C4	7	10, 10, 9, 9, 9, 9, 7
	C5	7	10, 10, 9, 9, 9, 9, 7
	C6	7	10, 10, 9, 9, 9, 9, 7
	C7	7	10, 10, 9, 9, 9, 9, 7
	C8	7	10, 9, 9, 9, 9, 9, 8
	C9	7	10, 9, 9, 9, 9, 9, 8
	C10	7	11, 9, 9, 9, 9, 9, 7
	F1	9	12, 10, 8, 8, 7, 6, 6, 5, 1
	I1	9	12, 12, 9, 9, 9, 7, 2, 2, 1

The revised results were submitted after the deadline of the formal run.

Table 10: Performance of each system of Subtask 3 for MedTxt-RR (CI) in multiple evaluation metrics. Bold font indicates the best score for each evaluation metric.

System ID		NMI	AMI	FM
RR-JA	D1	0.3569	0.1988	0.2674
	E1	**0.5415**	0.1489	0.1814
	F1	0.1744	-0.0117	0.1170
	J1	0.4161	0.2838	0.3044
	J1*	0.4622	**0.3409**	**0.3622**
RR-EN	C1	**0.8725**	**0.8437**	**0.8436**
	C2	0.8463	0.8116	0.8110
	C3	0.8468	0.8122	0.8126
	C4	0.8468	0.8122	0.8126
	C5	0.8468	0.8122	0.8126
	C6	0.8468	0.8122	0.8126
	C7	0.8581	0.8261	0.8166
	C8	0.8468	0.8123	0.8119
	C9	0.8468	0.8123	0.8119
	C10	0.8576	0.8255	0.8150
	F1	0.2172	-0.0045	0.1085
	I1	0.7879	0.7309	0.6992
Extreme prediction (isolate-all)		0.6392	-4.7901	0.0000

The revised results were submitted after the deadline of the formal run.
NMI, normalized mutual information; AMI, adjusted normalized mutual information; FM, Fowlkes-Mallows score.

of clinically similar documents into numerous clusters. This is supported by the scores at the bottom of Table 10. With AMI and FM, the J1 system achieved the highest scores, suggesting the effectiveness of sentence classification in determining TNM staging even in a limited availability of a knowledge base.

In summary, the CI challenge results show a difference in effective strategy for case clustering between RR-JA and RR-EN. For RR-EN, embedding distance with the help of a knowledge base works well and can be applied to other clinical specialties beyond lung cancer. For RR-JA, the lack of a knowledge base motivated participants to adopt a more dataset-specified approach, resulting in comparatively lower performance and limited possibility of application beyond lung cancer.

6 CONCLUSIONS

This study introduced the Real-MedNLP task setting, which is a medical NLP shared task handling three different tasks (named entity recognition (NER), case identification (CI), and adverse drug event extraction (ADE)) in a bilingual language (English and Japanese). The basic approach is twofold: (1) to employ data augmentation and (2) to utilize domain-specific language models such as BioBERT and ClinicalBERT. These approaches partially solve the low-resource problem in MedNLP. However, the performance in an extremely low resource-setting task (Subtask 2 guideline learning) is insufficient. In particular, for newly designed tasks (Subtask 3 CI challenge and ADE challenge), we needed to start by deciding how to evaluate them and still work for suitable evaluation systems. Considering that not only our three tasks but also various medical

tasks are waiting for NLP solutions, it is important to organize and share the approach and results across the world. We believe that our datasets and the results of all participants will contribute to the boost of future research.

ACKNOWLEDGEMENTS

This work was supported by JST, AIP Trilateral AI Research, Grant Number JPMJCR20G9, and JST AIP-PRISM Grant Number JPMJCR18Y1, Japan.

REFERENCES

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 72–78. https://doi.org/10.18653/v1/W19-1909

[2] Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, and Yuta Nakamura. 2022. Natural Language Processing: from Bedside to Everywhere. *IMIA Yearbook of Medical Informatics 2022* (2022), 257–267.

[3] Alan R. Aronson. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. (2001). http://skr.nlm.nih.gov/papers/references/metamap_01AMIA.pdf

[4] 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)*. Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[5] E. B. Fowlkes and C. L. Mallows. 1983. A Method for Comparing Two Hierarchical Clusterings. *J. Amer. Statist. Assoc.* 78, 383 (1983), 553–569. https://doi.org/10.1080/01621459.1983.10478008 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/01621459.1983.10478008

[6] 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 Transactions on Computing for Healthcare (HEALTH)* 3, 1 (2021), 1–23.

[7] 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)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1375

[8] Colleen R Kelly, Sachin S Kunde, and Alexander Khoruts. 2014. Guidance on preparing an investigational new drug application for fecal microbiota transplantation studies. *Clinical gastroenterology and hepatology: the official clinical practice journal of the American Gastroenterological Association* 12, 2 (Feb. 2014), 283–288.

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, , and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[10] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana K. Savova. 2021. EntityBERT: Entity-centric Masking Strategy for Model Pretraining for the Clinical Domain. In *BIONLP*.

[11] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2014. Overview of the NTCIR-11 MedNLP Task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

[12] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2016. Overview of the NTCIR-12 MedNLPDoc Task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-12, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

[13] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, and Eiji Aramaki. 2013. Overview of the NTCIR-10 MedNLP Task. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-10, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

[14] E. Pons, L. M. Braun, M. G. Hunink, and J. A. Kors. 2016. Natural Language Processing in Radiology: A Systematic Review. *Radiology* 279, 2 (May 2016), 329–343.

[15] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[16] Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?. In *Proceedings of the 26th Annual International Conference on Machine Learning* (Montreal, Quebec, Canada) *(ICML '09)*. Association for Computing Machinery, New York, NY, USA, 1073–1080. https://doi.org/10.1145/1553374.1553511

[17] Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2017. Overview of the NTCIR-13 MedWeb Task. In *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-13, National Center of Sciences, Tokyo*. National Institute of Informatics (NII).

[18] Kawazoe Y, Shibata D, Shinohara E, Aramaki E, and Ohe K. 2021. A clinical specific BERT developed using a huge Japanese clinical text corpus. *PLoS One* 15, 11 (2021), e0259763.

[19] 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 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4565–4572. https://www.aclweb.org/anthology/2020.lrec-1.561

[20] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Chinese Information Processing Society of China, 1218–1227. https://aclanthology.org/2021.ccl-1.108

APPENDIX

Table 11: NEs of the training set in MedTxt-CR and MedTxt-RR.

Data-set		CR-JA	CR-EN	RR-JA	RR-EN
# of texts		148	148	72	72
# of chars (ave.)		84471(570)	40383(272)	16861(234)	8488(117)
<a>	total	823	819	464	465
<d>	total	2348	2346	884	883
	"positive"	1695	1693	465	462
	"suspicious"	80	80	191	191
	"negative"	251	251	149	148
	"general"	302	302	1	1
<t-test>	total	387	388	26	27
	"scheduled"	0	0	0	0
	"executed"	362	363	19	19
	"negated"	7	7	2	2
	"other"	18	18	5	6
<timex3>	total	1353	1353	29	29
	"date"	539	539	26	26
	"time"	53	53	0	0
	"duration"	82	82	2	2
	"set"	34	34	0	0
	"age"	189	189	0	0
	"med"	428	428	1	1
	"misc"	28	28	0	0
<m-key>	total	344	344	0	0
	"scheduled"	0	0	0	0
	"executed"	266	266	0	0
	"negated"	27	27	0	0
	"other"	51	51	0	0
<m-val>	total	64	64	0	0
	"scheduled"	0	0	0	0
	"executed"	0	0	0	0
	"negated"	2	2	0	0
	"other"	0	0	0	0
<t-key>	total	524	524	1	1
<t-val>	total	427	427	0	0
<f>	total	638	636	345	340
<c>	total	569	569	22	22
<r>	total	678	678	2	1
<cc>	total	266	266	16	15

CR indicates the case report corpus, and RR represents the radiology report corpus.
Although the corpora included general medical annotations, we only evaluated high-frequency labels (highlighted rows).