Is boundary annotation necessary? Evaluating boundary-free approaches to improve clinical named entity annotation efficiency

Gabriel Herman Bernardim Andrade, Shuntaro Yada, Eiji Aramaki

Submitted to: JMIR Medical Informatics
on: April 19, 2024

Table of Contents

Is boundary annotation necessary? Evaluating boundary-free approaches to improve clinical named entity annotation efficiency

Gabriel Herman Bernardim Andrade[1] MSc; Shuntaro Yada[1] PhD; Eiji Aramaki[1] PhD

[1]Graduate School of Information Science Nara Institute of Science and Technology Ikoma JP

Corresponding Author:
Eiji Aramaki PhD
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5, Takayama-cho
Ikoma
JP

Abstract

Background: Named Entity annotation is a fundamental task in Natural Language Processing. However, it has associated challenges. Especially in the clinical domain, determining entity boundaries is one of the most common sources of disagreements between annotators, as questions such as whether modifiers or peripheral words should be annotated appear. If unresolved, these can induce inconsistency in the produced corpora, but, on the other hand, strict guidelines or adjudication sessions prolong what is already a slow and convoluted process.

Objective: To address these challenges, we evaluate two novel annotation methodologies: Lenient Span and Point annotation, aiming to relieve the pain of precisely determining entity boundaries.

Methods: We evaluate their effects through an annotation case study on a dataset of Japanese medical case reports. We compare annotation time, annotator agreement, and the quality of the produced labeling and assess the impact on the performance of a NER system trained on the annotated corpus.

Results: We saw significant improvements in the labeling process efficiency, with up to a 25% reduction in overall annotation time and even an 8% improvement in annotator agreement compared to the traditional boundary-strict approach. However, even the best-achieved NER model presented some drop in performance.

Conclusions: Our findings demonstrate a balance between annotation speed and model performance. Although disregarding boundary information affects model performance to some extent, this is counterbalanced by significant reductions in the annotator's workload and notable improvements in the speed of the annotation process. These benefits may prove valuable in various applications, offering an attractive compromise for developers and researchers.

(JMIR Preprints 19/04/2024:59680)
DOI: https://doi.org/10.2196/preprints.59680

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Gabriel Herman Bernardim Andrade[1], MSc. <herman_bernardim_andrade.hi1@is.naist.jp>
Shuntaro Yada[1], Ph.D. <s-yada@is.naist.jp>
Eiji Aramaki[1], Ph.D. <aramaki@is.naist.jp>

[1] Social Computing Lab, Graduate School of Science and Technology, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma, Nara, Japan, 630-0192

Corresponding author: Eiji Aramaki, +81-743-72-5250, aramaki@is.naist.jp

Is boundary annotation necessary? Evaluating boundary-free approaches to improve clinical named entity annotation efficiency

Abstract

Background: Named Entity Recognition (NER) is a fundamental task in Natural Language Processing. However, it is typically preceded by Named Entity annotation, which poses several challenges, especially in the clinical domain. For instance, determining entity boundaries is one of the most common sources of disagreements between annotators due to questions such as whether modifiers or peripheral words should be annotated. If unresolved, these can induce inconsistency in the produced corpora, yet, on the other hand, strict guidelines or adjudication sessions can further prolong an already slow and convoluted process.

Objective: To address these challenges, we evaluate two novel annotation methodologies, *Lenient Span* and *Point annotation*, aiming to mitigate the difficulty of precisely determining entity boundaries.

Methods: We evaluate their effects through an annotation case study on a Japanese medical case report dataset. We compare annotation time, annotator agreement, and the quality of the produced labeling and assess the impact on the performance of an NER system trained on the annotated corpus.

Results: We saw significant improvements in the labeling process efficiency, with up to a 25% reduction in overall annotation time and even a 10% improvement in annotator agreement compared to the traditional boundary-strict approach. However, even the best-achieved NER model presented some drop in performance compared to the traditional annotation methodology.

Conclusions: Our findings demonstrate a balance between annotation speed and model performance. Although disregarding boundary information affects model performance to some extent, this is counterbalanced by significant reductions in the annotator's workload and notable improvements in the speed of the annotation process. These benefits may prove valuable in various applications, offering an attractive compromise for developers and researchers.

Keywords: natural language processing; named entity recognition; information extraction; text annotation; entity boundaries; lenient annotation; case reports.

Introduction

The Electronic Health Record (EHR) can be an important source of data for health-related research as it contains information on a patient's condition and complaints, performed procedures and administered drugs, the outcome of the treatment, and more [1].

Clinical narratives are a fundamental part of EHRs. Due to their free and unstructured format, Natural Language Processing (NLP) methods are essential for extracting the information from such documents in a way that is comprehensible and useful for computer systems. Although machine learning-based NLP systems can achieve high performance, these often require large amounts of in-domain annotated data for proper training [2]. Recent few-shot approaches empowered by Large Language Models (LLMs) have also been shown to be performant. Yet, these can also benefit from fine-tuning with in-domain examples, yielding notable improvements [3].

Named Entity (NE) annotation, as an inherently manual process, allied to the sheer volume of data that must be meticulously labeled to produce an accurate model, makes it an exhausting and time-consuming task [4]. Particularly when annotating clinical data, workers must possess not only linguistic understanding, but specialized medical knowledge is also required. Recruiting such capable workforce can make the process rather costly [5].

Furthermore, annotation is accompanied by a set of practical issues. For instance, it is natural that contributors disagree on how certain information is annotated or even whether it should be annotated [6]. Determining entity boundaries, meaning where a concept starts and ends, is one of the primary sources of conflict during the process, as so-called *boundary words*, such as articles or adjectives, can induce ambiguity [7].

Especially in medical texts, it is common for annotators to be unsure whether adjectives or modifiers should be included in the annotation. For example, in the sentence presented in Figure 1, some may annotate only the core symptom ("inflammation"). Conversely, others may consider adding all modifiers necessary for a complete encapsulation of the condition.

While entity boundary definition is a problem that affects all languages, scriptio continua languages (which do not have spaces between words), such as Japanese, Chinese, and Korean, are particularly impacted due to the increased difficulty in separating concepts and modifiers.

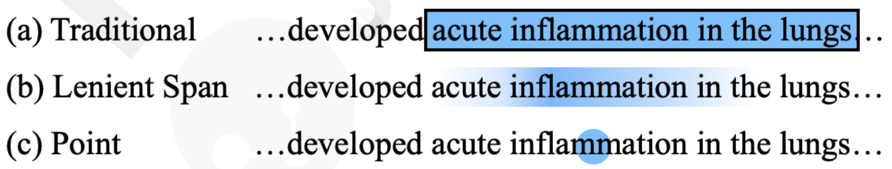

(a) Traditional
…developed acute inflammation in the lungs…
(b) Lenient Span
…developed acute inflammation in the lungs…
(c) Point
…developed acute inflammation in the lungs…

Figure 1. Example of different annotation paradigms. Traditional annotation (a) requires precisely labeling the beginning and end of the span, while boundary-free (b and c) methods focus on only identifying the core term.

One can employ strict annotation guidelines to delineate precisely how information should be annotated and even implement adjudication sessions to resolve disagreements. Yet, this can increase the workload and complexity of an already slow and convoluted process.

As an alternative to mitigate such issues, we propose to reformulate the annotation task by eliminating the need to define specific span boundaries when annotating an NE. By demanding less precision from the annotators, we expect to minimize the required decision-making during labeling, thus improving annotation speed and relieving conflicts.

Although this approach may reduce annotation quality, Named Entity Recognition (NER) performance should not be significantly impacted, as previous research found that models are resilient to a certain amount of boundary imprecision in their training data [8].

In this paper, we leverage this phenomenon by introducing two *boundary-free* annotation methodologies: *Lenient Span*, which relieves the emphasis on entity boundary precision, and *Point*, which uses a single position to represent the annotation. Figure 1 presents a visual comparison between the methodologies. We performed a case study to evaluate the efficiency of the proposed methods when annotating a corpus of Japanese medical case reports to create training data for an NER system.

Our contributions are summarized as follows: (1) We present two novel boundary-free annotation methodologies. (2) We evaluate the efficiency of the annotation process by metrics of annotation time and annotator agreement. (3) We analyze the impact on the performance of a NER system trained with annotated corpora.

Related Work

Annotation Efficiency Improvements

Attempts to improve the annotation workflow are a common theme in NER-related research.

Pre-annotation depicts the automatic labeling of the text prior to the annotator work [9]. This technique can not only reduce the required annotation time and workload required but also minimize errors [10]. Active Learning (AL) [11] can further optimize automatic labeling by iteratively incorporating the data produced during the annotation process to re-train the pre-annotation model. Kholghi et al. [12] ascertained that AL reduced the annotation time by up to 35% during experiments.

While these are well-established approaches, recent studies also explore alternative ideas. Tokunaga et al. [13] analyzed eye-tracking data during NE annotation to identify characteristics that can help design effective features for an annotation tool. Saxena et al. [14] introduced a hybrid search-enhanced software that allows users to look for similar terms and annotate related information simultaneously, shortening work time by around 30% compared to standard tools.

In recent years, generative large language models have transformed NLP research and applications, becoming state-of-the-art NLP techniques. While the potential of LLMs to improve the text annotation workflow has also been evaluated in a few different studies [15–17], Tan et al. [18] point out that their effectiveness is still strongly affected by model hallucinations and the gap in performance between proprietary and open-source LLMs.

Although crowdsourcing platforms allow the convenient annotation of vast amounts of data [19], they do not improve task execution or reduce the workload of an individual worker. In addition, as Snow et al. [20] noted, inconsistent or low-quality annotations require effective quality control measures. Li [21] found that LLMs can be used to improve the quality of annotation generated by crowdsourcing. Yet LLM annotation quality is still shy of what can be produced manually; thus, combining the automated technique and human effort is still the best approach to creating a high-quality dataset [22].

Entity Boundary Imprecision

When addressing boundary imprecision, most studies regard it as a form of noise that should be corrected or circumvented. For instance, Liu et al. [23] use confidence scores and normalization techniques based on the labeling structure to estimate the correct span.

Zhu et al. [7] introduced a boundary regularization technique, redistributing a portion of the probability assigned to an annotated span to its neighboring words. This process produces a smooth transition between entity annotations and their non-entity surroundings, mitigating annotation boundary inconsistencies.

Shem et al. [24] propose the NER task as a boundary-denoising diffusion process, where a model is trained to derive precise NEs from noisy spans. The authors added controlled noise to gold entity boundaries and used the imprecise data to teach a model to apply a reverse diffusion process to recover the original entity boundaries.

On the other hand, Andrade et al. [8] identified that imprecise boundary annotation may not have an extensive impact in some applications. The authors evaluated the effect of various levels of imprecise boundary annotation on NER and Entity Linking. They identified that models are resilient to a certain amount of noise, showing a small performance drop in that range.

Methods

Dataset

We used the MedTxt-CR-JA corpus [25] in our experiments. This dataset comprises 148 open-access case reports in Japanese. Table 1 presents an example document from the dataset.

A case report is a detailed description of a patient’s medical condition, containing, among other information, the temporal progression of the disease and its treatment. Its format is similar to a discharge summary and is frequently used in medical NLP, such as in MIMIC-III [26] or n2c2 shared tasks [27].

Table 1. Example of a case report from MedTxt-CR-JA and its English translation.

□□□□□□□ □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□ □□ □□ □□ □□□□□□□□□□□□□□□□□□□□□□□ □□
A 58-year-old female. The patient visited this hospital due to the appearance of a skin rash which worsened about 2 weeks before her first visit. At the initial examination, the patient had extensive edematous erythema on her torso and extremities, with forming blisters. Keratinized erythema was uniformly observed around the joints on the back side of the fingers, erythema and purpura were observed around the nails, and mild purplish-red spots were observed around the eyes. At this point, there were no abnormalities other than mildly elevated CPK and LDH and 20-

fold increase in antinuclear antibodies. Consequently, follow-up with a topical steroid ointment was carried out. However, by the third week, the skin rash on the torso and extremities changed to keratotic red plaques, and edematous erythema of the upper eyelids became prominent by approximately the first month and became a typical rash. The patient died 1 year and 2 months after the onset of illness due to complications of lung cancer. Based on the clinical history, the erythema multiforme or eczema-like skin rash seen at the time of the initial examination is considered to be an early-stage skin rash of dermatomyositis.

This corpus was used in previous studies [28] and contains pre-existing annotations for diseases and symptoms names, drugs, anatomical parts, etc. Although we discarded these labels for our experiments, we use them as a gold standard (GS) for evaluation purposes. From now on, this set of annotations will be identified as *Gold Standard Corpus* (GSC).

We randomly selected a subset of 100 documents from the full corpus, referred to from now on as the *dataset*. To minimize the difference in difficulty between texts, we selected documents with similar lengths and quantity of gold-standard entities. Texts are, on average, 554 characters long, roughly equivalent to 250-300 English words, containing around ten entities per text.

Even though the set of documents for annotation may be considered small, it's worth noting that a scenario with such a small amount of data is not uncommon in the clinical setting, where strong data restrictions usually limit the amount of data available to work with [29].

Annotation Guidelines

It is common to define a set of guidelines before an annotation process to minimize the divergences between annotators and guarantee consistency.

We followed the annotation schema as defined by [30]. To simplify the evaluation process, we considered the annotation of only one entity type in this. Annotators were asked to label only positive (non-negated) entities of the "Diseases and symptoms" category. We provided the participants with a document describing what should be annotated and some examples, as summarized in Table 2.

Table 2: Annotation Guidelines. In the examples, entities that should be annotated are marked in bold.

What to annotate	
Description	**Examples**
Reported symptoms, disease names, and clinical findings (pathology, CT, and other images)	• Patient visited this hospital due to the appearance of a **skin rash.**
Clinical suspicion, even if there is a slight possibility of disease occurrence	• **Epicarditis** was *suspected* and the patient was hospitalized on July 2.
The locus of a condition, such as an anatomical structure or location, body substance, or physiologic function	• Abdominal CT scan revealed **many enlarged intra-abdominal lymph nodes**.
Adjectives and other modifier words that alter the characteristics or intensity of a condition	• Patient had no subjective symptoms other than a **high fever**. • There was **spotty necrosis** in the lobules.

What should not be annotated	
Description	**Examples**
Absence of symptoms or diseases. Basically, a negation of a clinical concept	• Abdominal findings were unremarkable. • The rash disappeared in about two months.
General Discussion of a condition merely as a reference and not as a clinical finding	• There is a possibility of primary biliary cholangitis when elevated hepatobiliary enzymes are detected.
Numeric or qualitative findings of an investigation, such as laboratory test values	• The measured blood pressure was abnormal.

Annotation Methodologies

Our goal is to evaluate whether relieving the emphasis on entity boundary improves annotation speed while maintaining the overall quality of the produced labels. Thus, we compared the *Traditional* (boundary-strict) annotation method against two proposed boundary-free approaches: *Lenient Span* and *Point annotation*. Figure 1 presents a comparative example of each annotation method.

Traditional annotation requires precise annotation of each NE's exact start and end positions.

Lenient Span annotation introduces flexibility to the annotation boundaries. While the annotation is still composed of a span, start and end positions are not required to be exactly aligned with the NE boundaries.

Point annotation Unlike span-based paradigms, this method requires selecting a single point at any position within the NE span without explicitly specifying the span. It prioritizes speed and simplicity in scenarios where it is not straightforward to determine the NE span precisely. On the other hand, it may introduce ambiguity in the information captured by the annotation.

Note on LLM Annotation

While the use of generative LLMs for text annotation is gaining traction, in this work, we seek ways to aid human annotation and reduce the necessary effort as much as possible where LLMs cannot be used.

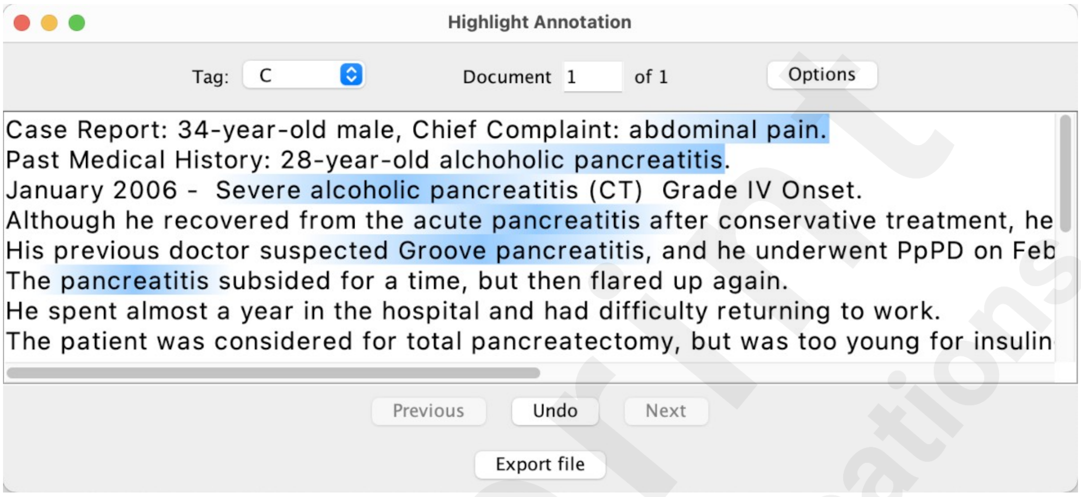
The employment of LLMs still raises concerns about privacy and security issues, as due to the necessary infrastructure and computational power needed, these models are usually held in the cloud and owned by third-party companies [31]. Given the sensitive nature of clinical data, the usage of LLMs in NLP tasks on real-world data is usually constrained by the policy of medical institutions. Thus, there is still a need for manual annotations until performant medical LLMs can be accessed through a secure private network or hosted inside hospital facilities at a reasonable cost.

Annotation Task

We asked four annotators with medical background and different levels of annotation experience to participate in the experiments. They produced three annotated corpora by labeling the documents from the dataset using each evaluated methodology. We measured the time taken for each annotation session and computed agreement metrics. We then used each produced corpora to fine-tune a BERT-based (Bidirectional Encoder Representations from Transformers) [32] NER system and evaluated its performance to assess the corpora quality.

Annotation Tool Development

We developed a Java-based annotation tool to support the proposed boundary-free approaches [33]. Annotations can be presented with smoothed edges using a gradient of color to represent a *soft boundary* and encourage the annotators to be less meticulous when marking the boundaries of the concept. Figure 2 shows a screenshot of the main annotation window.



Case Report: 34-year-old male, Chief Complaint: abdominal pain.
Past Medical History: 28-year-old alchoholic pancreatitis.
January 2006 - Severe alcoholic pancreatitis (CT) Grade IV Onset.
Although he recovered from the acute pancreatitis after conservative treatment, he
His previous doctor suspected Groove pancreatitis, and he underwent PpPD on Feb
The pancreatitis subsided for a time, but then flared up again.
He spent almost a year in the hospital and had difficulty returning to work.
The patient was considered for total pancreatectomy, but was too young for insulin

Figure 2. Screenshot of the Annotation Tool.

The text is displayed in its original style, keeping line breaks, spacing, and special characters. Since there is no pre-tokenization of the texts, annotators can select text spans with character-level precision.

The tool has two modes to annotate a concept:

Click and drag The user clicks on the location where the concept begins and drags the mouse up to where it ends. After releasing the mouse, the area becomes highlighted, representing the labeling.

Click-only The user clicks on an entity to label it. While the annotation is stored as a single point, the position will be expanded to a *simulated* span on the interface, representing approximately the labeled concept, as shown in Figure 3.

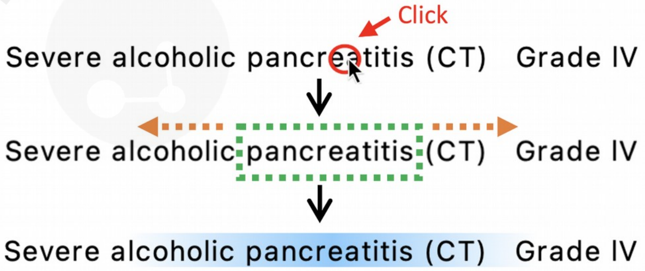

Click
Severe alcoholic pancreatitis (CT) Grade IV
Severe alcoholic pancreatitis (CT) Grade IV
Severe alcoholic pancreatitis (CT) Grade IV

Figure 3. Example of a *click-only* annotation. The selected position, represented by the red circle, is expanded to the word boundaries (in green) plus a random span (orange arrows).

The annotators received instructions on how to use the tool and a video demonstrating the annotation of a document (available in the Multimedia Appendix). They were also supplied with ten test documents to familiarize themselves with the tool.

Labeling Workflow

The annotation was conducted in three sessions. In the first two sessions, each annotator was assigned two sets of 50 documents to work on using the boundary-free methodologies.

To minimize the number of times each annotator would annotate the same document yet allow us to have at least two sets of annotations for a given methodology, we divided our dataset of 100 documents into four splits.

For each annotation session, each participant received a file containing two splits and the annotation methodology that should be used (totaling 50 documents per annotator), as presented in Table 3. We attempted to maximize the mixing between the annotator and the methodology used.

The work was executed in three different sessions, the first for Point Annotation, followed by the Lenient Span annotation, and lastly, the Traditional annotation. During the first two sessions, the annotation tool was configured to show smooth edges, and annotators were instructed not to fix slightly incorrect annotations as long as the core concept was highlighted in the tool's interface.

Table 3. Data split for crossover experiment design.

Annotator	**Documents**			
	1-25	26-50	51-75	76-100
A	P / T	P	S / T	S
B	S / T	S / T	P	P
C	S	P / T	P / T	S / T
D	P	S	S	P / T

T = Traditional Annotation, S = Lenient Span Annotation, P = Point Annotation

Although the same annotator worked on the same document more than once, the Traditional annotation (third) session was conducted six months later to avoid memory bias affecting annotation time measurement. This time, annotators were instructed to be as precise as possible when selecting the entity spans and not to refrain from undoing incorrect annotations. The annotation tool was configured beforehand to present the annotations with precise hard boundaries, as any other standard annotation software.

Across all sessions, participants were instructed to annotate the broadest expression whenever in doubt about whether some words should be included in the annotation. Each session produced two parallel sets of annotations for each document, unified in a single corpus for each annotation method.

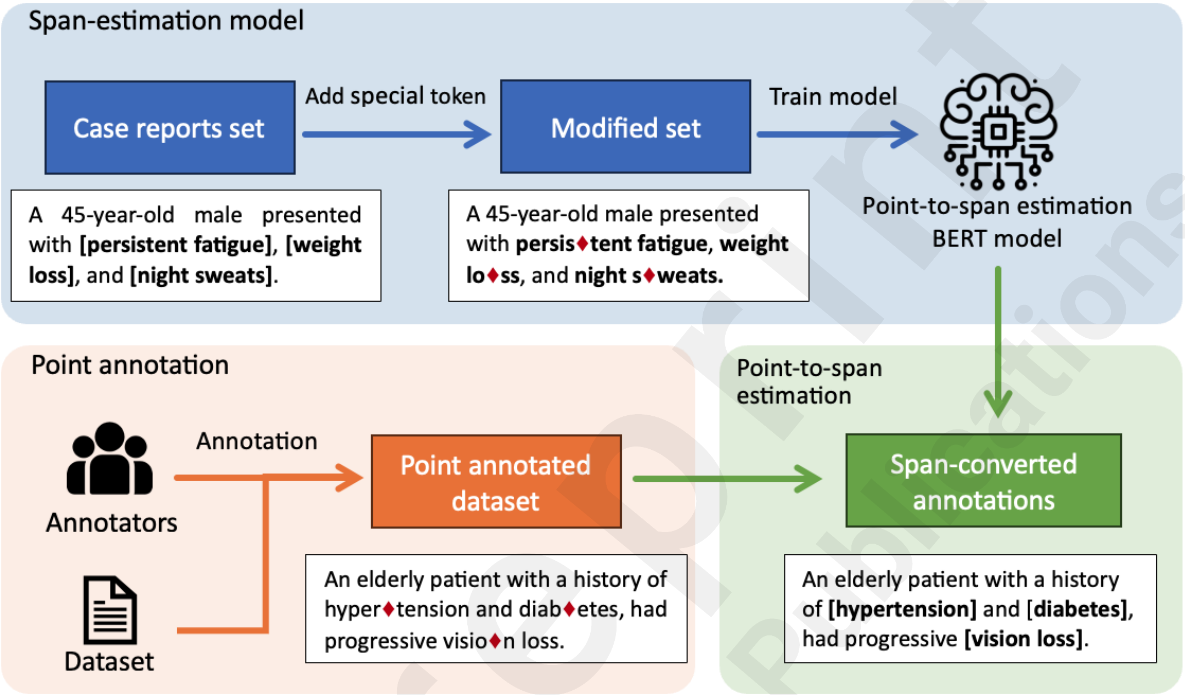
We resolved all disagreements between the two sets automatically. We accepted all annotations made by either annotator, even if there is no matching counterpart. Whenever there is boundary disagreement, we choose the broadest span possible when combining the two annotations.

For *Point* annotations, we grouped annotations that refer to the same NE and averaged their positions. We consider annotations as referring to the same concept when located within six characters of distance from each other. The distance limit was chosen based on the average Japanese word length, around three characters. We chose a larger value to account for multi-word concepts.

Point-to-span estimation

Being aware that the single-position label produced by the *Point* annotation method may not convey enough information about the adequate range of the NE to be extracted when training the model, we developed a *Point-to-span* estimation method [34]. It can complement the annotation with span information without additional manual work.

We used a BERT model (referred to as the *expansion model*) that receives the positional annotation and attempts to predict the original NE span. Effectively, it works as a method to convert Points into Span-based annotations, as illustrated in Figure 4.



A 45-year-old male presented with **[persistent fatigue]**, **[weight loss]**, and **[night sweats]**.

A 45-year-old male presented with **persis♦tent fatigue**, **weight lo♦ss**, and **night s♦weats**.

An elderly patient with a history of hyper♦tension and diab♦etes, had progressive visio♦n loss.

An elderly patient with a history of **[hypertension]** and **[diabetes]**, had progressive **[vision loss]**.

Figure 4. Flow of the Point-To-Span estimation process.

The *Point-to-span* estimation model is based on the pre-trained *tohoku-nlp/bert-base-japanese-char-v2* model [35], and it was fine-tuned using the training parameters presented in Table 4. Training was performed on a server with two NVIDIA Quadro RTX 8000 GPUs.

Table 4. Hyper-parameters used for model training.

Parameter	**Value**
Max epochs	10
Training batch size	16
Learning rate	3e-5
Optimizer	AdamW
Max sentence length	512 characters
Model selection	Early stopping
Training time	Approximately 30 min.

As training data, we used a large dataset of Japanese medical texts with labeled diseases and symptoms consisting of 1027 synthetic medication history notes generated through crowdsourcing. Ten experienced dispensing pharmacists were hired as writers to craft the corpus. Each writer was assigned one of 285 drug names and tasked with creating a “typical” clinical narrative.

Before being fed to the model, each annotation of the training data was replaced by an identifier token (⧫) in a random location within its span based on a truncated normal distribution. A different distribution was used for each annotation, centered on the middle point, with the standard deviation being a sixth of the annotation length. Due to the randomicity of the data, we augmented the dataset 10 times by re-executing the annotation replacement module and generating different valid positions for the ⧫ token.

The expansion model was then trained to identify this token and output the start and end positions of the concept based on the word containing the token and its surrounding context.

We evaluated the model by predicting the spans for annotations on the GSC. We pre-processed the GSC annotations using the same method to replace the annotations with ⧫ tokens. Our best model was able to achieve an F1 score of 0.77.

We applied the expansion model to the Point-annotated dataset to infer spans for each annotation, producing a *Point-expanded* corpus. Effectively, the combination of point annotation and expansion allows the generation of a span annotated dataset with less human effort.

Evaluation

Annotation Method Efficiency

We evaluated the annotation methods according to the following:

Annotation Quality We assessed the percentage of GSC concepts that were correctly annotated. We consider an annotation correct when at least one token overlaps with the GS span.

Annotation Time Annotators manually measured the time they took to work on the data during each session. They were instructed to start the timing after loading the texts in the annotation software.

Inter-Annotator Agreement (IAA) We use Cohen’s Kappa [36], one of the most common metrics for gauging agreement between annotators. Kappa is a function of the proportion of observed and expected agreement, and it may be interpreted as the proportion of agreement corrected for chance [37].
Given that the *Point* annotation methodology allows for multiple correct annotations within the NE span, we computed an additional *adjusted variant* of the metrics specifically for these annotations. In this variant, we considered annotations to agree if they were within a 3-character range of each other, reflecting the average word length in the Japanese language.

Downstream Task Performance

As one of the typical downstream tasks, we developed a NER system to benchmark each annotation approach. We again employed the pre-trained *tohoku-nlp/bert-base-japanese-char-v2* model [35] and fine-tuned it using our annotated corpora.

We used the same training parameters for all models, as presented in Table 4. To minimize the variability between results, we used 5-fold cross-validation and averaged the obtained values.

We evaluated model predictions on the MedTxt-CR-JA test set, comprised of 75 documents, by the metrics of *precision*, *recall*, and *F-score*. We employ two variants of the metrics:

Strict metrics follow CoNLL criteria [38] and only consider predictions where the span exactly matches the ground truth. These metrics allow us to estimate how closely the model fits the GS.

Relaxed metrics [39] accept partial matches or extra tokens as long as at least one token of the predicted span overlaps with the GS span. This variant allows assessing the model's capability of identifying the presence of concepts of interest in the text.

Results

Annotation Method Efficiency

Upon merging the data received from the annotators, we produced the final version of the annotated corpus for each one of the methodologies. Table 5 shows some statistics of the produced corpora.

There is no substantial difference between *Traditional* and *Lenient Span* methods when comparing the average length of the produced annotation. However, both produced annotations slightly larger than the gold annotations due to the disagreement resolution approach adopted in this study.

Table 5. Statistics of the produced corpora.

Method	**Total Annotations**	**Avg. Annotation Length (char)**
Gold Standard	1167	6.31
Traditional	1065	7.30
Lenient Span	1012	7.30
Point	1066	NA

Annotation Quality

Figure 5 shows the percentage of GSC annotations covered by each corpus. Although none of the methodologies captured all the ground truth concepts, the percentage of entities captured was similar for every method, with less than a 10% difference between the best (Lenient Span) and worst (Point).

As the value of missed entities is consistent for all methodologies, we attribute it to some divergence between the guidelines for annotating the GSC and the one used in this study. Differences in the interpretation may have led the annotators to skip some of the entities.

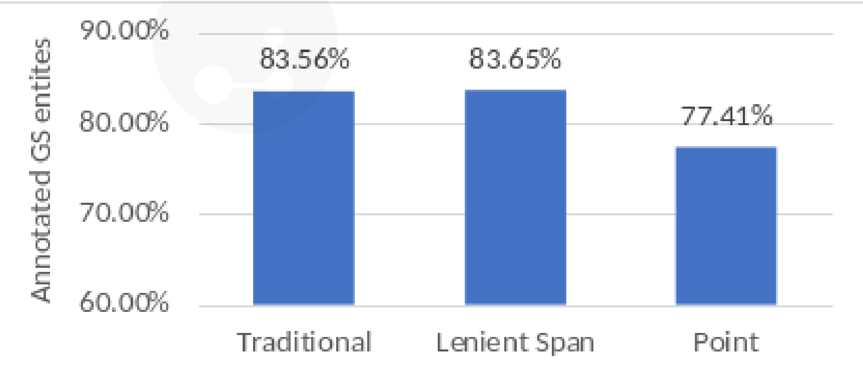


Figure 5. Percentage of correctly annotated GS entities.

Figure 6 presents the accuracy of the annotations of each participant in relation to GSC on each methodology. We noticed that the traditional methodology presented a more constant accuracy throughout the annotators, while the boundary-relaxed methods had more variation, especially for annotators C and D.

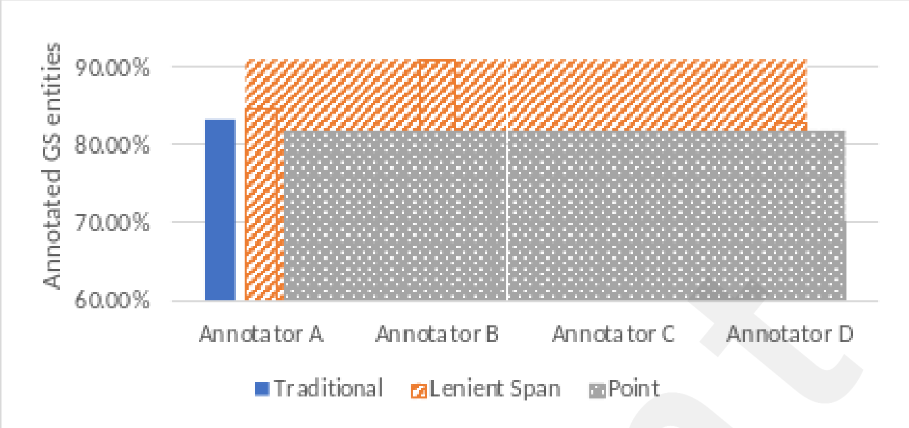


Figure 6. Annotation accuracy per annotator.

Annotation Time

The time measurement results in Table 6 demonstrate that both boundary-free annotation techniques can provide time-saving benefits. On average, reductions of around 25% (around 28 minutes) and 20% (around 21 minutes) were observed when using *Point* and *Lenient Span* methods, respectively, compared to the *Traditional* annotation process.

Table 6. Comparison of the individual annotation time per annotation method. Times are presented in the HH:MM:SS format, with the percentage comparison to the Traditional method in parenthesis.

Annotator	**Traditional**	**Lenient Span**	**Point**
A	1:23:44	1:03:23 (-24%)	0:54:35 (-35%)
B	1:09:14	0:52:07 (-25%)	0:48:45 (-30%)
C	3:16:58	2:10:20 (-34%)	2:15:27 (-31%)
D	1:10:23	1:31:29 (+30%)	1:10:40 (+0%)
Average	1:45:05	1:24:20 (-20%)	1:17:22 (-26%)

Inter-Annotator Agreement

As evidenced by the results presented in Figure 7, the IAA measured for both boundary-free annotation methods overcame the *Traditional* methodology.

Point annotations recorded the lowest agreement due to the inherent low probability of annotators precisely pinpointing the exact same position within an NE. Despite that, it achieves the highest measured agreement using the *adjusted variant of the metrics*.

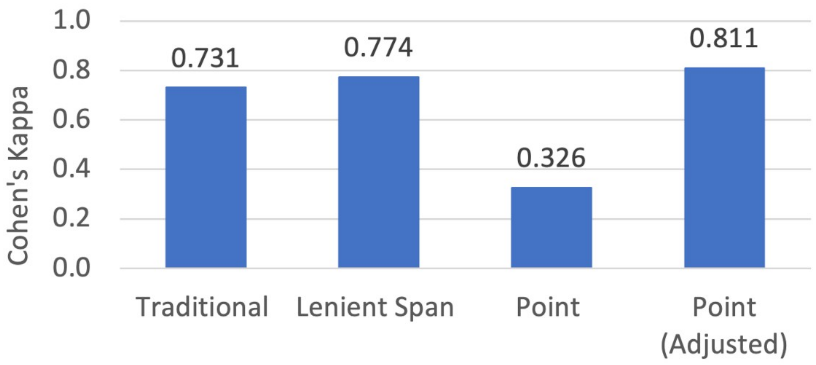


Figure 7. Average IAA per annotation methodology.

Downstream Task Performance

Table 7 presents the NER model evaluation results. We trained a *Gold Standard Model* (GSM) using the Gold Standard data as a reference for our system's best possible performance.

The data produced in our annotation experiments probably has lower quality due to the lack of proper curation and review sessions. Thus, when comparing the *Traditional* annotation approach against the GSM, there is a slight decrease in performance, 15% and 11% on strict and relaxed metrics, respectively. Nevertheless, the relation between precision and recall remains the same, as both models were trained on similarly boundary-strict annotations.

Table 7. Evaluation of the trained NER models.

Annotation Approach	**Strict**				**Relaxed**		
	Precision	**Recall**	**F1**		**Precision**	**Recall**	**F1**
Gold Standard Model	0.72	0.78	**0.75**		0.90	0.89	**0.89**
Traditional	0.60	0.69	**0.64**		0.77	0.81	**0.79**
Lenient Span	0.56	0.54	**0.55**		0.67	0.62	**0.64**
Point	0.00	0.00	**0.00**		0.60	0.45	**0.51**
Point (Expanded)	0.34	0.35	**0.35**		0.73	0.71	**0.72**

Discussion

Principal Findings

Throughout the experiments, it was noticeable that simplifying the annotation process contributed to a more comfortable experience for the participants. We observed increased annotation speed, annotator agreement, and overall positive feedback from the annotators regarding the changes.

Although we showcase our proposal in clinical data, the annotation methodologies are both domain and language-agnostic, so they can be applied to texts of different domains and idioms.

Annotation Speed Improvements

The results in Table 6 show that simplifying the constraints under which annotators work can effectively increase the speed at which they execute the task. By virtually removing the need to decide on entity boundaries, both proposed methodologies allowed the annotation of our dataset in less time than the *Traditional* method.

However, while an overall decreasing trend in annotation time was observed, different annotators experienced varying degrees of time reduction. Notably, Annotator C experienced a significant increase in efficiency when using these methodologies. Conversely, Annotator D was quicker with the *Traditional* annotation scheme. Still, his precision was lower than other annotators, as shown by the individual accuracy results presented in Figure 6.

Annotator Agreement Improvements

Meanwhile, the IAA evaluation (Figure 7) revealed some interesting insights into the annotation consistency of each methodology. Both the *Lenient Span* and the adjusted *Point* agreement overcame the *Traditional* methodology by 5.88% and 10.94%, respectively.

While we believe that slightly different interpretations of what information should be annotated may have diminished *Traditional* approach agreement, such a finding was still unexpected due to the higher flexibility given to the annotators when removing the need for entity boundaries. However, this improvement can be attributed to the ease with which annotators can consistently agree on the core parts of mentions (or the "main words") compared to determining the precise boundaries of entire entities. Such boundaries may or may not encompass adjectives, modifiers, etc., which often contribute to annotation disagreements.

Notably, *Point* annotations perceived a large difference in the agreement values measured using the default and *adjusted* variants of the IAA metrics. This is explained by the fact that, even though it is virtually impossible for annotators to select the same character in an NE for all annotations, they generally selected positions close to each other for the same NE. Such finding is evidenced by the distribution of annotation pairs based on the number of characters of difference between them, as depicted in Figure 8.

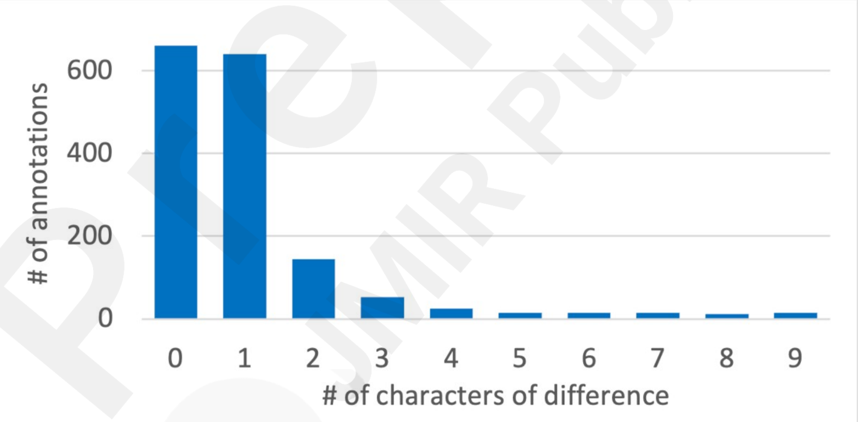


Figure 8. Distribution of annotation pairs based on the distance between them.

Such a small distance is due to annotators’ diligence in positioning the annotation close to the center of the NE’s core word. As in the sentence shown in Figure 9 (which translates to “Current symptoms: Diffuse **dark red infiltration** is observed on both cheeks.”), even though the span of the desired annotation is quite large, both annotators placed its label near the most relevant set of words, “dark red infiltration”.

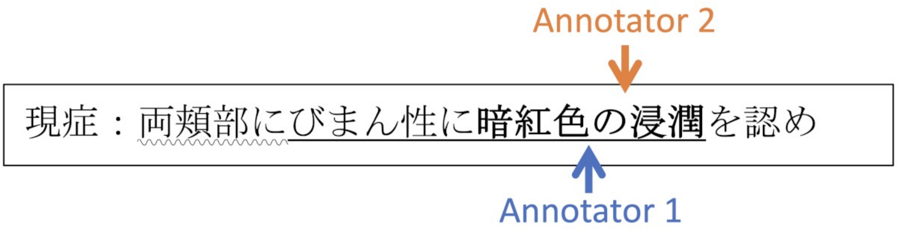


Figure 9. Example of two distinct Point annotations in an NE with a large span (underscored). The annotations are located near the center of the core word (in bold).

Annotator's Opinions

Annotator feedback was positive especially regarding the *Point* annotation, given its simplicity. The participants highlighted the easiness of the single-click selection mode, particularly due to the reduced mouse manipulation needed.

However, the participants expressed difficulty in understanding the correctness of their annotations and whether the chosen range was indeed accurate. They felt that the soft boundaries displayed by the annotation tool turned the annotations ambiguous, making them unsure whether they matched the range they intended to select.

Impacts on Model Performance

While achieving significant improvements in annotators' work quality, the additional flexibility from boundary-free methods considerably impacted model performance, particularly in strict evaluation, due to the imprecise training data, as seen in Table 7.

The *Lenient Span*-trained model exhibited a significant subside in its recall, which hindered strict and relaxed evaluations. We did not expect that the ambiguity in NE boundaries could affect the model's capability of locating NEs in the text.

While such performance drop may be acceptable for some applications, we believe additional annotation post-processing methods could restore the accuracy to levels similar to the *Traditional* schema.

Point-to-Span Estimation

In particular, the insights from *Point* annotation experiments underscore the potential of automated methods to supplement human annotations. We believe that *Point-to-span estimation* can be pivotal for improving annotation speed, but beyond that, it can be proven beneficial to aid in addressing other annotation problems.

Given the lackluster nature of the annotation task, it is not uncommon that annotators make mistakes, such as including punctuation markers or failing to label part of the NE simply for a lack of focus. The span estimation model can be a tool to "normalize" such annotations.

Furthermore, the estimation could be integrated into the actual annotation process by coupling it with our annotation tool, enabling the "click-only" annotation interface to present the predicted span directly and allowing the annotator to correct its mistakes.

However, there is potential for enhancements in the expansion model. Although expanding a point to the expected word seems to be a simple task, as we are evaluating our methods on a Scriptio continua language, which makes the definition of the word boundaries not as obvious as in space-delimited languages, such as English.

Through analysis of the model's output, we have observed that the estimation model exhibited a tendency to choose spans larger than the Gold Standard (GS) entities, particularly when characters that act like qualitative adjectives (such as "□" for high, "□□" for acute, "□□" for huge) were connected to the concept of interest.

For instance, the model outputted “□□□□□□" (Severe liver atrophy) instead of only “□□□" (Liver atrophy). Another example was the expansion of the term “□□□□□□□□□" (Giant splenorenal shunt), where □□□ (Giant) was included.
Yet, even though the model output in these examples can be regarded as “incorrect” when compared to the GSC, from a clinical point of view, it is not uncommon that some diseases are distinguished by such modifier words. For example, “□□□□□” (acute cholecystitis) and “□□□□□ (chronic cholecystitis), which even have different International Classification of Diseases (ICD) codes, K81.0 and K81.1, respectively.

Error Analysis

Tables 8 and 9 present example comparisons between all the evaluated models in two different sentences.

We could not identify any unusual behavior when inspecting the Traditional annotation model output. Yet, we highlight that the Lenient Span model portrayed a tendency to overly extend the span lengths. In some cases (as shown especially in Table 9), multiple NEs are “merged” into a single continuous extraction.

As seen in both examples, the model trained with raw point annotations could not extract NE spans, denoting that the single position annotation contains insufficient information to train the model properly.

In contrast, the model trained on expanded point annotations showcases the effectiveness of the *Point-to-span* estimation method. Although strict metrics are still substantially lower than other approaches, relaxed results are comparable to the *Traditional* annotation approach. The analysis of the model output evidenced that, while it could locate most concepts of interest, it struggled in correctly extracting multi-word concepts.

Table 8. Comparison of model output for the sentence “*While waiting for a CT scan, patient went into cardiopulmonary arrest (CA), but could not be resuscitated and died*”. Gold standard entities and model extractions are marked in bold and underscored. [a]

Gold Standard	□□□□□ □□□□□□□□ □□□□ □□□□□□□ □□□□□□□□ (CT scan) (while waiting) (**CA**) (had) (cannot resuscitate) □□□□ (**died**)
Model	**Example**
Traditional	□□□□□ □□□□□□□□ □□□□ □□□□□□□ □□□□□□□□□□□□□
Lenient Span	□□□□□ □□□□□□□□ □□□□ □□□□□□□ □□□□□□□□□□□□□

Point	□□□□□ □□□□□□□ □□□□ □□□□□□ □□□□□□□ □□□□
Point (Expanded)	□□□□□ □□□□□□□ □□□□ □□□□□□ □□□□□□□□□□□

[a] White space tokenization was added to the Japanese text to enhance readability for non-Japanese readers. The original text does not contain spaces.

Table 9. Comparison of model output for the sentence "*History of hypertension (HTN), diabetes, hyperlipidemia (HLD), or atrial fibrillation (AFib)*". Gold standard entities and model extractions are marked in bold and underscored. [a]

Gold Standard	□□□ □□□□ □□□□ □□□□□ □□□□ □□ (History) (**HTN**) (**diabetes**) (**HLD**) (**AFib**) (had)
Model	**Example**
Traditional	□□□ □□□□ □□□□ □□□□□ □□□□ □□
Lenient Span	□□□ □□□□ □□□□ □□□□□ □□□□ □□
Point	□□□ □□□□ □□□□ □□□□□ □□□□ □□
Point (Expanded)	□□□ □□□□ □□□□ □□□□□ □□□□ □□

[a] White space tokenization was added to the Japanese text to enhance readability for non-Japanese readers. The original text does not contain spaces.

Limitations

While our research focused on exploring novel approaches to text annotation and revealed promising findings, a few concerns and limitations need further investigation. Our investigations were only conducted in the Japanese language. Though our proposal is language-independent, applying our techniques in a space-delimited language, such as English, could introduce some bias. Evaluation using different languages is, thus, encouraged. Since our dataset in this study has an English variant, we plan to conduct additional experiments.

We concentrated on a singular entity class, disease, and symptom names to streamline the analysis. Even though our texts contain a large number of entities, a single class annotation may not represent a real use case. Exploring our methodologies in a multi-class scenario would enhance the robustness of our findings and conclusions.

Furthermore, we acknowledge that automated labeling techniques, such as pre-annotation, can affect the improvements observed in annotation time by adopting boundary-free methodologies. We chose not to incorporate these features in our annotation tool to minimize the number of variables affecting the annotation process.

The observed performance of the trained NER models could have been impacted by our choice of using a simple and automatic approach to solve disagreements. Although it avoids additional annotator work and simplifies the research flow, implementing adjudication or review sessions with the annotations would be preferred, as it could have provided a better annotation quality.

LLMs are prevalent in the current NLP research scenario, and their application has led to the development of systems that push state-of-the-art performance in many different tasks. In the current state of our work, we have not adopted LLMs. Still, we acknowledge that the accuracy of our methods may be improved by employing such methods in our workflow, possibly replacing the Point-to-Span BERT model.

Conclusions

In this study, we investigated the effects of reducing the emphasis on entity boundary annotations while labeling NEs in a medical dataset. We proposed two novel boundary-free annotation methodologies, *Lenient Span* and *Point* annotation. We evaluated the impact of their application in an annotation process regarding annotation efficiency and the quality of the labeling produced.
We also publicly released our developed annotation tool [33] and point-to-span estimation model [34].

Our results demonstrate a trade-off relation between annotation efficiency and model performance. Although not surprising, it unveils the weak points of each methodology and uncovers potential adjustments that can be made to each approach. We underscore that completely disregarding boundary information may ease the annotator’s work while it sacrifices performance to some extent.

We plan to evaluate the proposed methodologies in other languages in future work. Also, we intend to explore the impact of post-processing techniques, such as normalization or boundary-regularization, to enhance model output performance.

Acknowledgments

This work was supported by JST CREST Grant Number JPMJCR22N1, JSPS KAKENHI Grant Number JP19H01118, and Cross-ministerial Strategic Innovation Promotion Program (SIP) on “Integrated Health Care System” Grant Number JPJ012425, Japan.

Data Availability

This study makes use of two datasets:

- MedTxt-CR is publicly available and contains case reports extracted using OCR from PDF files of open-access articles in the Japanese journal repository, J-Stage. Access to the dataset requires an email request.
- A dataset of synthetic medication history notes generated through crowdsourcing, which is not yet publicly available.

Author Contributions

GHBA designed the study, performed the computational experiments and data analysis, and wrote the manuscript. SY and EA discussed the results and reviewed the manuscript. EA supervised the study. All the authors have approved the final manuscript.

Conflicts of Interest

All authors declare that they have no conflicts of interest.

Abbreviations

AL: Active learning
BERT: Bidirectional encoder representations from Transformers
EHR: Electronic health records

GS: Gold standard
GSC: Gold standard corpus
GSM: Gold standard model
NE: Named entity
NER: Named entity recognition
NLP: Natural language processing

References

1. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, Godtliebsen F. Challenges and opportunities beyond structured data in analysis of electronic health records. WIREs Computational Statistics 2021;13(6):e1549. doi: 10.1002/wics.1549
2. Gomes I, Correia R, Ribeiro J, Freitas J. Effort Estimation in Named Entity Tagging Tasks. Proceedings of the Twelfth Language Resources and Evaluation Conference Marseille, France: European Language Resources Association; 2020. p. 298–306. Available from: https://aclanthology.org/2020.lrec-1.37
3. Monajatipoor M, Yang J, Stremmel J, Emami M, Mohaghegh F, Rouhsedaghat M, Chang K-W. LLMs in Biomedicine: A study on clinical Named Entity Recognition. 2024; doi: 10.48550/arXiv.2404.07376
4. Marrero M, Urbano J, Sánchez-Cuadrado S, Morato J, Gómez-Berbís JM. Named Entity Recognition: Fallacies, challenges and opportunities. Comput Stand Interfaces 2013;35(5):482–489. doi: https://doi.org/10.1016/j.csi.2012.09.004
5. Chapman WW, Nadkarni PM, Hirschman Lynette and D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc Oxford University Press (OUP); 2011 Sep;18(5):540–543. PMID:21846785
6. Baledent A, Mathet Y, Widlöcher A, Couronne C, Manguin J-L. Validity, Agreement, Consensuality and Annotated Data Quality. Proceedings of the Thirteenth Language Resources and Evaluation Conference Marseille, France: European Language Resources Association; 2022. p. 2940–2948. Available from: https://aclanthology.org/2022.lrec-1.315
7. Zhu E, Li J. Boundary Smoothing for Named Entity Recognition. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) Dublin, Ireland: Association for Computational Linguistics; 2022. p. 7096–7108. doi: 10.18653/v1/2022.acl-long.490
8. Andrade GHB, Yada S, Aramaki E. Comparative evaluation of boundary-relaxed annotation for Entity Linking performance. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) Association for Computational Linguistics; 2023. p. 8238–8253. doi: 10.18653/v1/2023.acl-long.458
9. Ganchev K, Pereira F, Mandel M, Carroll S, White P. Semi-Automated Named Entity Annotation. Proceedings of the Linguistic Annotation Workshop Prague, Czech Republic: Association for Computational Linguistics; 2007. p. 53–56. Available from: https://aclanthology.org/W07-1509
10. Komiya K, Suzuki M, Iwakura T, Sasaki M, Shinnou H. Comparison of Methods to Annotate Named Entity Corpora. ACM Transactions on Asian and Low-Resource Language Information Processing New York, NY, USA: Association for Computing Machinery; 2018 Jul;17(4). doi: 10.1145/3218820
11. Dasgupta Sanjoy and Kalai AT and MC. Analysis of Perceptron-Based Active Learning. Learning Theory Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. p. 249–263.
12. Kholghi M, Sitbon L, Zuccon G, Nguyen A. Active learning reduces annotation time for clinical concept extraction. Int J Med Inform 2017 Oct;106:25–31. PMID:28870380
13. Tokunaga T, Nishikawa H, Iwakura T. An Eye-tracking Study of Named Entity Annotation. Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017 Varna, Bulgaria: INCOMA Ltd.; 2017. p. 758–764. doi: 10.26615/978-954-452-049-

6_097

14. Saxena K, Sunkle S, Kulkarni V. Hybrid Search based Enhanced Named Entity Annotation Tool. Proceedings of the 15th Innovations in Software Engineering Conference New York, NY, USA: Association for Computing Machinery; 2022. doi: 10.1145/3511430.3511455
15. Kim H, Mitra K, Li Chen R, Rahman S, Zhang D. MEGAnno+: A Human-LLM Collaborative Annotation System. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations St. Julians, Malta: Association for Computational Linguistics; 2024. p. 168–176. Available from: https://aclanthology.org/2024.eacl-demo.18
16. Goel A, Gueta A, Gilon O, Liu C, Erell S, Nguyen LH, Hao X, Jaber B, Reddy S, Kartha R, Steiner J, Laish I, Feder A. LLMs Accelerate Annotation for Medical Information Extraction. Proceedings of the 3rd Machine Learning for Health Symposium PMLR; 2023. p. 82–100. doi: 10.48550/arXiv.2312.02296
17. Kholodna N, Julka S, Khodadadi M, Gumus MN, Granitzer M. LLMs in the Loop: Leveraging Large Language Model Annotations for Active Learning in Low-Resource Languages. 2024. doi: 10.48550/arXiv.2404.02261
18. Tan Z, Beigi A, Wang S, Guo R, Bhattacharjee A, Jiang B, Karami M, Li J, Cheng L, Liu H. Large Language Models for Data Annotation: A Survey. 2024. doi: 10.48550/arXiv.2402.13446
19. Sabou M, Bontcheva K, Derczynski L, Scharl A. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) Reykjavik, Iceland: European Language Resources Association (ELRA); 2014. p. 859–866. Available from: http://www.lrec-conf.org/proceedings/lrec2014/pdf/497_Paper.pdf
20. Snow R, O'Connor B, Jurafsky D, Ng A. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing Honolulu, Hawaii: Association for Computational Linguistics; 2008. p. 254–263. Available from: https://aclanthology.org/D08-1027
21. Li J. A Comparative Study on Annotation Quality of Crowdsourcing and LLM via Label Aggregation. 2024; doi: 10.48550/arXiv.2401.09760
22. Pangakis N, Wolken S, Fasching N. Automated Annotation with Generative AI Requires Validation. 2023; doi: 10.48550/arXiv.2306.00176
23. Liu K, Fu Y, Tan C, Chen M, Zhang N, Huang S, Gao S. Noisy-Labeled NER with Confidence Estimation. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Online: Association for Computational Linguistics; 2021. p. 3437–3445. doi: 10.18653/v1/2021.naacl-main.269
24. Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y. DiffusionNER: Boundary Diffusion for Named Entity Recognition. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) Toronto, Canada: Association for Computational Linguistics; 2023. p. 3875–3890. doi: 10.18653/v1/2023.acl-long.215
25. Yada S, Nakamura Y, Wakamiya S, Aramaki E. Real-MedNLP: Overview of real document-based medical natural language processing task. Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies 2022. p. 285–296. Available from: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/01-NTCIR16-OV-MEDNLP-YadaS.pdf
26. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data Springer Science and Business Media LLC; 2016 May;3(1):160035. PMID:27219127
27. Mahajan D, Liang JJ, Tsou C-H. Toward understanding clinical context of medication change events in clinical narratives. AMIA Annual Symposium Proceedings 2021;2021:833–842. PMID:35308981
28. Nishiyama T, Nishidani M, Ando A, Yada S, Wakamiya S, Aramaki E. NAISTSOC at the NTCIR-16

Real-MedNLP Task. Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies 2022. Available from: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/07-NTCIR16-MEDNLP-NishiyamaT.pdf

29. Spasic I, Nenadic G. Clinical text data in machine learning: Systematic review. JMIR Med Inform JMIR Publications Inc.; 2020 Mar;8(3):e17984. PMID:32229465

30. Yada S, Joh A, Tanaka R, Cheng F, Aramaki E, Kurohashi S. Towards a Versatile Medical-Annotation Guideline Feasible Without Heavy Medical Knowledge: Starting From Critical Lung Diseases. Proceedings of the Twelfth Language Resources and Evaluation Conference European Language Resources Association; 2020. p. 4565–4572. Available from: https://aclanthology.org/2020.lrec-1.561

31. Ollion E, Shen R, Macanovic A, Chatelain A. ChatGPT for Text Annotation? Mind the Hype! SocArXiv; 2023 Oct; doi: 10.31235/osf.io/x58kn

32. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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) Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–4186. doi: 10.18653/v1/N19-1423

33. Andrade GHB. GitHub - gabrielandrade2/FuzzyAnnotationTool: Annotation Tool for Fuzzy NER research. 2023. Available from: https://github.com/gabrielandrade2/FuzzyAnnotationTool [accessed Apr 16, 2024]

34. Andrade GHB. GitHub - gabrielandrade2/Point-to-Span-estimation. 2023. Available from: https://github.com/gabrielandrade2/Point-to-Span-estimation [accessed Apr 16, 2024]

35. Tohoku NLP Group. Hugging Face - tohoku-nlp/bert-base-japanese-char-v2. 2023. Available from: https://huggingface.co/tohoku-nlp/bert-base-japanese-char-v2 [accessed Apr 16, 2024]

36. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 1960;20(1):37–46. doi: 10.1177/001316446002000104

37. Warrens MJ. Five ways to look at Cohen's kappa. J Psychol Psychother OMICS International; 2015 Jul;5. doi: 10.4172/2161-0487.1000197

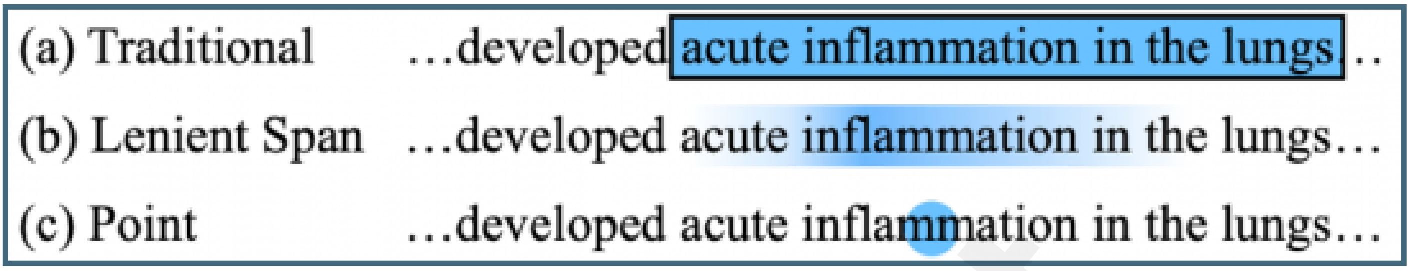
38. Tjong Kim Sang EF, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 2003. p. 142–147. Available from: https://aclanthology.org/W03-0419

39. Ghiasvand O, Kate RJ. Learning for clinical named entity recognition without manual annotations. Inform Med Unlocked 2018;13:122–127. doi: https://doi.org/10.1016/j.imu.2018.10.011

Supplementary Files

Figures

Example of different annotation paradigms. Traditional annotation (a) requires precisely labeling the beginning and end of the span, while boundary-free (b and c) methods focus on only identifying the core term.

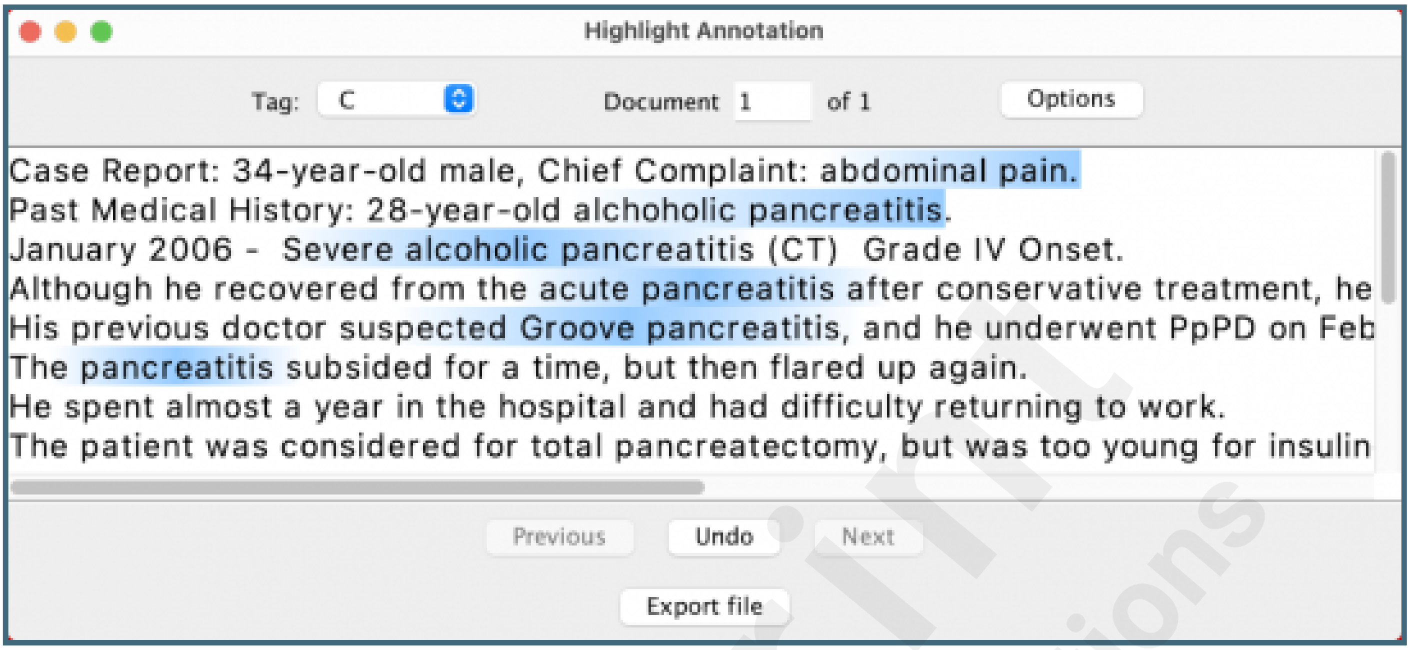


(a) Traditional …developed acute inflammation in the lungs…

(b) Lenient Span …developed acute inflammation in the lungs…

(c) Point …developed acute inflammation in the lungs…

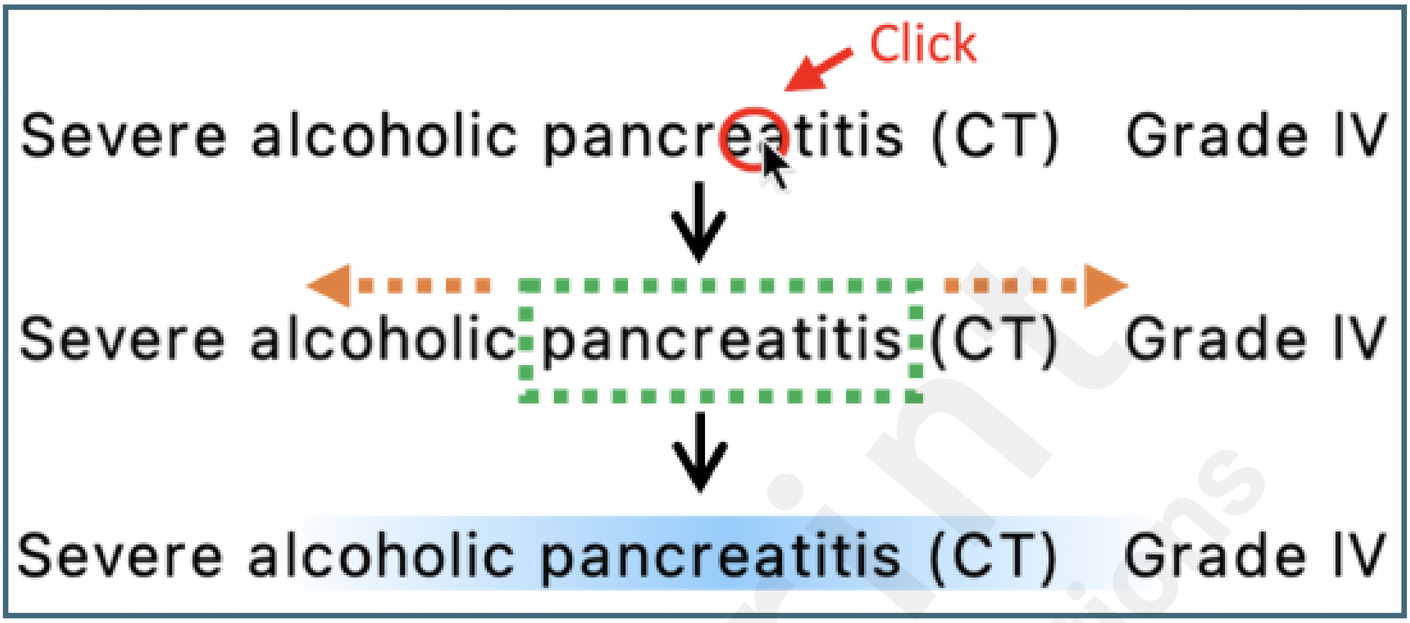
Screenshot of the Annotation Tool.



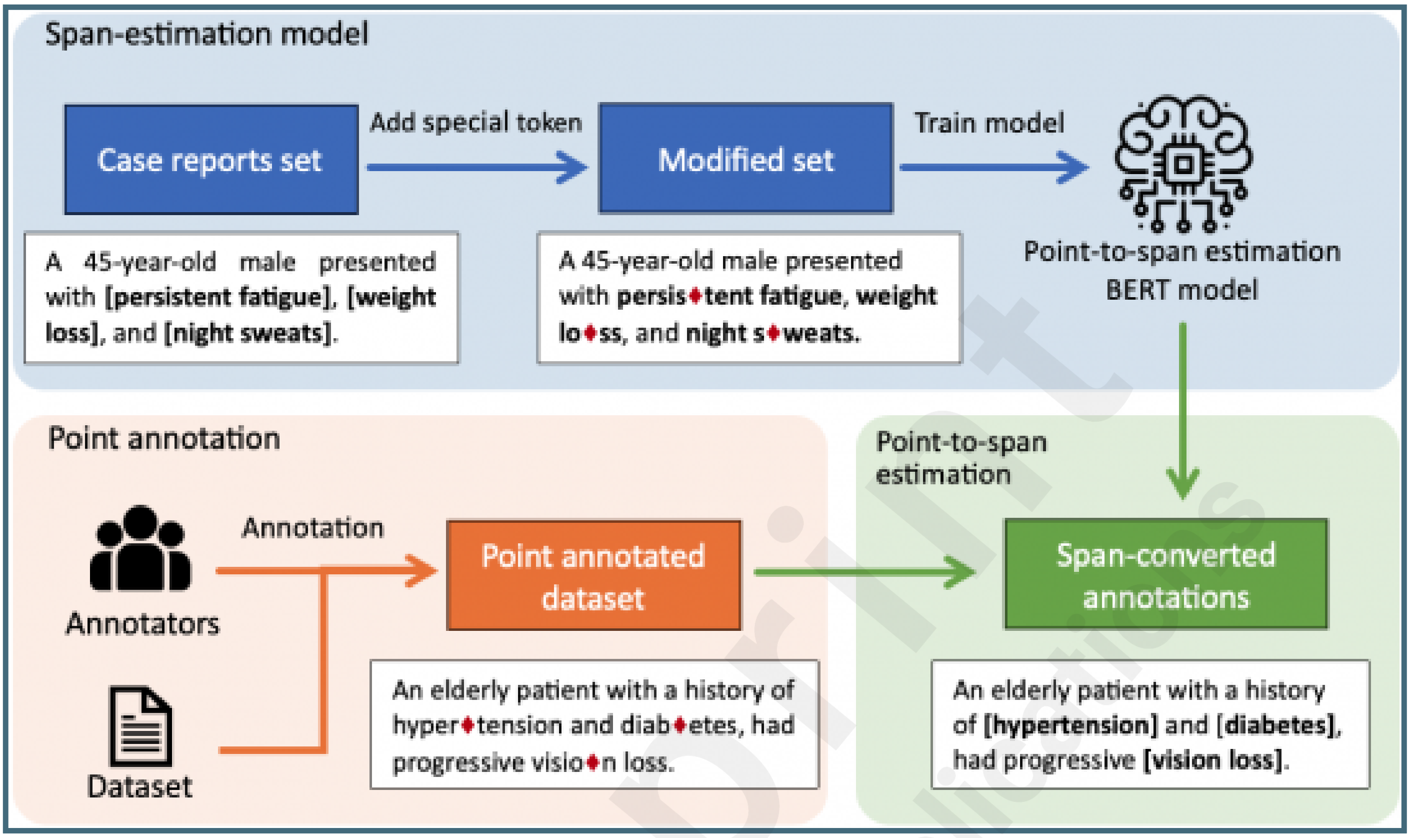
Tag: C Document 1 of 1

Case Report: 34-year-old male, Chief Complaint: abdominal pain.
Past Medical History: 28-year-old alchoholic pancreatitis.
January 2006 - Severe alcoholic pancreatitis (CT) Grade IV Onset.
Although he recovered from the acute pancreatitis after conservative treatment, he
His previous doctor suspected Groove pancreatitis, and he underwent PpPD on Feb
The pancreatitis subsided for a time, but then flared up again.
He spent almost a year in the hospital and had difficulty returning to work.
The patient was considered for total pancreatectomy, but was too young for insulin

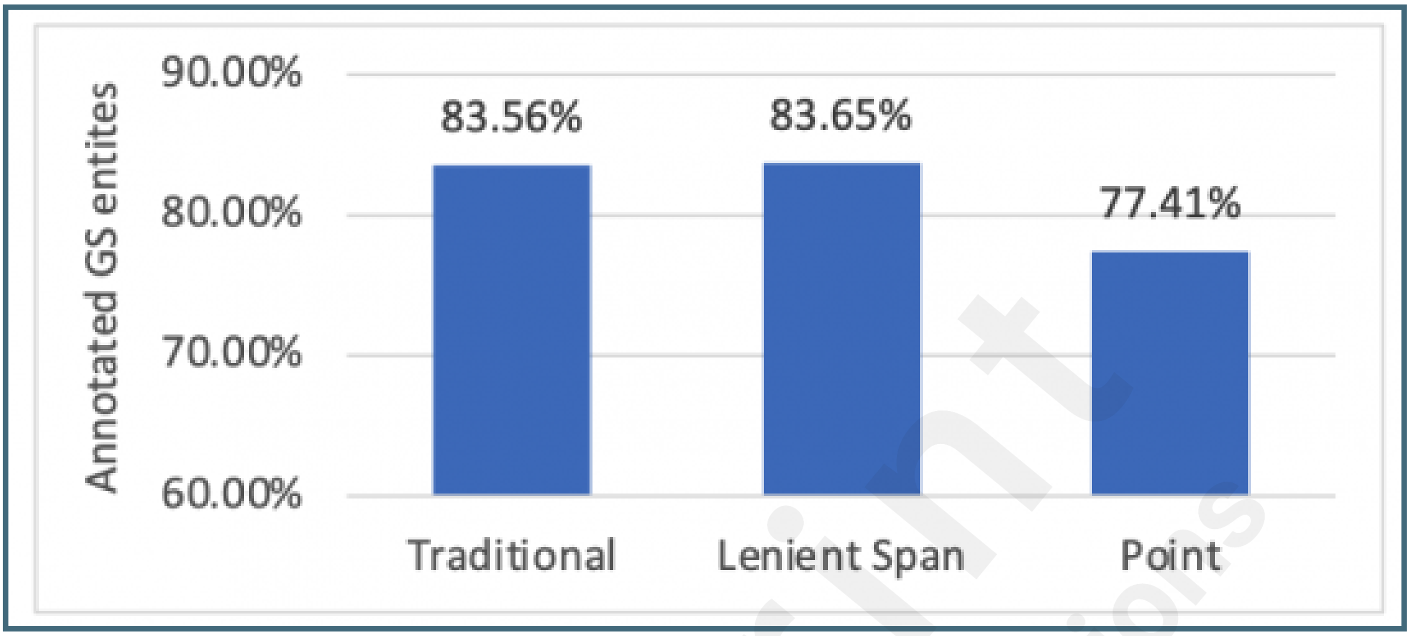
Example of a click-only annotation. The selected position, represented by the red circle, is expanded to the word boundaries (in green) plus a random span (orange arrows).


Click
Severe alcoholic pancreatitis (CT) Grade IV
Severe alcoholic pancreatitis (CT) Grade IV
Severe alcoholic pancreatitis (CT) Grade IV

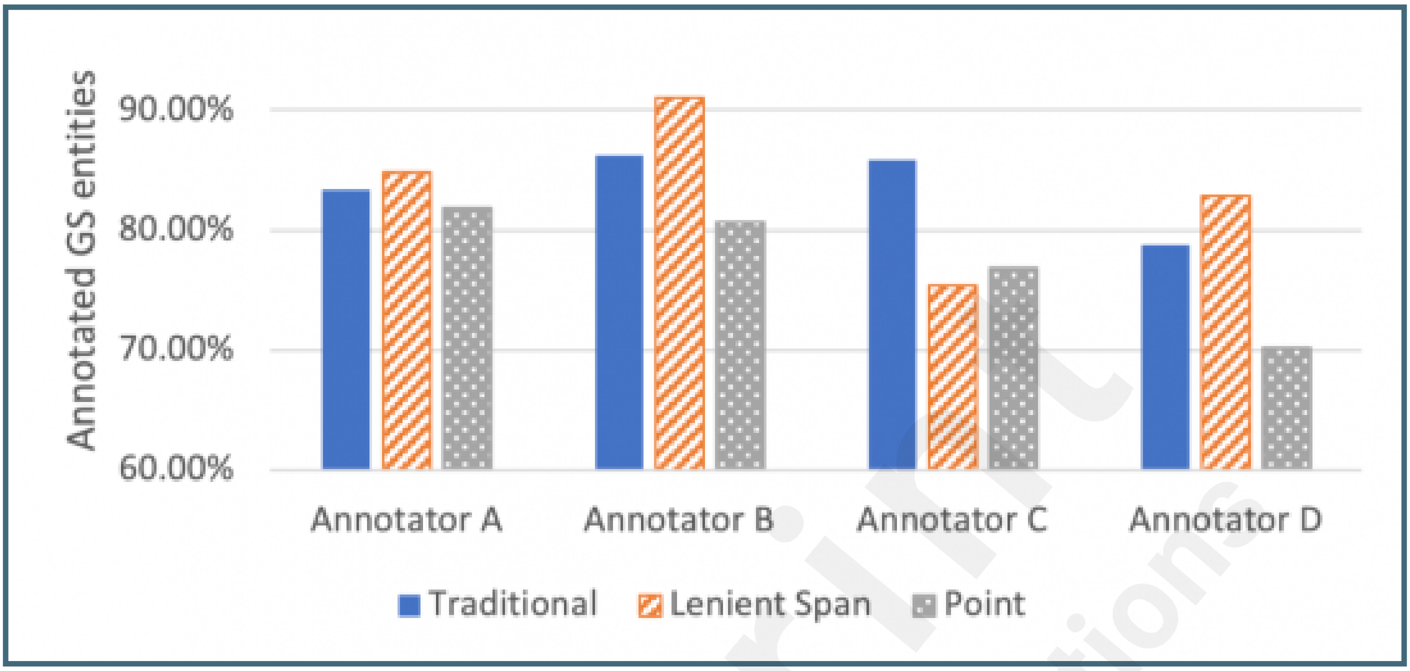
Flow of the Point-To-Span estimation process.


Span-estimation model
Case reports set
Add special token
Modified set
Train model
A 45-year-old male presented with [persistent fatigue], [weight loss], and [night sweats].
A 45-year-old male presented with persis♦tent fatigue, weight lo♦ss, and night s♦weats.
Point-to-span estimation BERT model
Point annotation
Annotation
Annotators
Dataset
Point annotated dataset
An elderly patient with a history of hyper♦tension and diab♦etes, had progressive visio♦n loss.
Point-to-span estimation
Span-converted annotations
An elderly patient with a history of [hypertension] and [diabetes], had progressive [vision loss].

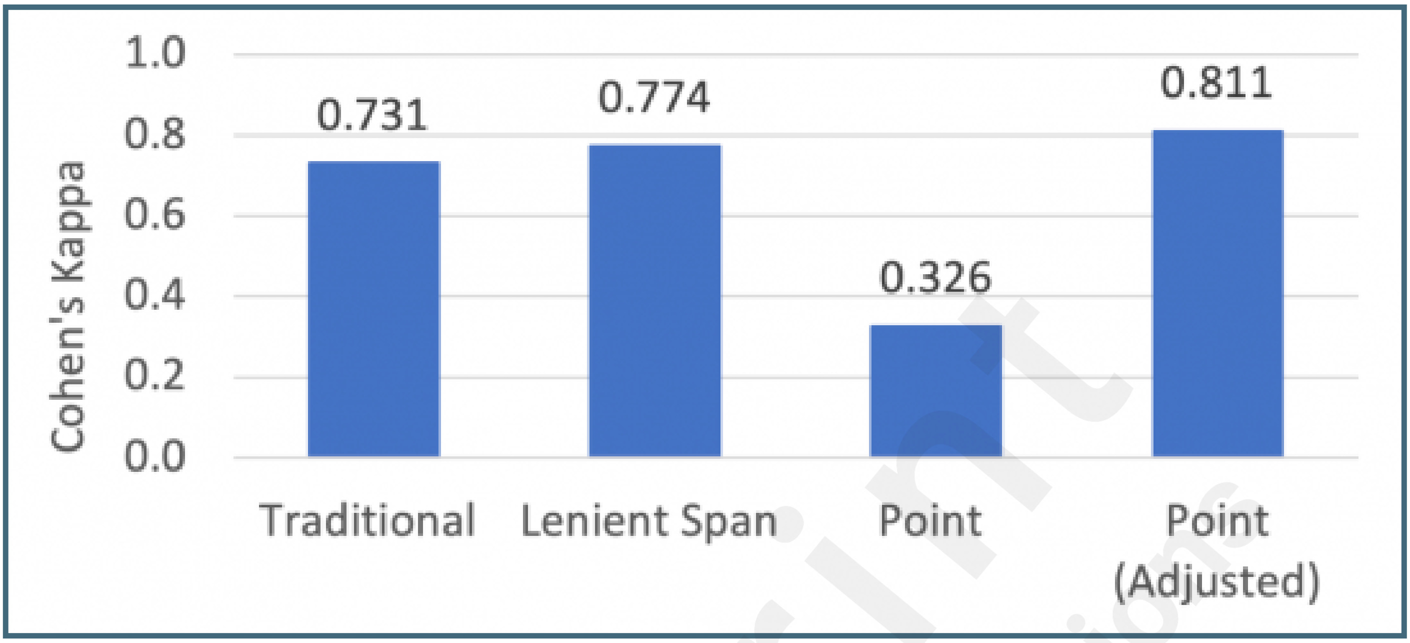
Percentage of correctly annotated GS entities.



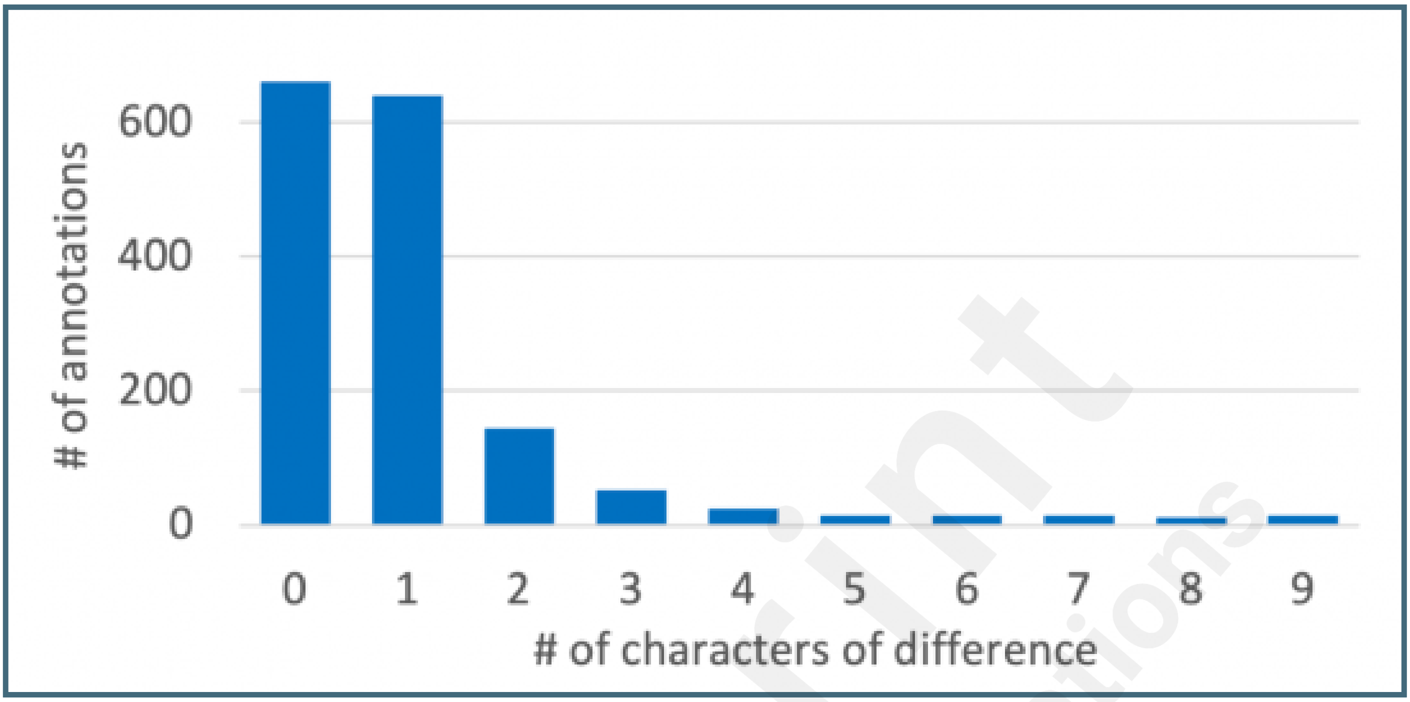
Annotation accuracy per annotator.



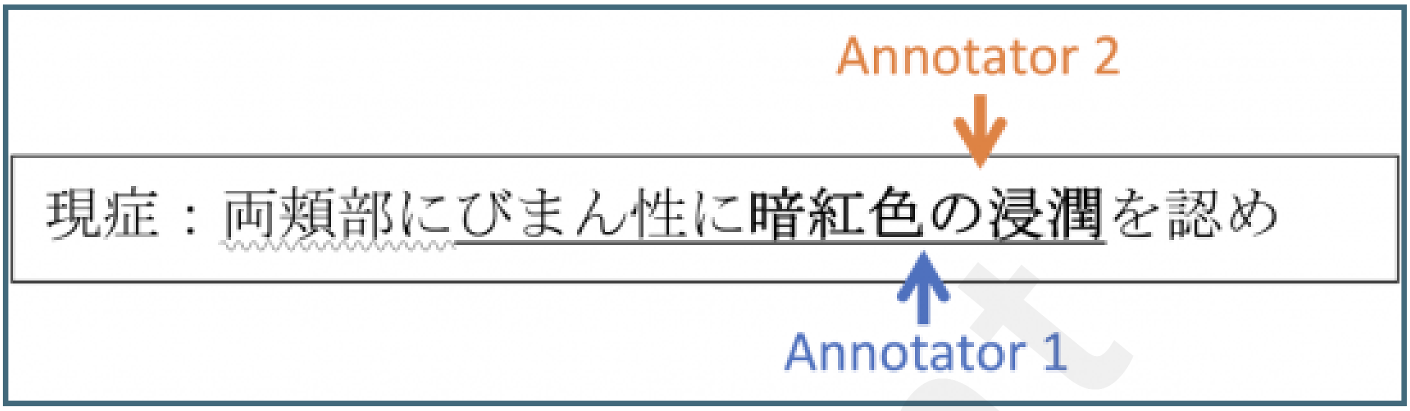
Average Inter-Annotator Agreement per annotation methodology.



Distribution of annotation pairs based on the distance between them.



Example of two distinct Point annotations in an NE with a large span (underscored). The annotations are located near the center of the core word (in bold).



現症：両頬部にびまん性に**暗紅色の浸潤**を認め

Multimedia Appendixes

Instructions on how to use the annotation tool.
URL: http://asset.jmir.pub/assets/a8040422368c3583892565f7e9b0d9cb.mp4

TOC/Feature image for homepages

TOC Placeholder.