Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4565–4572
Marseille, 11–16 May 2020

Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge: Starting From Critical Lung Diseases

Shuntaro Yada,[1*] Ayami Joh,[1*] Ribeka Tanaka,[2*] Fei Cheng,[2] Eiji Aramaki,[1] Sadao Kurohashi[2]
[1]Institute for Research Initiatives, Nara Institute of Science and Technology (NAIST)
8916-5, Takayama-cho, Ikoma, Nara 630-0192, Japan
{s-yada, ajoh, aramaki}@is.naist.jp
[2]Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{ribeka.tanaka, feicheng, kuro}@i.kyoto-u.ac.jp

Abstract

Applying natural language processing (NLP) to medical and clinical texts can bring important social benefits by mining valuable information from unstructured text. A popular application for that purpose is named entity recognition (NER), but the annotation policies of existing clinical corpora have not been standardized across clinical texts of different types. This paper presents an annotation guideline aimed at covering medical documents of various types such as radiography interpretation reports and medical records. Furthermore, the annotation was designed to avoid burdensome requirements related to medical knowledge, thereby enabling corpus development without medical specialists. To achieve these design features, we specifically focus on critical lung diseases to stabilize linguistic patterns in corpora. After annotating around 1100 electronic medical records following the annotation scheme, we demonstrated its feasibility using an NER task. Results suggest that our guideline is applicable to large-scale clinical NLP projects.

Keywords: Electronic medical records, Medical annotation, Medical health record, Named entity recognition, Radiography interpretation report

1. Introduction

Electronic medical records (EMR) are now replacing paper-based records in hospitals. Correspondingly, natural language processing (NLP) techniques in the medical field have radically gained in importance. Current NLP in medical fields, however, particularly addresses fundamental tasks, such as named entity (NE) recognition, NE normalization (also known as *ontology linking*), and event factuality detection. These tasks have been investigated individually, thereby yielding less-standardized task definitions across different studies. Table 1 portrays a summary of existing corpora. As the table shows, the corpora are designed for an idiosyncratic task. Only task-specific annotation was done for each task. For example, the i2b2 Deid corpus includes only personal health information (PHI).

The definition (range) of tags also presents numerous variations. In medical NE recognition (NER) tasks, for instance, *disease names* are a popular NE category. The definition of ‘diseases,’ however, varies substantially along with downstream applications. On the one hand, disease names refer strictly to diagnoses of patients (Wakamiya et al., 2019; Patel et al., 2018). On the other hand, disease names encompass not only diagnoses, but also symptoms and even complaints (Morita et al., 2013; Aramaki et al., 2014).

Several corpora have been designed for general purposes covering multiple categories. An example is a Japanese medical corpus (Aramaki et al., 2009) consisting of date-time information (TIMEX), diseases, medication (drug names & dosages), test names, and test values, as well as the relation between drug names and their adverse effects. Because this complex corpus covers multiple NE categories, the annotation cost for 435 discharge summaries was extremely high (over 10,000 USD; during two years). In addition, a French medical corpus, QUAERO (Suominen et al., 2013) uses UMLS[3] semantic categories consisting of anatomy, chemical and drugs, devices, disorders, geographic areas, living organisms, objects, phenomena, physiology, and procedures. Such UMLS-based classification requires deep knowledge encompassing the entire UMLS ontology.

Similarly, for event factuality detection, which aims at assessment of whether a clinical event occurred or not, task setups range from simple binary classification, i.e., POSITIVE or NEGATIVE adopted in (Chapman et al., 2001; Savova et al., 2010; Morita et al., 2013), to more complex five-way classification comprising NEGATION, PURPOSE, SUSPICIOUS OF, POSSIBLE, and OTHER (Aramaki et al., 2009).

To characterize the matter succinctly, no standard annotation scheme exists, especially for non-English languages. The objective of this research is to propose an annotation scheme that is useful across various tasks in different languages.

A recent study (Patel et al., 2018) has proposed general-purpose annotation guidelines which covers several domains and formats. Most existing schemes including them, however, presume that the annotators have a high level of medical and linguistic knowledge. Considering the high costs and burdens of employing medical staff, we also intend to design a guideline that is feasible for use by people with less clinical expertise.

We specifically examine two difficult diseases that have high death rates in Japan: lung cancer and idiopathic pul-

[*] Equal contributions

[1] `https://mimic.physionet.org`

[2] `https://www.i2b2.org/NLP`

[3] `https://www.nlm.nih.gov/research/umls/index.html`

Table 1: Summary of Related Work (NER, Named Entity Task; REL, relation; SIM, similarity estimation; IE, information extraction; and CODING, (ICD) coding)

Corpus	Language	Task	Description
MIMIC[1]	English	NER	Clinical Text in ICU
2006 – i2b2[2] De-identification & Smoking	English	NER	Discharge Summary
2008 – i2b2 Obesity	English	NER	Discharge Summary
2009 – i2b2 Medication	English	IE	Discharge Summary
2010 – i2b2 Relations	English	REL	Discharge Summary
2011 – i2b2 Coreference	English	REL	Discharge Summary
2012 – i2b2 Temporal Relations	English	REL	Discharge Summary
2014 – i2b2 De-identification & Heart Disease	English	NER	Clinical narratives
2019 – n2c2/OHNLP Clinical Semantic	English	SIM	EHR
NTICR10 MedNLP (Morita et al., 2013)	Japanese	NER (PHI+disease name)	Case Report
NTCIR11 MedNLP2 (Aramaki et al., 2014)	Japanese	NER + CODING	Case Report
NTCIR12 MedNLPDoc (Aramaki et al., 2016)	Japanese	CODING	Discharge Summary
UTH-FX (Aramaki et al., 2009)	Japanese	NER (V.A.)	Discharge Summary
ShARe (Névéol et al., 2014)	English	NER	V.A.
QUAERO (Suominen et al., 2013)	French	NER	Research Paper
Patel et al., 2019	English	NER + REL	Six domains and six report types

monary fibrosis (IPF). This constraint stabilizes linguistic patterns in corpora, and thereby provides a basis for designing a task-independent annotation scheme and makes annotation procedures more independent of workers' clinical knowledge. The resulting annotation might instead awaken us to what domain specificity we relied on afterwards, which is expected to contribute to future domain versatility. In other words, by virtue of target disease fixation, the proposed annotation possesses two characteristics:

Non task-specificity (versatility) – We assume no specified task. We design the annotation for general objectives. To achieve various objectives, we merged the medical categories that have already been proposed. Then we simplify them.

Feasibility without medical knowledge – We specifically examine grammatical clues, which can be inferred even by people who have little medical knowledge. This emphasis enables us to reduce annotation costs, which are often high because of the necessity of employing medical experts.

This paper presents our annotation guidelines, which we applied to actual EMRs written in Japanese, one of the languages with relatively few available medical corpora. The NER results for the annotated corpora demonstrate the quality and feasibility of the proposed scheme.

2. Annotation Schemes and Guidelines

2.1. Format

Corpora annotated according to our guidelines follow the XML standard. Each clinical entity forms XML tags. Some entities have their own attributes. Our basic policy to annotate entities is to mark expressions in a large (long) unit. We apply some rules to annotate expressions as a single entity if the expression comprises several entity types. Entity types such as "diseases and symptoms" and "time expressions" are assigned priority over others. Technical medical terms should be annotated as one certain entity, as mentioned later in Section 2.3. This annotation also enables us to capture frequent compositions. In some domains of EMR, tracking such frequently appearing compositions is more useful for end applications than keeping small entity units. We elaborate on the particular benefits of this policy for each particular entity section below. We carefully designed these tags based on discussion with medical experts and trial annotations on a small set of medical records.

2.2. Entity types

Diseases and symptoms `<D>`: One main entity in medical and clinical texts includes expressions related to diseases and symptoms, such as 'tumors' and 'coughs.' For this entity, we consider its *certainty* as the attribute, i.e., whether the disease or symptom is confirmed to exist in the patient at the time of EMR writing: positive, suspicious, negative, and general. This can describe the existence of diseases and symptoms, which is useful for mining large EMR corpora to generate, e.g., patient medical history without a need for deeper NLP techniques such as dependency parsing or semantic role labeling. The certainty 'general' is applied when the disease or symptom expression is presented merely as a reference and not as actually occurring in the patient. Particularly for the other values (i.e., positive, suspicious, negative), our guidelines tell annotators to examine the predicates of medical descriptions specifically. Details are explained below:

- For a predicate indicating existence, such as "X is recognized," "X is apparent," or "Xs are diffused," assign 'positive'
- For a predicate indicating uncertainty, such as "X is suspicious," "X cannot be denied," " It could be X," or "X should be considered in the differential diagnosis," assign 'suspicious'
- For a predicate indicating nonexistence, such as "X is not recognized," "No X," or "Xs have disappeared," assign 'negative'

When appending this tag to TNM classification of Malignant Tumors, which is a globally recognized standard for classifying the extent of spread of cancer, one can merely use `<D>` tag without its certainty.

Examples:

- There is <D certainty="positive">a mucus plug</D> in <A>the peripheral bronchus</A>.
- From a first phase image, <D certainty="suspicious">an organizing pneumonia</D> is suspected.
- No <D certainty="negative">pleural effusion</D>.
- I explained <D certainty="general">IP</D> by a pamphlet.
- I consider it <D certainty="positive">primary lung cancer</D>. <D>cT1c</D>

Anatomical entities <A>: We label expressions specifying anatomical parts of the body such as 'lung,' 'stomach,' and 'kidney.' Relative expressions such as 'inside,' 'edge,' and 'right under' are also annotated as this entity type. We treat complex words such as 'left third rib' and 'both lung lower superior' as independent entities rather than decomposing them. This treatment helps to keep track of anatomical locations within the entity itself. Otherwise, it might become necessary to conduct dependency parsing to reconstruct the entry.

Examples:

- <D certainty="positive">Small lymph nodes</D> <F>are diffused</F> in <A>the right lung</A>.
- <D certainty="positive">A part of solid nodule</D> is recognized <A>along the verge of the internal cavitation</A>.

Features and measurements <F>: We identify expressions which describe disease characteristics and symptoms such as their values, degrees, ranges, amounts, and sizes. Grammatical clues to annotate this entity type are stem words of predicates and modifiers including verbal nouns.

Examples:

- A <F>well-defined and smooth marginated</F> <D certainty="positive">nodule shadow</D> is recognized.
- A <F>slight</F> <D certainty="positive">shadow</D> is observed.
- <D certainty="positive">Small lymph nodes</D> <F>are diffused</F> in <A>the right lung</A>.

Change <C>: Expressions stating changes of diseases and symptoms, or values of test items and medicines are annotated as this entity type. Expressions such as 'not changed' and 'no remarkable change' are also annotated as this entity type because it is also important information for diagnosis that a certain disease or symptom has not changed. We merge them into <C> (e.g., 'slightly increasing') if feature-describing expressions appear immediately prior. However, even if the term "change" is used, some instances are treated as an expression of a disease or symptom, such as "change after fracture." In addition, even if it represents a change, such as "contractive change," which forms a noun phrase highly related to a disease/symptom, this entity type is not used; we instead use the entity type "diseases and symptoms."

Examples:

- It <C>has increased</C> from <TIMEX3 type="DATE">last time</TIMEX3>.
- <D certainty="negative">The pneumothorax cyst</D> at <A>the left lung apex</A> <C>has disappeared</C>.
- <C>No remarkable change</C> from <TIMEX3 type="DATE">last time</TIMEX3>.

TIMEX3: We adopt TIMEX3, a standard scheme for time expressions (Pustejovsky et al., 2003), for EMR corpora. It has a *type* attribute that distinguishes the kind of time expression: 'date,' 'time,' 'duration,' and 'set.' We extend these *type* attributes by adding clinical context ('cc'), age ('age'), and miscellaneous ('misc') attributes. For cc and age, we deal with 'after surgery' and '48 years old,' respectively, whereas misc encompasses cases in which all other types do not match well but where the entity is apparently a time expression. As the examples below illustrate, "3 months later" and "2 years ago" should be annotated together, whereas "from H35 8/3" is annotated without "from," because the latter is not regarded as a phase. This entity type is not annotated only to time expressions such as "after," but also in units of compound words such as "after surgery" and "after resection."

Examples:

- It <C>has increased</C> from <TIMEX3 type="DATE">last time</TIMEX3>.
- In <TIMEX3 type="TIME">the morning</TIMEX3>, he has <D certainty="positive">difficulty opening</D> his <A>right eyelid</A>.
- <T-key>Smoking</T-key>: <T-val>20 / day</T-val> × <TIMEX3 type="DURATION">40 years</TIMEX3>. He <T-val>quit smoking</T-val> at <TIMEX3 type="AGE">59 years old</TIMEX3>[4]
- Getting a <D certainty="positive">fever</D> <TIMEX3 type="SET">frequently</TIMEX3>
- <TIMEX3 type="CC">Postoperatively</TIMEX3>, he has <D certainty="positive">shortness of breath</D> with stairs or trotting.

Test <T>: This entity comprises three sub-entities: <T-test>, <T-key>, and <T-val>. We refer to them generically as <T>, but we do not allow standalone <T> annotation; annotators must label one of <T-test/key/val>. T-test denotes general names of clinical tests such as 'CT scan.' <T-key> and <T-val> are fundamentally assumed to appear as a pair: a test entry and its measured value, respectively. Sometimes one <T-key> has several <T-val>s. In addition, <T-test> has the *state* attribute for distinguishing whether the test is conducted: 'executed,' 'scheduled,' 'negated' (not conducted), and 'other.' This attribute is used

[4] "Smoking" is a clinical interview items to be attached <T-key> tags. "20 / day × 40 years" and "He quit smoking at 59 years old" are results to be assigned <T-val> tags. The <T-val> tag, however, is added to the remaining words after adding <TIMEX3> tag according to the basic policy in our annotation guidelines, i.e., "a nested structure is not allowed" and "TIMEX3 is assigned priority over other tags."

based on the document time. The state 'other' is applied to exceptional cases. For the decision of the *state*, the SOAP structure in medical records is useful as a hint.
Examples:

- <T-test state="executed">Chest CT </T-test>
- <T-key> FEV1 </T-key>: <T-val> 1.97L (80.0%) </T-val>

Medicine <M>: Similarly to <T>, two tags constitute this entity, which represents medications: <M-key> and <M-val>. The former is used for labeling medicine names. The latter is applied to the amount of the corresponding <M-key> medicine dosage. <M-key> has the *state* attribute with the same specification as that of <T-test>. When multiple state-related expressions appear around medications within a sentence, the last applicable *state* attribute is applied.
Examples:

- Planning to dose him up with <M-key state= "scheduled">Pulmocare</M-key> <M-val>100mL/1hr</M-val> from <TIMEX3 type="DATE">tomorrow</TIMEX3>.
- At <TIMEX3 type="DATE">2027/7</TIMEX3>, <M-key state="negated">Pirespa</M-key> <M-val>1200 mg</M-val> had been caused <D certainty="positive">anorexia</D>, so it was abandoned. (In this example, add the state attribute of "negated" because the final state of medication is "abandoned.")

Remedy <R>: We annotate the expressions of medical treatments (e.g., 'defibrillation' and 'abscission') using this entity type. It has *state* attributes similar to <T-test/key> and <M-key>. We can differentiate <R> from <T> and <M> by non-accompaniment of 'values.'
Examples:

- <T-test state="executed">CAG </T-test> was implemented but <R state= "negated">PCI</R> was not performed.
- <TIMEX3 type="cc">After operation </TIMEX3> of <R state="executed">resection of superior lobe of left lung in the region of chest</R>

Clinical Context <CC>: In EMR, the state of patients in relation to the hospital is described frequently, such as 'rehospitalization' and 'follow up.' We also apply the *state* attribute to this tag.
Examples:

- He <CC state="executed">was going</CC> to the cardiovascular clinic.
- He is going to <CC state="scheduled">be discharged from the hospital</CC> at <TIMEX3 type="DATE">this weekend (8/27)</TIMEX3>.

Pending <P>: Annotators sometimes tentatively find some expressions that are apparently clinical terms, such as abbreviations (e.g., " 'LNs' are found"). We allow annotators to label such expressions using this <T> tag to skip, if confident references are not found. After the annotation, these pending expressions can be resolved with assistance by someone with medical expertise.
Examples:

- <A>Directly under the pleura</A> was <P>spared</P>.
- <A>Within the mediastinum</A>, <P>LNs</P> up to <F>minor axis 4cm</F> are recognized.

2.3. Annotation Procedures

We have developed a systematic annotation procedure that allows even annotators who have little medical knowledge to annotate EMR texts. Fundamentally, the annotation scheme asks annotators to label a span of sentences one by one if they find an entity that matches our definitions presented above. The following guidelines also support annotators to label entities without extensive medical knowledge, while maintaining coherent annotation:

1. nested structures in which another tag is labeled inside one tag are not allowed; in other words, technical medical terms should be annotated as a single entity
2. most informative entity types such as "diseases and symptoms <D>" and "time expressions <TIMEX3>" are assigned priority over others
3. an easy-to-use reference dictionary for diseases and symptoms such as J-MeDic (Ito et al., 2018) is used when annotators are not confident about the exact span of the entity
4. annotators can take a longer span of a single entity if unsure, especially under the case of complex compound words

Although these guidelines might reduce the granularity of annotated entities, it can be controlled after the annotation is completed. It is noteworthy that the granularity, range, and definition of entities depends on the downstream application. We instead assign importance to the ease of labeling for non-medical professionals. Overly finer-grained annotation might impose burdens, especially for such workers.

2.4. Technical comparison to major existing schemes and guidelines

Patel et al. (2018) proposed a similar annotation scheme. Our scheme merges some of their individual tags for simplicity. For example, their scheme splits <D> into "Problem" (i.e., *major* problem) and "Finding" (*minor* problem), which might be difficult to distinguish even with medical knowledge. They might therefore cause incoherent annotation. <T-test> and <T-key> of our scheme corresponds to "lab data," "body measurement," and "medical device" in their definition, which makes sense for two reasons. (a) The granularity of these concepts depends on the end clinical applications. It can be restored afterwards, almost automatically using medical dictionaries or manually by the assessment of clinical professionals with a light load. (b) These concepts appeared much less frequently than the other tags, according to their result.

In addition, the *certainty* attribute we defined enables us to conduct, for example, disease state recognition without conducting syntax or relation annotation, which usually requires much more complex labeling effort than entity tagging. This capability is particularly useful for generating structured data tables from medical documents (Aramaki et al., 2009).

3. Pilot Experiment

We applied our guidelines to actual clinical text. This section presents description of corpus statistics, annotation costs, and results of a preliminary experiments with NER.

3.1. Material

Among the various types of EMR text used in hospitals, this study targets two types: (1) medical records and (2) radiography interpretation reports. We employed the same annotation scheme for them. We are ordering annotation for around 400 medical records and around 1400 radiography reports. Examples of the corpora are presented in Table 2.

Table 3 presents statistics of some already annotated corpora: 156 for medical records and 1000 for radiography reports. The results reflect the report-type characteristics: medical records include rich time expressions and treatment information of patient histories; radiography reports intensively describe diseases of particular anatomical locations with features based on radiographs.

In Table 4, the count of appearing attributes is reported. It also shows the characteristics of the report types. For *certainty* in <D>, medical records are primarily about found disease (positive) whereas radiography reports put as much importance on non-existence of disease (negative) as actual findings (positive). Note that the sum does not equal to the total <D> occurrence because no *certainty* annotation is permitted for TNM notations. For *type* in <TIMEX3>, medical records again shows the rich diversity in time expressions. For *state*, finally, medical records are relatively characterized by more scheduling expressions in contrast to radiography reports.

3.2. Annotation cost

In the Japanese annotation market, for example, annotation by medical expertise costs around 4,000 JPY *per medical record*. Because professional medical knowledge is unnecessary for our annotation guidelines, we can order the annotation to the labeling companies not specialized in clinical text. In fact, our cost is less than the amount (i.e., 3,000 JPY per medical record; 25% cost reduction from 4,000 JPY) required by general workers to do the annotation. Furthermore, the period to annotate 156 medical records by one person was three weeks (83 hours in total), which can be regarded as reasonably short; recall that clinically professional annotations on 435 discharge summaries cost more than 10,000 USD and two years (see Section 1).

3.3. Named Entity Recognition

To estimate corpus quality, we experimentally applied NER to the annotated corpora. We chose radiography interpretation reports for this experiment because of the current availability of a sufficient amount of annotated data. In this experiment, where a trained model labels clinical tags we defined, we skip estimating the attributes of these tags, such as `certainty`, for brevity.[5]

We adopt BERT (Devlin et al., 2019), a state-of-the-art NER model based on transfer learning. Transfer learning in the NLP field refers to a set of methods that explore various pre-training methods to capture contextualized word representations from large-scale raw text data and to generalize (i.e., fine-tune) pre-trained information to target tasks or other text domains. During the past two years, several successful transfer learning models, e.g., ELMo, GPT, and BERT (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019), have been proposed. They have markedly improved the state-of-the-art of widely various NLP tasks.

Next we explain details of the application of the pre-trained Japanese BERT (Shibata et al., 2019) to our NER task. Japanese language uses no space to delimit words in sentences. To segment sentences into words, we first applied morphological analysis[6] (Tolmachev et al., 2018). Japanese BERT further applies the byte-pair encoding model to tokenize words into sub-words. As a pilot experiment, we followed the standard BERT practice to formulate NER as a sequential tagging task. The last-layer representations of sub-words are fed into the NER label classifier. The following fine-tuned hyper-parameters were empirically applied according to the original BERT paper:

Batch size: 16

Number of epochs: 5

Learning rate: 5e-5

We split the 1000 medical records randomly into 80% training data and 20% test data. We applied the 'BIO' position tags for experiments. We compared our BERT classifier to a strong baseline bidirectional LSTM model with CRF constraints (Huang et al., 2015; Lample et al., 2016) for measuring the relative performance. The baseline adopted the word2vec (Mikolov et al., 2013) embedding pre-trained on Japanese Wikipedia, hidden size equal to 256 and batch equal to 64. The optimizer setting is the same as the BERT classifier.

Table 5 presents the results of our BERT-based NER classifier. Our model produced satisfactory performance when recognizing medical tags from medical records. It achieved 95.30 overall F-score, significantly outperforming the baseline in all the types of medical tags. For the most frequently used tags, our model achieved F-scores of 93–96. Both the BERT classifier and baseline achieved high classification performance, which underscored the high quality of our proposed annotation scheme for medical record data. We also observed that some tags, such as <T-key>, <T-val>, <CC>, and <M>, rarely appeared among our 1000 radiography interpretation reports. We assume the tag distribution strongly depends on the types of medical records. Further exploration will present interesting topics when we annotate medical records of other types in future work.

Transfer learning models usually lean towards over-fitting in a small number of fine-tuning epochs, because they are already pre-trained over large steps or epochs. For the BERT classifier, we additionally reports the training losses and test F-scores over fine-tuning epochs. As shown in Figure 1, the epoch loss is continuously dropping during train-

[5] One solution to attribute-level NER is to train another model that predicts attributes from the tag-level NER result.

[6] `https://github.com/ku-nlp/jumanpp`

Table 2: Examples of annotated corpora: Medical records & Radiography interpretation reports

Medical records	Radiography interpretation reports
S) When exercising, can not walk long, breathing. Want to be cured and go home early O) Vital signs (blood pressure: 126/84, pulse rate: 78 / min, respiratory rate: 16 / min), chest breathing, increased activity of respiratory assist muscles, auscultation to listen to a rale in S2, 6.10, 6 min walk test: 240 m, MRC: Grade 3, F-H-J Classification: Ⅲ degree Summary: Breathing difficulty, respiratory muscle weakness, and inefficient breathing caused dyspnea. Walking distance decreased. Short-term goal (after 2 weeks): Reducing dyspnea by reducing dryness and learning breathing (modified Borg scale 7 to 4). Long-term goal (1 month later): Further reduction of dyspnea (modified Borg scale 4 to 2), extending walking distance and home discharge P) (1) Exclusion training, (2) Breathing assistance method, (3) Breathing method guidance at rest and walking, (4) Voluntary training guidance	It will be taken asymptomatically. After 6 months, glomeruli will improve neoplastic lesions percutaneously. Abdominal pain. Diarrhea. CT scan showed uneven contrast in the bladder, large intestine, gallbladder artery (SMV), and ascending colon. The gallstone caused deterioration of the general condition. Clean water exchange and high-dose infusion were started. Subsequently, emergency surgery was performed on lower left extremity stenosis on the day of lower stomach hypogastric. During the surgery, hemorrhage symptoms disappeared. Aggravated dyspnea was recognized by administering aftinib on the first day of steroid treatment during surgery. On day 33 of disease, superior mesenteric vein splenic deterioration was confirmed by contrast-enhanced CT.

Table 3: Corpus statistics

	Medical records	Radiography reports
Total documents annotated	156	1000
Average sentence count per document	30.89	13.36
Average word count per document	268.73	142.02
Total tag count		
Disease	2008	13897
Anatomical	742	7123
Feature	325	5345
Change	678	1100
TIMEX3	1820	1550
T-Test	716	852
T-Key	1957	40
T-Val	2116	3
M-Key	399	0
M-Val	170	0
Remedy	439	137
Clinical Context	331	28

ing while obtaining close to the best F-score 95.30 in the 4th and 5th epochs. To be clarified, the best epoch setting should be estimated based on the validation set in practice. However, as a preliminary experiment, we directly reported the test F-scores. The results suggested the model did not over-fit the training data in 5 epochs.

4. Conclusion

We proposed an annotation guideline for medical and clinical texts with two crucially important benefits: capability of supporting various downstream NLP tasks, and feasible use without medical expertise of annotators. The proposed annotation comprises two levels of clinical information: (1) the Named Entities (NE) level, i.e., disease names, time expressions, and clinical verbs; and (2) modality level, i.e., positive vs. negative, suspicious, and general NE factuality. This paper also described preliminary results of NE Recognition. Its performance was satisfactorily high, supporting the quality of our annotation guidelines. This annotation is expected to be widely applicable to large-scale medical NLP research. For future studies, we plan to publish the corpus with the proposed annotation scheme.

We specifically examined lung cancer and IPF as target diseases in the corpora. As one future direction of study to enhance its versatility, we expect to extend our scheme to other diseases. We intend to analyze other clinical corpora including our own to obtain greater insight into means of extending the scheme.

5. Acknowledgments

This work was supported by JST AIP-PRISM Japan Science and Technology Agency. We thank every member of this project for constructive advice on clinical text interpretation. We also thank Dr. Kaoru Ito (Kyushu University)

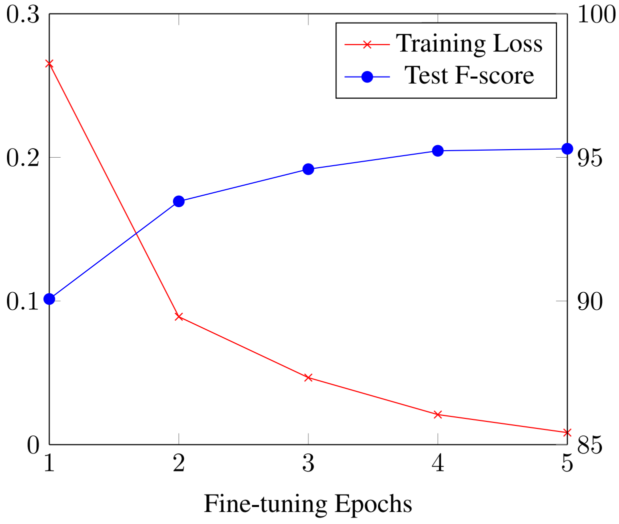
Table 4: Attributes statistics.

	Medical records	Radiography reports
Disease: *certainty*		
positive	1407	5941
negative	258	5491
suspicious	283	1155
general	60	-
TIMEX3: *type*		
DATE	1378	1187
DURATION	94	3
CC	144	358
TIME	142	-
AGE	34	-
SET	28	-
MISC	-	1
Test/Medicine/Remedy/CC: *state*		
executed	1091	870
scheduled	259	2
negated	36	1
other	100	144

Table 5: NER results of our BERT-based classifier. '**Baseline**' denotes the F-score of the baseline system.

Tag	**Precision**	**Recall**	**F-score**	**Baseline**
Diseases	95.90%	96.83%	**96.36**	95.44
Anatomical	95.08%	95.93%	**95.50**	94.21
Features	92.48%	94.77%	**93.61**	93.09
Change	88.56%	91.67%	**90.09**	89.73
TIMEX3	95.22%	97.39%	**96.30**	95.27
T-Test	94.80%	95.35%	**95.07**	92.13
T-Key	66.67%	100.00%	**80.00**	66.67
T-Val	-	-	-	-
Remedy	81.48%	81.48%	**81.48**	63.64
Clinical Con.	83.33%	71.43%	**76.92**	44.44
Overall	94.65%	95.95%	**95.30**	94.26

Figure 1: Training loss and test F-score over fine-tuning epochs.



for cooperating in the design of our annotation scheme and for writing a Japanese draft of the guideline. Finally, we acknowledge Faith Mutinda's assistance with proofreading and with advising us on related work.

6. Ethics statement

The ethics committee of Osaka University Hospital, Kyoto University, National Cancer Center Hospital, and NAIST approved this research, including the provision of proxy consent. Consent was obtained from a proxy when a patient lacked the autonomous competence to provide consent, in accordance with the Ministry of Health, Labour and Welfare's Ethical Guidelines for Medical and Health Research Involving Human Subjects.

7. Bibliographical References

Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Mashuichi, H., and Ohe, K. (2009). TEXT2TABLE: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the BioNLP 2009 Workshop*, pages 185–192, Boulder, Colorado, June. Association for Computational Linguistics.

Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T. (2014). Overview of the NTCIR-11 MedNLP-2 task. In *In Proceedings of the 11th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, Tokyo, Japan, June. Natinal Institute of Informatics.

Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T. (2016). Overview of the NTCIR-12 MedNLPDoc task. In *In Proceedings of the 12th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, Tokyo, Japan, June. Natinal Institute of Informatics.

Chapman, W., Bridewell, W., Hanbury, P., Cooper, G., and Buchanan, B. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34:301–310, November.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Ito, K., Nagai, H., Okahisa, T., Wakamiya, S., Iwao, T., and Aramaki, E. (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*, pages 2365–2369, Miyazaki, Japan, May. European Language Resource Association.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Morita, M., Kano, Y., Ohkuma, T., Miyabe, M., and Aramaki, E. (2013). Overview of the NTCIR-10 MEDNLP task. In *In Proceedings of 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, Tokyo, Japan, June. Natinal Institute of Informatics.

Névéol, A., Grouin, C., Leixa, J., Rosset, S., and Zweigenbaum, P. (2014). The Quaero french medical corpus: A resource for medical entity recognition and normalization. In *In Proceedings of Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, pages 24–30.

Patel, P., Davey, D., Panchal, V., and Pathak, P. (2018). Annotation of a large clinical entity corpus. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2033–2042, Brussels, Belgium, November. Association for Computational Linguistics.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G., and Radev, D. (2003). TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the Fifth International Workshop on Computational Semantics*, pages 337–353, Tilburg, Netherlands, January. Association for Computational Linguistics.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *Preprint*.

Savova, G., Masanz, J., Ogren, P., Zheng, J., Sohn, S., Kipper-Schuler, K., and Chute, C. (2010). Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17:507–513.

Shibata, T., Kawahara, D., and Kurohashi, S. (2019). Improvement of Japanese sentence parsing with BERT. In *Proceedings of the Japanese Annual Conference on Natural Language Processing*, pages 205–208, Nagoya, Japan. Association for Natural Language Processing. (in Japanese).

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., Pradhan, S., South, B. R., Mowery, D. L., Jones, G. J. F., Leveling, J., Kelly, L., Goeuriot, L., Martinez, D., and Zuccon, G. (2013). Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In Pamela Forner, et al., editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231, Valencia, Spain, September. CEUR Workshop Proceedings.

Tolmachev, A., Kawahara, D., and Kurohashi, S. (2018). Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium, November. Association for Computational Linguistics.

Wakamiya, S., Morita, M., Kano, Y., Ohkuma, T., and Aramaki, E. (2019). Tweet classification toward Twitter-based disease surveillance: New data, methods, and evaluations. *Journal of Medical Internet Research*, 21(2):e12783.