Journal of Biomedical Informatics 116 (2021) 103729

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Original Research

Extracting clinical terms from radiology reports with deep learning

Kento Sugimoto [a,b,*], Toshihiro Takeda [a], Jong-Hoon Oh [b], Shoya Wada [a,b], Shozo Konishi [a], Asuka Yamahata [a], Shiro Manabe [a], Noriyuki Tomiyama [c], Takashi Matsunaga [d], Katsuyuki Nakanishi [e], Yasushi Matsumura [a]

[a] Department of Medical Informatics, Osaka University Graduate School of Medicine, Suita, Osaka, Japan
[b] National Institute of Information and Communications Technology, Seika, Kyoto, Japan
[c] Department of Diagnostic and Interventional Radiology, Osaka University Graduate School of Medicine, Suita, Osaka, Japan
[d] Department of Medical Informatics, Osaka International Cancer Institute, Osaka, Japan
[e] Department of Diagnostic and Interventional Radiology, Osaka International Cancer Institute, Osaka, Japan

ARTICLE INFO

Keywords:
Natural Language Processing
Radiology Report
Information Extraction
Deep Learning

ABSTRACT

Extracting clinical terms from free-text format radiology reports is a first important step toward their secondary use. However, there is no general consensus on the kind of terms to be extracted. In this paper, we propose an information model comprising three types of clinical entities: observations, clinical findings, and modifiers. Furthermore, to determine its applicability for in-house radiology reports, we extracted clinical terms with state-of-the-art deep learning models and compared the results. We trained and evaluated models using 540 in-house chest computed tomography (CT) reports annotated by multiple medical experts. Two deep learning models were compared, and the effect of pre-training was explored. To investigate the generalizability of the model, we evaluated the use of other institutional chest CT reports. The micro F1-score of our best performance model using in-house and external datasets were 95.36% and 94.62%, respectively. Our results indicated that entities defined in our information model were suitable for extracting clinical terms from radiology reports, and the model was sufficiently generalizable to be used with dataset from other institutions.

1. Introduction

Radiology examination is one of the important steps for diagnosing diseases and developing a treatment plan. Radiologists read a radiology examination image and prepare a diagnostic imaging report to communicate with referring clinicians. Radiology reports include clinical information about observed structures, diagnostic possibilities, and recommendations for treatment plans. Such information is also valuable for observational study and clinical decision support. However, since radiology reports usually have a free-text format and contain misspellings, abbreviations, and non-standard terminology, it is challenging to extract clinical information from radiology reports.

Information extraction (IE) [1], a natural language processing (NLP) task used to extract structured information from unstructured texts, is a promising approach to extract clinical information from radiology reports. For example, consider the following sentence in a radiology report:

A 3-cm increasing irregular nodule in the right upper lobe suggests lung cancer.

Here, the term "lung cancer" is a clinical finding identified by the radiologist observing the "nodule". The terms, "3-cm", "increasing", "irregular", and "right upper lobe" are modifiers of the term "nodule". We show an example of extracting structured information from the above sentence (Fig. 1). In Fig. 1, observation entity and clinical finding entity include the terms "nodule" and "lung cancer", respectively. Similarly, size entity, change entity, characteristics entity, and anatomic location entity, which are subclasses of the modifier entity, include the terms "3-cm", "increasing", "irregular", and "right upper lobe", respectively. The structured format allowed access to the clinical information that the radiologist suspects a diagnosis of "lung cancer" because of the presence of a "nodule", and its specific features. Thus, information extraction enables to organize clinical terms in a report.

However, there is no general consensus on the kind of clinical terms to be extracted and the ways in which clinical entities can be defined. Several previous works [2,3] have proposed an information model with defined "observation entities" for some observations and "modifier

* Corresponding author.at: Osaka University, 2-2 Yamadaoka Suita, Osaka, Japan.
E-mail address: sugimoto.kento@hp-info.med.osaka-u.ac.jp (K. Sugimoto).

https://doi.org/10.1016/j.jbi.2021.103729
Received 7 May 2020; Received in revised form 24 February 2021; Accepted 25 February 2021
Available online 9 March 2021

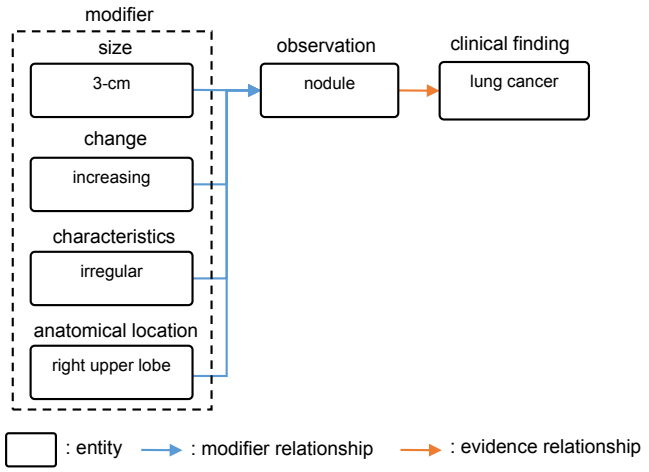

Fig. 1. . Example representing the sentence "*A 3-cm increasing irregular nodule in the right upper lobe suggests lung cancer*" in a structured format.

entities" for the "observation entities". These models are well organized for the extraction of clinical terms from reports. However, since these models do not distinguish between observation entities and clinical finding entities, "nodule" and "lung cancer" are assigned as the same entity, even though they have different semantic roles in the context.

We propose an information model mainly comprising observation entities, clinical finding entities, and modifier entities. We differentiated clinical finding entities from observation entities for assigning appropriate entities according to the semantics of the reports. This enables to represent free-text radiology reports in a structured format, as shown in Fig. 1. From the structured format, we can easily capture clinical information regarding what is observed on the image and which disease is suspected by the radiologist. Furthermore, we applied state-of-the-art deep learning models [4,5] and evaluated their performance to verify the applicability of our information model to in-house radiology reports. We also evaluated the generalizability of the deep learning model, using radiology reports from other institution. In this paper, we focus on defining the information model and applying the existing state-of-the-art deep learning models [4,5], rather than building novel deep learning models.

The rest of this paper is organized as follows. Section 2 provides a brief review of related works. Section 3 describes an information model, and Section 4 defines the corpus and annotation scheme. In Sections 5 and 6, we report and discuss the result of our experiments and limitations. Finally, Section 7 summarizes and concludes our work.

2. Related work

2.1. Clinical information extraction

Many clinical NLP systems have been proposed for information extraction from unstructured clinical texts [6]. In earlier work, heuristic methods such as dictionary-based and pattern matching methods were common. MedLEE utilizes a pre-defined dictionary to convert radiology reports into a structured format [7]. Mayo Clinic's Text Analysis and Knowledge Extraction System (cTAKES) [8] tool combines dictionary and machine learning methods. cTAKES uses the Unified Medical Language System (UMLS) [9] for dictionary inquiries. One major issue with dictionary-based NLP systems is that they do not always establish high performance when handling in-house raw clinical texts, particularly texts that contain misspellings, abbreviations, and non-standard terminology. In addition, MedLEE and cTAKES only cover English clinical texts and cannot handle non-English clinical texts. Other dictionary-based NLP systems, such as Health Information Text Extraction (HITEx) [10] and MetaMap, also cover only English texts. Languages other than English, including Japanese, do not have sufficient clinical resources like UMLS. This has been a major obstacle in developing clinical NLP systems in countries where English is not the official language [11].

Machine learning approaches have received significant attention in the past decade. Esuli et al. [12] applied the linear-chain conditional random field (CRF) [13] to extract clinical terms from mammography reports. Hassanpour and Langlotz [3] also trained CRF for chest computed tomography (CT) reports. They compared its performance with cTAKES and showed that the machine learning approach yielded better results than dictionary-based NLP systems.

Deep learning approaches have drawn a great deal of attention in recent years. Cornegruta et al. [14] built bidirectional long short-term memory (BiLSTM) [15] to extract clinical terms from chest X-ray reports. Miao et al. [16] also built a BiLSTM for breast ultrasound reports in Chinese. They both showed that BiLSTM yielded better results than the CRF. More recently, contextualized word representation models, such as ELMo [17] and BERT [5], have achieved better results in various NLP tasks, including clinical information extraction. Si et al. [18] reported that ELMo and BERT had higher performance than BiLSTM-based models in clinical information extraction tasks.

2.2. Information model

Some previous studies have proposed information models for extracting clinical information from unstructured clinical texts. Friedman et al. [2] defined the concept "*rad_finding*" to represent findings of radiology examinations. This comprises an observation and its modifiers (e.g., body location, location qualifiers, certainty, degree temporal, quantity, and property). Langlotz and Meininger [19] defined "*Description Set*" to represent clinical entities of radiology examination. "*Description Set*" includes a finding on an imaging examination, the location of the finding on the imaging study, and the anatomic location of the finding. Using this information model, they developed a structured reporting system. Based on their study, Hassanpour and Langlotz [3] proposed a new information model. Their model has five concepts: anatomy, anatomy modifier, observation, observation modifier, and uncertainty. They applied their information model to multi-institutional chest CT reports.

3. Our information model

In this section, we elaborate on our information model in detail. The information model was designed to extract clinical information in a structured format that mainly comprises three entity groups: observation entity, clinical finding entity, and modifier entity. Observation entities are specific terms representing abnormal features, such as lesions and opacities in the image. Clinical finding entities are prepared for any terms identified by the radiologists on the basis of observation entities. According to the semantic role of the entity, the modifier entity was categorized into five entities: anatomical location modifier, certainty modifier, change modifier, characteristics modifier, and size modifier. Thus, seven entities were defined in our information model (see Table 1). The description of RadLex [20], a comprehensive glossary of radiology domain terms, was referenced when defining each entity. Anatomical location modifier is based on the description of "anatomical entity", "anatomically-related descriptor" and "location descriptor" in RadLex. Anatomical location terms were included in this entity because preliminary experiments showed that determination of boundaries between anatomical terms (e.g., lung, lobe) and location terms (e.g., right, lower) was often difficult. This result was consistent with the results reported in a previous study [3]. In addition, we initially prepared more fine-grained characteristics modifier according to the sub-classes of "Radlex descriptor". However, in our preliminary experiments, we found it was very time-consuming and difficult to annotate accurately according to the fine-grained characteristics modifier. Thus, we

Table 1
. Example terms for entities and corresponding RadLex categories.

Entity	Example	RadLex category
Observation	nodule, mass, ground-glass opacity	Imaging observation
Clinical finding	lung cancer, emphysema	Clinical finding
Anatomical location modifier	right lung, left lung S1 + 2, lymph node	Anatomical entity
		Anatomically-related descriptor
		Location descriptor
Certainty modifier	uncertain, no, definite	Certainty descriptor
Change modifier	unchanged, increased, advanced	Status descriptor
		Temporal descriptor
Characteristics modifier	diffuse, coarse, flattened, irregular	Density descriptor
		Imaging observation descriptor
		Morphologic descriptor
		Quantity descriptor
Size modifier	4 mm, 3.0 cm, 30 × 14 mm	Size descriptor

prepared change modifier and characteristics modifier by integrating several sub-classes of "Radlex descriptor". As well as for accurate annotation, we believe that to leave the granularity in a coarse level is a practical choice in this step since appropriate granularity at post-coordination process highly depends on the downstream task. Finally, certainty modifier was added to our information model since negation detection is important in clinical information extraction.

4. Corpus and annotation scheme

4.1. Corpus development

For this study, 118,078 chest CT reports from 2010 to 2018 that were stored in the radiology information system at Osaka University Hospital (OUH) were used. Osaka University Hospital is a general hospital comprising six clinical divisions: Medicine; Surgery; Sensory; Cutaneous and Motor Organ Medicine; Clinical Neuroscience; Woman, Child Health, and Urology; and Radiology. Since requests for radiology examination come from various clinical departments, our dataset includes contents of different kinds of diseases. More details about the corpus can be found in the supplementary material. For evaluation of the generalizability of our model, we collected 77 chest CT reports from the Osaka International Cancer Institute (OICI). Both datasets are written in Japanese. This study was approved by institutional review boards of Osaka University Hospital (Permission number: 19276), and Osaka International Cancer Institute (Permission number: 19175).

Data cleansing was performed on all reports. First, header and footer information was removed from reports. Next, full-width characters were converted to half-width characters. In addition, strings related to date were converted to "$" using regular expressions. Each report was segmented into a single sentence, and word tokenization was implemented by MeCab [21].

4.2. Annotation scheme

The annotation process to create a gold standard requires domain knowledge. Therefore, three medical experts (two clinicians and one radiological technologist) performed the annotation process. To achieve consistent annotation, the annotators were provided with a guideline describing rules and terms to annotate for each entity. Several sample reports were given to each annotator to improve the understanding of the task before the annotation process. Each report was annotated by all three annotators to avoid annotator bias. Annotators independently annotated 540 reports from Osaka University Hospital and 77 reports from the Osaka International Cancer Institute. In the adjudication process, annotation disagreements were resolved by a majority vote of the annotators. After the first annotation process, we determined that some disagreement between annotators may be caused by the lacking guidelines. Therefore, the guidelines were improved based on reports in which all annotators had different annotation results. After improving our guidelines, the annotators re-annotated only sentences for which the results did not match in the first annotation process. The adjudicator (a clinician who did not attend the annotation process) resolved result disagreements among the three annotators if the gold standard could not be determined by a majority vote. Cohen's kappa [22] and Fleiss' kappa [23] are widely used metrics of inter-annotator agreement (IAA). However, these metrics are not suitable for sequence labelling tasks in which annotators have to classify entities and also determine their spans [24]. Instead of Cohen's kappa or Fleiss' kappa, micro F1-score was used to indicate IAA, following previous studies [24–26]. The mean IAA score for the three annotators was 91%,[1] which denotes very high agreement. Annotation work was performed using TALEN [27], a web-based annotation tool. Finally, annotation results were saved in the IOB2 format [28].

5. Experiments

5.1. Experimental settings

Named entity recognition (NER) identifies pre-defined entities from the text and is well suited to the preliminary task of extracting clinical information. BiLSTM-CRF [29], BERT, and BERT-CRF were used as state-of-the-art deep learning models in our experiments.

For BiLSTM-CRF, two different experimental setups about word embeddings were explored to evaluate the effectiveness of pre-training on domain-specific corpora. Both word embeddings were obtained using *word2vec* [30].

- BiLSTM-CRF (Wiki)

Word embeddings were pre-trained using Japanese Wikipedia articles[2] (12,056,502 sentences). Words that occurred more than five times in the corpus were included in the vocabulary.

- BiLSTM-CRF (CT)

Word embeddings were pre-trained using chest CT reports from Osaka University Hospital (117,970 reports; 952,292 sentences). All words in the reports were included in the vocabulary.

For BERT and BERT-CRF, two different pre-trained BERT models that were fine-tuned for our NER task were prepared.

- BERT(Wiki)

The publicly available pre-trained Japanese BERT model[3] was used. The model was pre-trained on Japanese Wikipedia. $\text{BERT}_{\text{BASE}}$ subword tokenization model with whole word masking was chosen. The model has 12 layers of transformer blocks, 768 hidden units, and 12 self-attention heads.

- BERT(CT)

This model was initialized from pre-trained **BERT(Wiki)** and used chest CT reports from Osaka University Hospital (117,970 reports; 952,292 sentences) for the second phase of pre-training for

[1] We confirmed that three F-1 scores between annotators (annotator A vs B, annotator B vs C, and annotator A vs C) were almost same.
[2] jawiki-latest-pages-articles.xml.bz2 (03-Jan-2020)
[3] https://github.com/cl-tohoku/bert-japanese

Table 2
. Number of entities in each dataset.

Entity	Train	Dev	Test
Observation	1,095	137	313
Clinical finding	2,258	324	683
Anatomical location modifier	2,321	365	700
Certainty modifier	2,180	298	647
Change modifier	835	127	237
Characteristics modifier	541	88	169
Size modifier	247	36	98
Total	9,477	1,375	2,847

Table 3
. Hyper-parameter search space for BiLSTM-CRF.

Hyper-parameter	Range
Batch size	{16, 32, 64, 128}
Dropout rate	{0.3, 0.5, 0.7}
Word embedding dim	{50, 100, 300}
Hidden layer dim	{64, 128, 256}

approximately 10,000 steps using a batch size of 32, maximum sequence length of 128, and Adam optimizer with a learning rate of 1e-4. The dataset for the second phase of pre-training was augmented by randomly masking different words. The second phase of pre-training took roughly five days with a single NVIDIA GeForce GTX 1080 GPU.

- BERT-CRF(Wiki)

This model concatenated the CRF layer to the output of BERT to result in the same settings as **BERT(Wiki)** except for the CRF layer.

- BERT-CRF (CT)

This model also had the same settings as **BERT(CT)** except for the CRF layer.

A total of 540 annotated reports were divided into 378 reports for training, 54 reports for development (dev), and 108 reports for testing. Table 2 shows the number of entities in each dataset.

The best hyper-parameter settings were chosen using a development dataset with Optuna [31]. Table 3 shows the hyper-parameter search space of BiLSTM-CRF. We searched the best hyper-parameter settings for **BiLSTM-CRF (Wiki)** and **BiLSTM-CRF (CT)**. Mini-batch stochastic gradient descent (SGD) with momentum was employed for parameter optimization, and the initial learning rate and momentum were set to 0.1 and 0.9, respectively. Learning rate was reduced when the F1-score of the development dataset stopped improving. Dropout [32] was applied on both input and output vectors of BiLSTM. Gradient clipping of 5.0 was used.

The hyper-parameter search space for fine-tuning of BERT and BERT-CRF is shown in Table 4.

To mitigate the effects of a random seed, all experiments were repeated five times for each model, and their mean scores were compared.

5.2. Comparison of performance of deep learning models

Entity level F1-score for each model was used as an evaluation metric, and the results were aggregated by micro-averaging (Table 5). Post-hoc tests using the Tukey–Kramer method were used for multiple comparison analysis (Table 6). Table 5 shows that **BiLSTM-CRF (CT)** achieved the best performance by the mean F1-score. This model was significantly different ($p < 0.05$) from other models except for **BERT-CRF(CT)**. Tables 5 and 6 show that pre-training with CT reports resulted in significantly higher F1-scores in each model compared to using only Wikipedia.

BiLSTM-CRF (CT) achieved the highest mean F1 score; thus, we used this model for the remainder of the experiments.

Table 4
. Hyper-parameter search space for BERT and BERT-CRF.

Hyper-parameter	Range
Epochs	{5, 10, 20}
Batch size	{16, 32, 64, 128}
Learning rate	{3e-5, 5e-5, 1e-4}

Table 5
. Comparison of models in mean F1-score (%).

Model	Pre-trained	
	Wikipedia	CT
BiLSTM-CRF	94.32	95.36
BERT	94.62	95.01
BERT-CRF	94.83	95.18

Table 6
. Statistical test results using the Tukey–Kramer method.

		BiLSTM-CRF		BERT		BERT-CRF	
		Wiki	CT	Wiki	CT	Wiki	CT
BiLSTM-CRF	Wiki	–	< **0.001**	0.034	< 0.001	< **0.001**	< **0.001**
	CT	–	–	< **0.001**	0.009	< **0.001**	0.405
BERT	Wiki	–	–	–	0.003	0.218	< **0.001**
	CT	–	–	–	–	0.358	0.418
BERT-CRF	Wiki	–	–	–	–	–	0.007
	CT	–	–	–	–	–	–

Notes: Wiki stands for Wikipedia; $P < 0.001$ are in bold, and $P < 0.05$ are underlined.

Table 7
. Performance metrics of each entity.

Entity	Precision	Recall	F1-score
Observation	94.19	94.25	94.22
Clinical finding	95.08	96.19	95.61
Anatomical location modifier	94.75	96.46	95.60
Certainty modifier	98.88	98.15	98.51
Change modifier	90.89	91.73	91.30
Characteristics modifier	91.11	87.22	89.12
Size modifier	93.15	96.94	95.00
Micro-averaging	95.11	95.61	95.36

5.3. Detailed performance for each entity

The detailed performance of **BiLSTM-CRF(CT)** is shown in Table 7. The certainty modifier established higher performance, while the change modifier and characteristics modifier had lower F1-scores than other entities.

5.4. Effect of training data size

The effect of training data size on performance was further explored (Fig. 2). When training data size was reduced to 10%, the certainty modifier maintained a high F1-score, while the F1-score of the characteristics modifier dropped significantly.

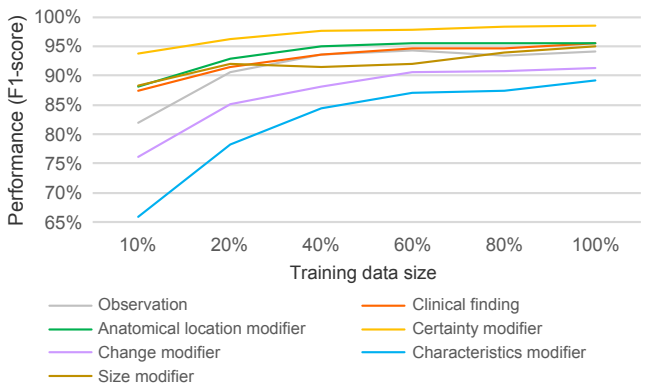

Fig. 2. Learning curves for performance versus training data size.

5.5. Comparison of performance with previous information models

To demonstrate that our information model is as well suited to extracting clinical terms from radiology reports as previous information models [2,3], we built another information model which was similar to previous implementations, and re-annotated. This information model integrated observation entities and clinical finding entities into one entity. Modifier entities were the same as in the original. Table 8 shows the result of a comparison between a dataset based on the original information model and a dataset based on the new one. The F1-scores of the observation entities and the clinical finding entities were 1.47% lower and 0.08% lower than those of the integrated entity, respectively.

5.6. Comparison of performance dataset from another institution

Finally, we evaluated the model generalizability by using a dataset from another institution. The difference in the micro-averaging F1-score was less than 1% between the in-house dataset and the other institutional dataset (Table 9).

6. Discussion

6.1. Annotation agreement

We asked three medical experts to attend annotation process. As mentioned in Section 4, we could obtain satisfactory IAA score to build gold standard dataset. From the annotation result, we found disagreements among the annotators often occurred at the boundaries of a term. For example, one clinical finding entity was assigned to "*nodular lymphoid hyperplasia*" by the majority vote. Conversely, the characteristics modifier was assigned to "*nodular*" in case of "*nodular thickening*". Wang [33] also pointed out that disagreements often occur at term boundaries when annotating clinical texts. In RadLex,[4] while no terms like "*nodular lymphoid hyperplasia*" and "*nodular thickening*" exist, the terms "*nodular hyperplasia*" and "*nodular ganglioneuroblastoma*" belong to "Clinical finding". The term "*nodular*" belongs to "Morphologic descriptor". This suggests difficulties in defining strict boundaries of clinical terms despite using RadLex. Although small disagreements are inevitable in build a dataset with the cooperation of several medical experts, we still believe that our annotation scheme has underwent enough consideration, and which leads to a high IAA score.

Table 8
. Comparison between our information model and previous models.

Entity	Our model	Previous models [2,3]
Observation	94.22	95.69
Clinical finding	95.61	
Anatomical location modifier	95.60	95.33
Certainty modifier	98.51	98.35
Change modifier	91.30	90.92
Characteristics modifier	89.12	89.85
Size modifier	95.00	96.65
Micro-averaging	95.36	95.49

Table 9
. Comparison with a dataset from another institution.

Entity	OUH	OICI	Delta
Observation	94.22	94.73	+0.51
Clinical finding	95.61	94.20	−0.14
Anatomical location modifier	95.60	95.68	+0.08
Certainty modifier	98.51	97.50	−0.10
Change modifier	91.30	90.50	−0.80
Characteristics modifier	89.12	88.76	−0.36
Size modifier	95.00	94.67	−0.34
Micro-averaging	95.36	94.62	−0.74

Notes: OUH and OICI stand for Osaka University Hospital and Osaka International Cancer Institute, respectively.

6.2. Comparison of performance between deep learning models

Tables 5 and 6 highlight that pre-training with CT reports resulted in significantly higher F1-scores in each model compared to using only Wikipedia. This result indicates that pre-training using domain corpora helps to improve performance, consistent with previous results [34]. BERT and BERT-CRF had significantly higher F1-scores than BiLSTM-CRF when pre-trained with only Wikipedia. This finding showed the effectiveness of contextualized word representation models for NER. Conversely, in contrast to previous results, BERT had a significantly lower F1-score than BiLSTM-CRF when pre-trained with CT reports, in contrast to previous results [18,35]. The size of the corpus we used for the second phase of pre-training of BERT may have caused this result. For example, Si et al. [18] initialized BERT with a pre-trained model provided by Devlin et al.[5] and used the MIMIC-III corpus [36] consisting of approximately 2 million clinical notes for the second phase of pre-training. However, we could only use almost 100,000 CT reports, which was probably not enough to adjust our NER task.

No significant difference was found between BERT-CRF and BiLSTM-CRF when pre-trained with CT reports. However, when considering the time cost of pre-training for domain adaptation, BiLSTM-CRF took only a few minutes, whereas BERT took roughly five days. Thus, we considered BiLSTM-CRF to be the appropriate model for our task.

As shown in Table 7, we established a comprehensive satisfactory result for NER. Observation entities and clinical finding entities were extracted with sufficient accuracy. We believe that extracting clinical terms of these entities is more important since they are more relevant to diagnosis and treatment. Table 7 also shows that F1-scores of change modifier and characteristics modifier are lower than those of other entities. Their low performance may be due to lexical diversity in their entities. Type Token Ratio (TTR) is a measure of lexical diversity within the corpus. TTR is the ratio of the total number of words (tokens) divided by the total number of unique words (types). High TTR indicates that entities are composed of various words. Table 10 shows the TTR of each entity in the training data. The results indicate that the TTR of characteristics modifier and change modifier are higher than those of other entities, which may lead to a lower F1-score.

6.3. Error analysis

We further analyzed error cases using the results of **BiLSTM-CRF (CT)** and revealed two frequent error patterns. The first pattern often occurs at the boundaries of a term. As we mentioned in section 6.1, we

[4] version 4.1, published November 2020

Table 10
. Type Token Ratio (TTR) of each entity in training data.

Entity	# of Type	# of Token	TTR
Observation	1,095	137	0.125
Clinical finding	2,258	357	0.158
Anatomical location modifier	2,321	292	0.126
Certainty modifier	2,180	96	0.044
Change modifier	835	193	0.231
Characteristics modifier	541	161	0.298
Size modifier	247	33	0.137
Total	9,477	1,269	0.133

considered that the misclassification about this pattern is inevitable since disagreements occur even among the medical experts. The second pattern is due to the clinical terms which play different roles according to the context despite the same surface form. For example, in the radiology reports in Japanese, the Chinese character "石灰化" was used not only as the noun "calcification", but also as the adjective "calcified". For "石灰化", the observation entity is appropriate when used as a noun while the characteristics modifier entity is appropriate when used as an adjective. Therefore, "石灰化" was assigned different entities according to the context. This difficulty sometimes accompanies with the wrong prediction.

6.4. Effect of training data size

The development of a large annotated corpus for clinical text requires medical experts. However, it is time-consuming and costly to recruit experts as annotators. To save on such costs, we explored the effect of training data size on performance. Our results will help to determine the size of new gold standard datasets in the future.

6.5. Comparison with previous information models

We differentiated clinical finding entities from observation entities in our information model. This allows to extract relationship between clinical finding entities and observation entities in the reports, which form the basic framework of semantic relations in the radiology reports. However, if the performance of NER of our information model was significantly lower than that of previous information models [2,3], it would suggest that our information model is not suitable for extracting clinical terms from reports. Thus, we performed additional experiment with our information model and previous information models. Table 8 shows that the performance of NER was almost the same as that of previous models, although our information model was more complicated than previous information models. As the next step for representing radiology reports in a structured format, extracting relationship between entities would be required. In addition, developing a controlled vocabulary and thesaurus are also expected to help organize extracted clinical terms. We believe that extracting clinical finding entity and observation entity separately at the entity level would benefit the next step.

6.6. Comparison of performance on a dataset from another institution

We showed that our trained model can be directly applied to a dataset from another institution without degradation of its performance. Although we used only in-house reports for the training and validation datasets, the difference between performance of the micro-averaging F1-score of the in-house dataset and that of a dataset from another institution was within 1% (Table 9). This result is probably due to the similarity of word distributions in the reports of the institutions. We observed that 98.7% words of the OICI dataset were in our vocabulary list built using the OUH dataset.

6.7. Limitations and future work

There are have some limitations of our dataset. First, we only used one institutional dataset for evaluating generalizability. Even though our experiments show that our model can be used for other institutions without using additional dataset, more datasets outside our institutions would be needed to ensure generalizability. Second, we used only chest CT reports, and have not evaluated radiology reports for other parts of the body, such as the abdomen. Direct tailoring of the trained model to reports related to other parts of the body would be difficult because of the differences in word distributions between the reports. Future work will focus on fine-tuning the trained model with a small amount of annotated data from reports pertaining to other parts of the body.

It is useful to extract both relationships and entities. However, we focused on extracting clinical terms in this study. Extracting relationships between entities is outside the scope of this paper. In the future, we plan to develop a relation extraction model.

7. Conclusion

In this paper, we defined an information model for radiology reports. Our information model enables observation and clinical findings to be distinguished. In addition, we applied state-of-the-art deep learning models to extract clinical terms from reports. This is an important first step towards representing radiology reports in a structured format. The micro F1-scores of our model for the in-house dataset and a dataset from another institution were 95.36% and 94.62%, respectively. We obtained satisfactory results showing that clinical terms could be accurately extracted, and entities in our information model are applicable for the NER of radiology reports. A further study is already in progress regarding reusing radiology reports.

CRediT authorship contribution statement

Kento Sugimoto: Conceptualization, Methodology, Software, Investigation, Resources, Writing - original draft, Visualization. **Toshihiro Takeda:** Conceptualization, Methodology, Data curation, Writing - review & editing, Supervision. **Jong-Hoon Oh:** Investigation, Writing - review & editing. **Shoya Wada:** Data curation, Validation, Writing - review & editing. **Shozo Konishi:** Data curation, Validation. **Asuka Yamahata:** Data curation. **Shiro Manabe:** Writing - review & editing. **Noriyuki Tomiyama:** Resources. **Takashi Matsunaga:** Resources. **Katsuyuki Nakanishi:** Resources. **Yasushi Matsumura:** Conceptualization, Methodology, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [This work was supported by the Council for Science, Technology and Innovation (CSTI), cross-ministerial Strategic Innovation Promotion Program (SIP), "Innovative AI Hospital System" (Funding Agency: National Institute of Biomedical Innovation, Health and Nutrition (NIBIOHN)), and partially supported by JSPS KAKENHI Grant Number 19H04496.]

Acknowledgment

This work was supported by the Council for Science, Technology and Innovation (CSTI), cross-ministerial Strategic Innovation Promotion Program (SIP), "Innovative AI Hospital System" (Funding Agency: National Institute of Biomedical Innovation, Health and Nutrition (NIBIOHN)), and partially supported by JSPS KAKENHI Grant Number 19H04496.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103729.

References

[1] S. Sarawagi, Information extraction, Found. Trends Databases 1 (2008) 261–377.
[2] C. Friedman, S.M. Huff, W.R. Hersh, E. Pattison-Gordon, J.J. Cimino, The Canon Group's Effort: Working Toward a Merged Model, J. Am. Med. Inform. Assoc. 2 (1995) 4–18.
[3] S. Hassanpour, C.P. Langlotz, Information extraction from multi-institutional radiology reports, Artif. Intell. Med. 66 (2016) 29–39.
[4] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer. Neural architectures for named entity recognition, 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference, 2016, pp. 260–270.
[5] J. Devlin, M.-W. Chang, K. Lee, K.T. Google, A.I. Language, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: North American Association for Computational Linguistics (NAACL), 2019, pp. 4171–4186.
[6] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research, Methods Inf. Med. 47 (2008) 128–172.
[7] C. Friedman, G. Hripcsak, W. DuMouchel, S.B. Johnson, P.D. Clayton, Natural language processing in an operational clinical information system, Nat. Lang. Eng. 1 (1995) 83–108.
[8] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.
[9] Unified Medical Language System (UMLS). http://www.nlm.nih.gov/research/umls.
[10] Q.T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S.N. Murphy, R. Lazarus, Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system, BMC Med. Inf. Decis. Making 6 (2006) 30.
[11] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical Natural Language Processing in languages other than English: opportunities and challenges, J. Biomed. Semant. 9 (2018) 12.
[12] A. Esuli, D. Marcheggiani, F. Sebastiani, An enhanced CRFs-based system for information extraction from radiology reports, J. Biomed. Inform. 46 (2013) 425–435.
[13] J. Lafferty, A. Mccallum, F.C.N. Pereira, F. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proc. ICML (2001).
[14] S. Cornegruta, R. Bakewell, S. Withey, G. Montana, Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks, Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI), 2016, p. 17–27.
[15] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (2005) 602–610.
[16] S. Miao, T. Xu, Y. Wu, H. Xie, J. Wang, S. Jing, et al., Extraction of BI-RADS findings from breast ultrasound reports in Chinese using deep learning approaches, Int. J. Med. Inf. 119 (2018) 17–21.
[17] M.E. Peters, M. Neumann, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: North American Association for Computational Linguistics (NAACL), 2018.
[18] Y. Si, J. Wang, H. Xu, K. Roberts, Enhancing Clinical Concept Extraction with Contextual Embeddings, J. Am. Med. Inform. Assoc. 26 (2019) 1297–1304.
[19] C.P. Langlotz, L. Meininger, Enhancing the Expressiveness and Usability of Structured Image Reporting Systems, Proceedings of the AMIA symposium, 2000, pp. 467–471.
[20] C.P. Langlotz, RadLex: A New Method for Indexing Online Educational Materials, RadioGraphics 26 (2006) 1595–1597.
[21] T. Kudo, MeCab, http://taku910.github.io/mecab/.
[22] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Measur. 20 (1960) 37–46.
[23] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (1971) 378–382.
[24] A. Savkov, J. Carroll, R. Koeling, J. Cassell, Annotating patient clinical records with syntactic chunks and named entities: the Harvey Corpus, Lang Resour. Evaluat. 50 (2016) 523–548.
[25] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, A. Setzer, et al., Semantic Annotation of Clinical Text: The CLEF Corpus, in: Proceedings ofbuilding and evaluating resources for biomedical text mining: workshop at LREC, 2008.
[26] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, et al., Building a semantically annotated corpus of clinical texts, J. Biomed. Inform. 42 (2009) 950–966.
[27] S. Mayhew, D. Roth, TALEN: Tool for Annotation of Low-resource Entities, ACL System Demonstrations, 2018.
[28] T.K. Sang, J. Veenstra, Representing Text Chunks. Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, 1999, pp. 173–179.
[29] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural Architectures for Named Entity Recognition, Proc. NAACL-HLT 2016 (2016) 260–270.
[30] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, ICLR, 2013.
[31] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, A Next-generation Hyperparameter Optimization Framework, Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.
[32] N. Srivastava, G. Hinton, A. Krizhevsky, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[33] Y. Wang, Annotating and Recognising Named Entities in Clinical Notes, Proceedings ofthe ACL-IJCNLP 2009 Student Research Workshop, 2009, pp. 18–26.
[34] I. Jauregi Unanue, E. Zare Borzeshi, M. Piccardi, Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition, J. Biomed. Inform. 76 (2017) 102–109.
[35] E. Alsentzer, J.R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, et al., Publicly Available Clinical BERT Embeddings, Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78.
[36] A.E.W. Johnson, T.J. Pollard, L. Shen, L.-W.H. Lehman, M. Feng, M. Ghassemi, et al., MIMIC-III, a freely accessible critical care database, Sci. Data 3 (2016).