QA-based Event Start-Points Ordering for Clinical Temporal Relation Annotation

Seiji Shimizu, Lis Kanashiro Pereira, Shuntaro Yada, Eiji Aramaki
Nara Institute of Science and Technology, Nara, Japan
{shimizu.seiji, kanashiro.lis, s-yada, aramaki}@is.naist.jp

Abstract

Temporal relation annotation in the clinical domain is crucial yet challenging due to its workload and the medical expertise required. In this paper, we propose a novel annotation method that integrates event start-points ordering and question-answering (QA) as the annotation format. By focusing only on two points on a timeline, start-points ordering reduces ambiguity and simplifies the relation set to be considered during annotation. QA as annotation recasts temporal relation annotation into a reading comprehension task, allowing annotators to use natural language instead of the formalisms commonly adopted in temporal relation annotation. Based on our method, most of the relations in a document are inferable from a significantly smaller number of explicitly annotated relations, showing the efficiency of our proposed method. Using these inferred relations, we develop a temporal relation classification model that achieves a 0.72 F1 score. Also, by decomposing the annotation process into QA generation and QA validation, our method enables collaboration among medical experts and non-experts. We obtained high inter-annotator agreement (IAA) scores, which indicate the positive prospect of such collaboration in the annotation process. Our annotated corpus, annotation tool, and trained model are publicly available: https://github.com/seiji-shimizu/qa-start-ordering.

Keywords: Temporal Relation Annotation, Clinical Domain, Annotation Method, Collaborative Annotation

1. Introduction

Reasoning the temporal relationships between events is essential for natural language understanding. Temporal reasoning is especially important in the clinical domain to discover potential causes of events. For example, to discover the cause of an adversarial drug effect, knowing which drugs were prescribed before the appearance of the target symptom is necessary.

There have been attempts to create open datasets for temporal relation extraction in the clinical domain, such as the i2b2 2012 (Sun et al., 2013) and the THYME (Styler IV et al., 2014) corpus. Still, the research community lacks annotated corpora, with most studies being conducted using only one dataset (Gumiel et al., 2021). Also, many languages other than English have no open dataset, including Japanese.

Two major factors that hinder the development of datasets are (1) the workload of temporal relation annotation and (2) the medical expertise required for the annotation. Clinical texts contain many events per document (147 per document in THYME and 94 per document in i2b2 2012, for instance). Thus, a large number of relations need to be annotated to cover all event pairs. Additionally, temporal relation annotation is complex because annotators need to consider various types of temporal relations (13, according to Allen's definition; Allen, 1983). Also, accurate annotation requires medical expertise. Hiring medical experts for the entire annotation process is costly and difficult since medical experts, such as doctors and nurses, have their primary duties.

To alleviate such challenges, we propose a new method for temporal relation annotation, integrating event start-points ordering (Ning et al., 2018) and question-answering (QA) as the annotation format (Ning et al., 2020). An overview of our annotation process is shown in Figure 1.

Given two events, ordering their start-points can simplify the temporal relation set to: *before*, *equal*, and *after*. We only annotated an event as "*vague*" when the relation of the event with any other event in a document cannot be annotated. This significantly reduces the number of event pairs annotated as "*vague*" and thus, we can infer a larger number of non-vague relations by iteratively applying closure (Verhagen, 2005).

QA as annotation recasts the temporal relation annotation task into a reading comprehension task, allowing annotators to use natural language instead of the formalisms commonly adopted in temporal relation annotation (Ning et al., 2020). In our method, in the **first step**, a non-medical expert generates a question about the start-points of events by choosing an event from a document. Then, the non-expert generates answers for the questions by choosing an event from the document (**QA-Gen**). This allows an annotation task without pre-defining event pairs to annotate. In thet **second step**, a medical expert validates the generated QA pairs using multiple-choice questions (**QA-Val**). The non-

LREC-COLING 2024, pages 13371–13381

medical expert can decide which event pairs should have temporal relations from linguistic cues even without medical expertise. Then, the medical expert can easily validate the relations annotated by non-experts since the relations are described in the QA format.

We show that our proposed method can reduce (1) the workload of temporal relation annotation and address the issue of (2) the medical expertise required for the annotation. Based on our proposed method, 90% of the relations among all possible event pairs are inferable from only 9% explicitly annotated relations, showing the efficiency of our approach. We empirically show the feasibility of using these inferred relations as training data in our experiment, resulting in a 0.72 F1 score for the classification of two start-points. We compare two results of **QA-Gen**: (a) QA pairs generated by a medical expert and (b) QA pairs generated by a non-expert, using (a) as the gold set. Our QA generation method is simple enough for a non-expert to approximate event pairs chosen by a medical expert. **QA-Val** by an additional medical expert increases the IAA of the QA pairs generated by a medical expert and a non-expert. The increased inter-annotator agreement (IAA) indicates the positive prospect of collaboration among experts and non-experts.

To the best of our knowledge, this is the first attempt to integrate event start-points ordering and QA as annotation for temporal relation annotation. Although we focused on the clinical domain, our method is domain and language-agnostic. The proposed annotation method can be effective for the temporal relation annotation of corpora with a large number of events. Also, it allows the collaboration of domain experts and non-experts. Our contributions can be summarized as follows:

- We propose a novel annotation method that enables simple and efficient annotation and collaboration among medical experts and non-experts.
- We empirically prove the feasibility of the proposed annotation method and corpus in our experiments.
- We created publicly open language resources: (i) the first temporal relation extraction dataset in Japanese; (ii) the annotation tool we built for **QA-Gen** and **QA-Val**, and (iii) the temporal relation extraction model trained on the developed corpus.

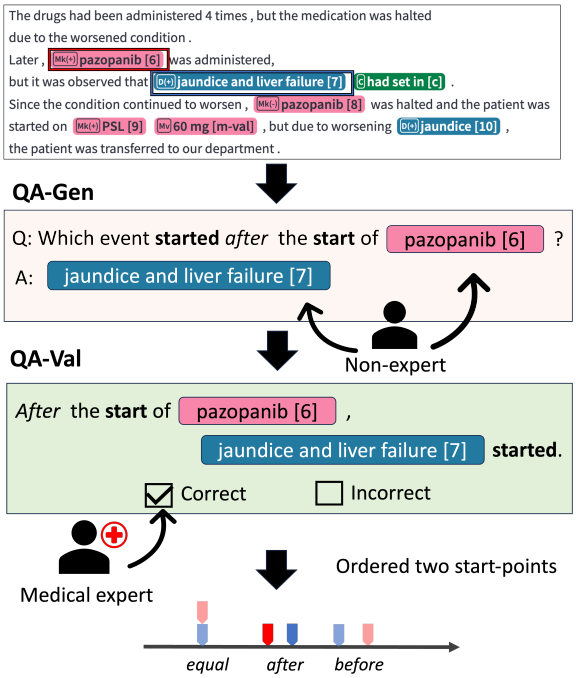

Figure 1: Overview of the annotation process. Events are marked in colors. A non-medical expert generates a QA pair in **QA-Gen** by simply choosing one event for the question and one for the answer from a document. Then, a medical expert validates the generated QA pair with a multiple-choice question in **QA-Val**. The relation of two event start-points is classified as *before, equal*, or *after*. If the relation between an event and any other event in a document can not be annotated, we annotate the event as "*vague*".

2. Related Work

2.1. Temporal Relation Annotation in the Clinical Domain

There is a long-standing interest in creating open datasets for the temporal reasoning task in the clinical domain. Extending the TimeML scheme (Pustejovsky et al., 2010), the i2b2 2012 corpus and the THYME corpus were developed.

i2b2 2012 (Sun et al., 2013) annotated the event to event and event to time expression temporal relations. Eight types of relations were annotated, and the BEFORE, AFTER, and OVERLAP relations were used for the shared task. In total, 61,214 relations were annotated for 310 discharge summaries. Although the corpus covers various relations, the complexity of the annotation task led to a low IAA score (0.3 kappa score for the relation type). Recently, Guan et al. (2021) utilized RoBERTa (Liu et al., 2019) for a subset of this task (event to time expression relation classification) and achieved a 0.72 F1 score.

The THYME corpus (Styler IV et al., 2014) em-

ployed the "narrative container" to reduce the annotation workload and complexity. Each event is associated with a container using the "CONTAINS" relation, and each container is linked to an anchor (dates or temporal expressions). Although five relation types were included in the original annotated corpus, only "CONTAINS" was used for the shared task. In total, 7,935 relations were annotated for 107 pathology reports. Extending the dataset with inferred relations, Lin et al. (2019) achieved a 0.68 F1 score using BioBERT (Lee et al., 2019).

The PRISM corpus (Yada et al., 2020, 2021; Cheng et al., 2022) annotated temporal relations of event to time expression pairs. Based on THYME, five relation types were considered. They annotated 13,884 relations for 1,000 radiology interpretation reports and 5,432 relations for 156 medical history reports. Classification models for each document type were developed. For the "on" relation, the models achieved a 0.82 F1 score in radiology interpretation reports and a 0.70 F1 score in medical history reports. Due to the lack of annotated events, the models achieved F1 scores below 0.5 in relations other than "on".

All of the aforementioned annotation methods resulted in a large number of annotated relations, making them not feasible for organizations with lower resources (e.g., individual hospitals). This hinders the development of datasets based on those methods. A lighter annotation method is necessary to promote the development of various datasets across different medical specialties and different text types. Currently, most of the publications are based on a single dataset (Gumiel et al., 2021). This could be a limitation since domain shift is common in the clinical domain (Laparra et al., 2020).

Also, the narrative containers used in THYME and PRISM cannot capture which event precedes the other within a container. In applications, this can be a drawback, for example, for capturing the potential cause of acute adversarial effects of drugs happened in the same narrative container.

2.2. Simplification of Temporal Relation Annotation

There have been attempts to simplify temporal relation annotation. Two schools of such attempts are: (1) **event start-points ordering** and (2) **QA as annotation**. The former simplifies the classification of two time-intervals into the classification of two event start-points. The latter simplifies the task of temporal relation annotation by recasting it into a reading comprehension task.

In **event start-points ordering**, the label set is simply *before*, *equal* and *after* (Ning et al., 2018). Raghavan et al. (2012) employed start/end-points ordering of all events within a clinical document. They showed that all relations in Allen's temporal relations can be expressed by ordering the start/end-points of two events. For example, the temporal relation of two events $(e_1, \textit{before}, e_2)$ can be represented by ordering the start-points ($e_1^{(s)}$ and $e_2^{(s)}$) and the end-points ($e_1^{(e)}$ and $e_2^{(e)}$) of those events, for example $(e_1^{(s)}, \textit{after}, e_2^{(s)})$ and $(e_1^{(e)}, \textit{after}, e_2^{(s)})$. MATRES (Ning et al., 2018) also used start-points ordering for temporal relation annotation. They only annotated start-points orders based on the observation that end-points are a major source of confusion for annotators. Following this, we employ start-points ordering for our annotation task. Unlike Raghavan et al. (2012), they adopt a dense annotation scheme, which is the annotation of all event pairs within a two-sentence window. Given event pairs to be annotated, they reported a high IAA (0.84 kappa) for relation annotation.

QA as annotation is proposed to transform the relation extraction task into a reading comprehension task (Levy et al., 2017; He et al., 2015; Michael et al., 2018). TORQUE (Ning et al., 2020) applied QA as annotation for temporal relation extraction. In this method, annotators were instructed to create a temporal question and choose all the answers from the list of events within two-sentence windows. It asks the annotator to choose event pairs to be annotated and to describe the relation in natural language. This allows an annotation without predefining event pairs or relation categories.

Event start-points ordering reduces the ambiguity of a temporal relation of two events, simplifying the classification of temporal relation. QA as annotation recasts temporal relation annotation into a reading comprehension task, allowing annotators to use natural language instead of the formalisms commonly adopted in temporal relation annotation. In this paper, we propose a way to integrate those two and apply it to clinical documents. To the best of our knowledge, this is the first attempt to integrate the QA as annotation and event start-points ordering for temporal relation annotation.

3. Corpus to Annotate

We chose MedTxt-CR-JA (Yada et al., 2022) as the annotation target. MedTxt-CR-JA is a collection of 148 Japanese case reports with entities and their attributes annotated based on Yada et al. (2020). We chose case reports as annotation targets as they describe holistic clinical narratives (similar to discharge summaries) and can make them public without de-identifications.

We extracted medication, remedy, disease, and test values as the events among other entity types, such as anatomical entities, considering these entity types are essential for constructing a compre-

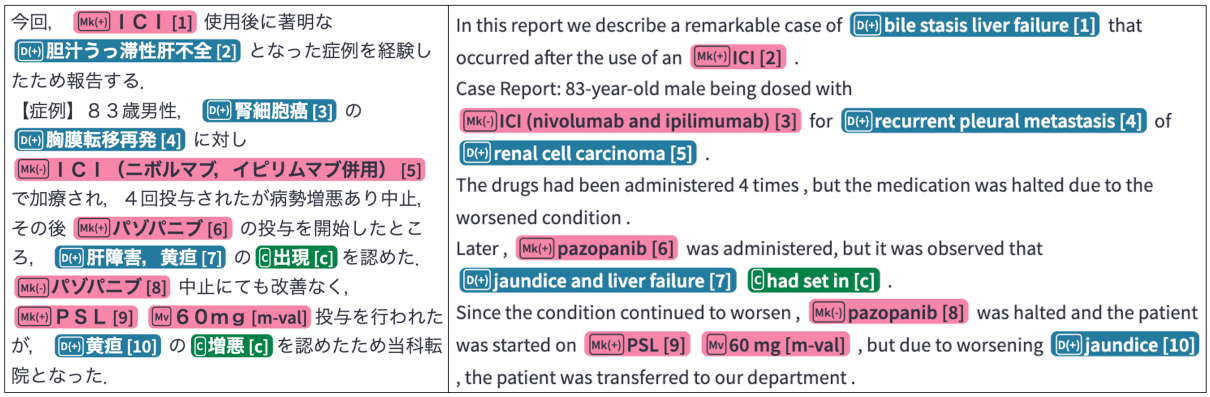

Figure 2: Example of a case report annotated in our study. The left part shows the original Japanese text and the right part shows its translation in English. Medication (Mk) and disease (D) are marked with pink and light blue, respectively. (+) and (-) represent the affirmation and negation of the event occurrences. Segments marked with green are changes that happened to the events which help annotators to understand the context of events. Event indices according to the orders in the document are attached.

hensive clinical timeline. We excluded negated or hypothetical entities with some exceptions. For example, if a medication was stopped being used, we considered that as an event, even if negated. Also, for the purpose of capturing the temporal dynamics of the four target event types, we only included the reports with:

- $\frac{\#remedy+\#medication}{\#disease+\#test} \geq 0.3$
- The number of events per document is greater than 10 and smaller than 30.

After the filtering, we obtained a total of 62 case reports. Figure 2 shows an example of a fraction of a target document. Annotation is done on the entirety of a single document. In the annotation, we displayed different types of events with corresponding markers. We also attached an event index to each event according to the order of appearance in a document.

4. Annotation Method

Given a clinical document, our aim is to reason temporal relations of as many event pairs as possible from the document. Following Ning et al. (2018), we reduce the temporal relation annotation to start-points ordering from the complex classification of two temporal intervals. Given two events, we classify the relation of two start-points into "*after*," "*equal*," or "*before*" (see Figure 1). If the relation of an event with any other event cannot be inferred, we treat it as "*vague*". Formally, two start-points and their relation can be represented as $(e_1^{(s)}, r, e_2^{(s)})$, where r = *{after, equal, before}*. Compared with the classification of two intervals, which requires 13 types of relations in the most complex case, this significantly simplifies the types of relations to consider.

For the annotation of start-points, we use QA as annotation, recasting temporal relation annotation into a reading comprehension task. We decompose the QA as annotation into (1) QA generation (**QA-Gen**) and (2) QA validation (**QA-Val**) and developed a system for each annotation process based on Maufe et al. (2022). We detail these two processes in the following section.

4.1. QA-Gen

In **QA-Gen**, we ask three questions to classify the temporal relation of two start-points:

1. Wich event **started** *after* the **start** of x?
2. Which event **started** *around the same time* as the **start** of x?
3. Which event refers to the *same event* as x?

The first question and the second question above are equivalent to the questions on "*after*" and "*equal*". Also, we ask a question concerning coreference (the third question). Though the temporal relation of the two start-points is "*equal*" in the third question, this makes the task more naturally understandable. For example, in Figure 2, it is more natural to say "*ICI [2] refers to the same event as ICI (nivolumab and ipilimumab) [3]*" rather than describing that they started at the same time. We omit the question on "*before*" because it can be represented using "*after*" with two events in the tuple being swapped: $(e_1^{(s)}, \textit{before}, e_2^{(s)}) = (e_2^{(s)}, \textit{after}, e_1^{(s)})$.

We represent a question as $(x^{(s)}, r, ?^{(s)})$. Annotators fix the "r" to work on by choosing one question from the above three questions. For example, if the first question is chosen, an annotator works on $(e_1^{(s)}, \textit{after}, e_2^{(s)})$. Then, the annotator first generates a question (Q) by choosing "x" and then generates

an answer (A) by choosing "?" in the tuple. The procedure is summarized in Figure 3.

We instruct annotators to cover as many events as possible from a document in either "x" or "?" in the tuple. Only when there is an event whose relation is vague to all the other events, we do not generate the QA pair for that event and annotate the event as "*vague*". For example, when the same medication was administered multiple times, and that medication name was mentioned in the later part of the case report without specifying the particular administration being referred to, an annotator can annotate this reference to the medication as "*vague*".

Generating the questions: In **QA-Gen**, first, annotators generate a question by choosing an event for "x" in the tuple, $(x^{(s)}, r, ?^{(s)})$. For example, by choosing "x" = "*pazopanib*", an annotator generates the question, "Which event **started** *after* the **start** of *pazopanib*?".

In the annotation system we developed, annotators can choose "x" from multiple options in a pull-down list that contains all the events in a document.

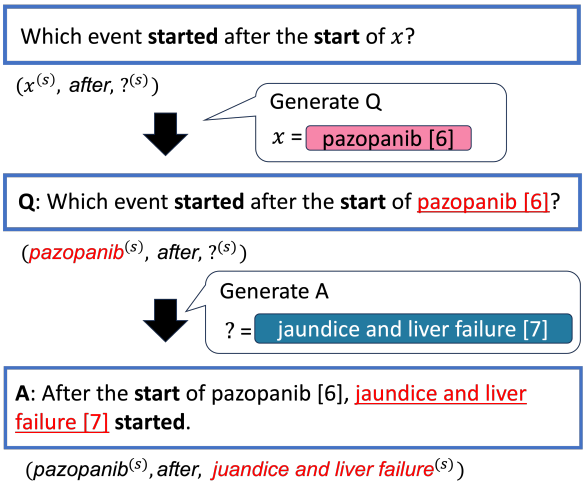

Figure 3: Process of **QA-Gen**. A non-expert annotator first chooses "x" (e.g., *pazopanib*) to generate a question (Q), and then the annotator chooses "?" (e.g., *jaundice and liver failure*) to generate the answer (A).

Generating the answers: Given a question with "x" being selected in $(x^{(s)}, r, ?^{(s)})$, annotators generate answers by choosing events for "?" from a document. For example, to the question "Which event **started** after the **start** of *pazopanib* ?", an annotator generates the answer "After the **start** of *pazopanib*, *jaundice and liver failure* started" by choosing "?" = "*jaundice and liver failure*". This completes the tuple $(e_1^{(s)}, r, e_2^{(s)})$. In the annotation system, annotators can choose "?" from multiple options in a pull-down list of all the events in a document.

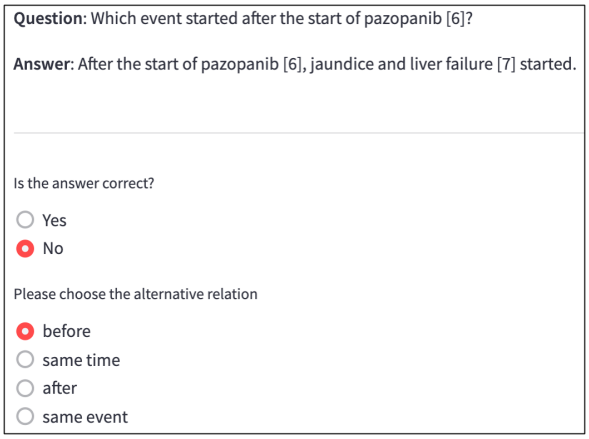

Figure 4: Screenshot of our proposed validation tool. Sentences are translated from Japanese into English. Given QA pairs previously annotated by a non-expert, a medical expert validates these pairs with multiple-choice questions.

Converting temporal relation annotation into reading comprehension encourages the annotators to choose event pairs using linguistic cues in a document. In this way, we do not need to pre-define pairs of "e_1" and "e_2" to be annotated in $(e_1^{(s)}, r, e_2^{(s)})$.

4.2. QA-Val

We validate the generated QA pairs with multiple-choice questions. Given the completed tuple $(e_1^{(s)}, r, e_2^{(s)})$, we ask a medical expert if the answer is correct or not, using the annotation tool we shown in Figure 4. For example, we ask whether the answer "After the start of *pazopanib*, *jaundice and liver failure* started" is correct or not. If the annotator chooses "No," we ask to modify the "r" by choosing the most appropriate relation among all options.

5. Corpus Statistics

We report the number of annotated relations in **QA-Gen** and the IAA for **QA-Gen** and **QA-Val**. For the calculation of IAA, we randomly sampled 10 documents from the original 62 case reports.

In addition to the number of explicitly annotated relations (Expl), we report the number of inferred relations (Infer). Although we only annotated a portion of all possible event pairs in the annotation, in practice, we want to infer the relation of any arbitrarily chosen event pair from a document. Given arbitrary two events, their relation can be "*before*", "*equal*", or "*after*". To cover as many event pairs as possible, we infer "r" in $(A^{(s)}, r, C^{(s)})$ from $(A^{(s)}, r, B^{(s)})$ and $(B^{(s)}, r, C^{(s)})$. More concretely, we create a graph (Figure 5) from the annotated

"*after*" and "*equal*" relations and infer the relations of unannotated event pairs. Given two events, their relation is "*after*" if "e_1" is connected with "e_2" with a forward-directed path and "*before*" if connected with a reverse-directed path. If they are connected with equalities, the relation is "*equal*". If the relation cannot be inferred, we considered them as "*vague*".

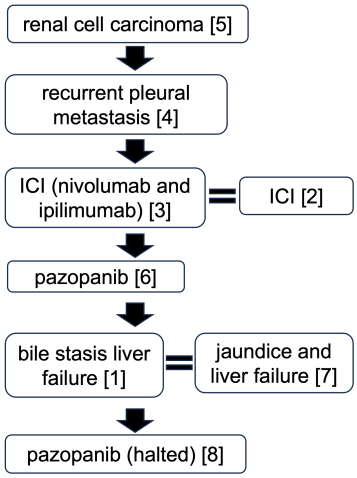

Figure 5: Example of the graph used for the inference of the relations in Figure 2. The arrows represent "*after*" and equalities represent "*equal*" relations. The numbers attached to events indicate the order of appearance of the events in the translated document.

5.1. The Number of Annotated Relations

We report the number of annotated relations in **QA-Gen**. We asked a medical expert (a nurse with more than 10 years of experience) and a non-expert to complete the task for comparing the annotation done by an expert and a non-expert. In total, each annotator was assigned 32 documents (including those for the calculation of IAA). It took about a month for both of the annotators to complete the annotation.

For Infer, we create a set of all possible event pairs and then inferred relations of the entire set on the graph as shown in Figure 5. For the overlapping 10 documents, we counted the exact same event pairs and relations only once, i.e., the exact same tuples $(e_1^{(s)}, r, e_2^{(s)})$ were not counted twice.

Table 1 shows the number of annotated relations for each relation category. Since inference on a graph itself is complex, 5% of "*after*" could not be converted into "*before*". We conduct an experiment (Section 6) using Infer to test its utility for developing a temporal relation classifier.

Using our method, most of the relations in a document are inferable with a relatively small number of explicitly annotated relations. The number of all possible event pairs obtained was 27,413, the explicitly annotated relations excluding "*vague*" were 2,554 (9% of all possible event pairs), and the inferable relations excluding "*vague*" were 24,737 (90% of all possible event pairs). In Lin et al. (2019), the number of inferred relations from the THYME corpus using a 60-token window and closure were 413,327 of NONE, 10,483 of CONTAINS, and 2,802 of CONTAINS-BY. Most relations could not be inferred, resulting in "NONE" consisting of 93% of all relations. We could infer most relations in a document since the number of "*vague*" in Expl was very small, and most events were included in the graph and used for the inference.

5.2. Agreement for QA-Gen

We compare two results of **QA-Gen**, QA pairs generated by the medical expert and QA pairs generated by the non-expert, using the pairs generated by the expert as a gold set. We calculated the inter-annotator agreement (IAA) of **QA-Gen** using 10 documents randomly sampled. As metrics, we used Krippenford's alpha and F1 score, following THYME (Styler IV et al., 2014). We did not use the inferred relations for the calculation of IAA.

Table 2 shows the IAA of Expl. "pairs" denotes the IAA for which event pairs were chosen, and "pairs + relation" denotes the IAA, including the relation types of the chosen event pairs. Though not directly comparable, we included those two IAA of the THYME corpus in the table for reference.

As one can see, in spite of the difference in expertise of the two annotators, the IAA shows the agreement on which event pairs to annotate. The high agreements indicate **QA-Gen** is simple enough for a medical expert and a non-expert to agree on which event pairs to annotate and their relations.

5.3. Agreement for QA-Val

We calculated the IAA of **QA-Val** using event pairs annotated by both the expert and the non-expert annotator in **QA-Gen** (223 pairs in total). We validated and modified the relations of event pairs using **QA-Val** only for the annotation done by the non-expert and compared the result with the annotation done by the expert. We asked an additional medical expert (a nurse also with more than 10 years of experience) to complete the task. The annotation took less than a day to complete.

As metrics, we used Cohen's kappa and F1 score following MATRES (Ning et al., 2018). Table 3 shows the IAA scores.

After **QA-Val**, IAA improved from 0.73 to 0.89 in Kappa and from 0.82 to 0.94 in F1. This indicates the utility of **QA-Val** and the positive prospect of collaboration among medical experts and non-experts in the annotation process.

	before	equal	after	vague	total
Expl	-	1,848	696	19	2,563
Infer	11,440	2,484	10,813	2,676	27,413

Table 1: The number of relations annotated in **QA-Gen**. Expl denotes the relations explicitly annotated by annotators. Infer denotes the relations inferred from explicitly annotated relations.

	Alpha	F1
pairs (THYME)	0.50	0.50
pairs + relation (THYME)	0.45	0.45
pairs (ours)	0.54	0.59
pairs + relation (ours)	0.51	0.64

Table 2: The IAA of **QA-Gen**. “pairs" denotes the IAA scores for which event pairs were chosen. “pairs + relation" denotes the IAA scores for the chosen event pairs and their relations.

	Kappa	F1
MATRES	0.84	0.90
w/o **QA-Val**	0.73	0.82
w/ **QA-Val**	0.89	0.94

Table 3: The IAA of **QA-Val**. “w/o **QA-Val**" denotes the scores without validation, and “w/ **QA-Val**" denotes the scores after the validation.

6. Experiment

6.1. Settings

To investigate the feasibility of using inferred relations to train a high-performance temporal relation classifier, we conducted an experiment using a subset of Infer. Given a document and two start points ($e_1^{(s)}$ and $e_2^{(s)}$), we classify their relation into r = *{after, equal, before}*. We excluded “*vague*” from the classification target because most of the relations could be categorized into “*after*", “*equal*" or “*before*". The “vague” relation consisted of only 3% of all the inferred relations.

Dataset: We used inferred relations as dataset. To avoid data imbalance, we down-sampled “*after*” and “*before*” to be the same size as “*equal*” per document. For each $(e_1^{(s)}, r, e_2^{(s)})$ in the dataset, we tagged two events in a document with start and end tags (<event1>,</event1> and <event2>,</event2>) as special tokens, signaling which event pair we are predicting on to the model. We used these tagged documents as inputs and the relations as target labels.

Method: Our method is based on Zhong and Chen (2021). We used a Japanese RoBERTa base model (Sugimoto et al., 2023) pre-trained on Japanese case reports as the encoder. Given a document with two events tagged, we encoded the document and used the final hidden layers' representations corresponding to the tokens in each event mention for the prediction. Specifically, we take the average of the outputs within start and end tags (<event1> to </event1> and <event2> to </event2>):

$$\boldsymbol{z_{event1}} = average([\boldsymbol{h}_{<event1>}, ..., \boldsymbol{h}_{</event1>}])$$
$$\boldsymbol{z_{event2}} = average([\boldsymbol{h}_{<event2>}, ..., \boldsymbol{h}_{</event2>}])$$

where $\boldsymbol{h}_{<event1>}$ denotes the encoder output for the start tag, and $\boldsymbol{h}_{</event1>}$ denotes the one for the end tag. Then, we concatenate the two average vectors and feed them to a multi-layer perceptron (MLP):

$$P(r|e_1, e_2) = softmax(MLP(\boldsymbol{z_{event1}} \oplus \boldsymbol{z_{event2}}))$$

where $\oplus$ denotes the concatenation operator. We used the highest value after the softmax operation as predictions. Hyper-parameters used in the experiment are training epoch as 10, batch size as 8, and AdamW Optimizer with learning rate as 1.0×10^{-5}. We used 20% of the training data to calculate the training loss and used the model checkpoints with the lowest loss for the evaluation per validation fold.

6.2. Evaluation

Table 4 shows the average scores of 5-fold cross-validation at the document level. We calculated accuracy, precision, recall, and F1 scores (macro averages) for the total scores. In “*after*” and “*before*”, the trained model achieved above 0.75 F1 scores. The classification of “*equal*” was harder than those two relations, with a F1 score of 0.63. Overall, the model achieved a 0.72 F1 score. Guan et al. (2021) reported a 0.72 F1 score using RoBERTa trained on a subset of i2b2 2012, and Lin et al. (2019) reported a 0.68 F1 score using BioBERT trained on THYME. While not directly comparable, our experimental results fall within a moderate range among those results. As for start-point ordering, Ning et al. (2018) reported a 0.77 overall F1 score on TimeBank-Dense (Cassidy et al., 2014) using two-sentence windows. Our scores are comparable to this, considering that we tackled

Relation	Accuracy	Precision	Recall	F1
before	0.83	0.77	0.75	0.76
equal	0.80	0.61	0.66	0.63
after	0.82	0.77	0.74	0.75
total	0.72	0.72	0.72	0.72

Table 4: Performance of the trained model on Infer. Scores are averaged across 5-fold validation. For the total scores, F1, Recall, and Precision are macro-averaged for each fold and averaged across 5-fold.

a more challenging scenario involving classification beyond two-sentence windows. The high performance of the trained model indicates the feasibility of this corpus.

7. Analysis of the Annotated Corpus

We analyze the difference between the number of event pairs chosen by the expert and the non-expert in **QA-Gen**. Also, we exemplify the modification done by the other expert in **QA-Val**.

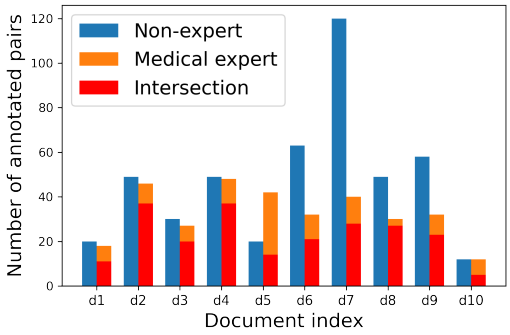

Figure 6: The number of annotated event pairs in 10 documents. “Non-expert" and “Medical expert" denote the number of annotations done by a non-expert and an expert, respectively. “Intersection" denotes the number of event pairs annotated by both annotators. Most of the event pairs chosen by the expert annotator are also chosen by the non-expert annotator.

Figure 6 shows the number of the annotated event pairs for each annotator and the number of event pairs annotated by both of the annotators (Intersection) in **QA-Gen**. As we can see, for most documents, the non-expert annotated the larger number of event pairs. Also, 69% of the chosen event pairs by the expert are also chosen by the non-expert on average. This implies that we can fairly approximate (1) the annotation entirely done by medical experts through (2) **QA-Gen** done by non-experts and **QA-Val** done by medical experts.

We manually analyzed the modifications made by the medical expert in **QA-Val**. We exemplify some of the representative cases where medical expertise was necessary for the annotation. Example sentences are shown in Figure 7.

Example 1: 内視鏡的治療を施行、ENPD留置目的に**ERCP**$_{(2)}$施行、体尾部の膵管は造影されず、**膵管癒合不全**$_{(1)}$を認めた。
(Endoscopic treatment was performed, **ERCP** $_{(2)}$ was performed for the purpose of ENPD implantation, the pancreatic duct at the tail was not contrasted, and **pancreatic duct fusion failure**$_{(1)}$ was observed.)

Example 2: 膵全摘及び自家膵移植、困難であれば自家膵島移植とし8月10日 **手術**$_{(2)}$施行。**癒着**$_{(1)}$は激しかったが、膵全摘＋自家膵移植を行った。
(We decided to **perform** $_{(2)}$ a total pancreatectomy and autologous pancreas transplantation, or autologous islet transplantation if difficult, on August 10. Total pancreatectomy plus autologous pancreas transplantation was performed, although **the adhesions** $_{(1)}$ were severe.)

Example 3: 同年5月、両側の**尿管狭窄**$_{(1)}$を認め. **腎後性腎不全**$_{(2)}$の診断で緊急入院した。
(In May of the same year, she was admitted to the hospital in an emergency with a diagnosis of **postrenal renal failure** $_{(2)}$ due to bilateral **ureteral structure** $_{(1)}$.)

Figure 7: Examples of event pairs whose relation is correctly modified in **QA-Val** by a medical expert. The two events in the event pair are shown in **bold**.

In **Example 1**, the non-expert annotated e_1 = “*pancreatic duct fusion failure*”, e_2 = “*ERCP*”, and $r =$ ”*equal*” for $(e_1^{(s)}, r, e_2^{(s)})$, then the medical expert modified it to $r =$ “*after*”. In this example, we can infer that two events happened around the same time from linguistic cues (sentences are connected by the conjunction “and"). Yet, “*ERCP*" is a procedure that combines upper gastrointestinal endoscopy and x-rays. Therefore, the procedure should come after “*pancreatic duct fusion failure*". This requires knowledge of the nature of “*ERCP*" as a procedure.

In **Example 2**, e_1 = “*adhesions*” and e_2 = “*perform*” was annotated as r = “*equal*” by the non-expert, and modified to be r = “*after*” by the expert. The event “*perform*" refers to the two operations mentioned in the next sentence. The annotator needs to know those operations do not cause “*adhesions*" in this case and should have started *before* the operation.

In **Example 3**, e_1 = “*ureteral stricture*” and e_2 = “*postrenal renal failure*” was annotated as r = “*equal*” by the non-expert and modified to be r = “*after*” by the expert. The annotation requires knowledge of the causal relation between “*ureteral stricture*" and “*postrenal renal failure*" and modified correctly in the validation.

All the above examples showcase the temporal relation annotation of clinical documents does require medical expertise, and **QA-Val** can properly compensate for the annotator's lack of expertise in **QA-Gen**.

8. Conclusion

We proposed a novel annotation method that integrates event start-points ordering and QA as annotation. We annotated clinical documents and showed that most of the temporal relations in a document can be automatically inferred, indicating the efficiency of our annotation method. Also, we showed that decomposing the annotation process into **QA-Gen** and **QA-Val** enables collaboration among medical experts and non-experts. Our experimental results suggest the utility of using inferred relations for developing a temporal relation classifier. The developed corpus (containing both the relations explicitly annotated by annotators, i.e., Expl, and the relations inferred from the explicitly annotated relations, i.e., Infer), developed annotation tools, and trained model are all openly available[1]. The proposed annotation method can be applied to the temporal relation annotation of corpora in other domains, which have a large number of events per annotation target and are costly to annotate due to the required expertise.

9. Ethics Statement

The corpus annotated in this research comprises case reports that have been anonymized, ensuring there is no possible threat to patient confidentiality. The annotators participating in the annotation process for this study received generous compensation. Non-expert annotators were remunerated at a rate equivalent to the average hourly wage in Japan, while expert annotators received triple the standard payment.

10. Limitations

In this paper, we chose case reports as the annotation targets. Unlike inner-hospital text, such as discharge summaries, case reports are well-written so readers can easily understand. The degree to which inner-hospital text has its own unique writing style differs among hospitals. This could be a limitation of this study when we try to apply our method to inner-hospital texts with only a few linguistic cues. Also, the number and the length of documents covered are a limitation of our study. Though we tested our trained model using the entire corpus by 5-fold cross-validation, extending test data distribution is desirable for further generalizability of our study. Finally, annotation on clinical texts written in English and end-points ordering is necessary for the direct comparison of our annotation method and existing methods. Although our annotation method needs no major modification to be applied to such annotations, we leave them for future work.

[1] https://github.com/seiji-shimizu/qa-start-ordering

11. Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR22N1, and Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425, Japan. The authors are grateful to Takako Fujimaki, Mutsumi Nakae and Kyoko Kawabata for their contributions to the annotation task.

12. Bibliographical References

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506.

Fei Cheng, Shuntaro Yada, Ribeka Tanaka, Eiji Aramaki, and Sadao Kurohashi. 2022. JaMIE: A pipeline Japanese medical information extraction system with novel relation annotation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3724–3731, Marseille, France. European Language Resources Association.

Hong Guan, Jianfu Li, Hua Xu, and Murthy Devarakonda. 2021. Robustly pre-trained neural model for direct temporal relation extraction. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 501–502. IEEE.

Yohan Bonescki Gumiel, Lucas Emanuel Silva e Oliveira, Vincent Claveau, Natalia Grabar, Emerson Cabrera Paraiso, Claudia Moro, and Deborah Ribeiro Carvalho. 2021. Temporal relation extraction in clinical texts: A systematic review. *ACM Comput. Surv.*, 54(7).

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natu-

ral language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, and Dongyan Zhao. 2023. More than classification: A unified framework for event temporal relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9631–9646, Toronto, Canada. Association for Computational Linguistics.

Egoitz Laparra, Steven Bethard, and Timothy A Miller. 2020. Rethinking domain adaptation for machine learning over clinical language. *JAMIA open*, 3(2):146–150.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pre-training approach.

Matt Maufe, James Ravenscroft, Rob Procter, and Maria Liakata. 2022. A pipeline for generating, annotating and employing synthetic data for real world question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 80–97, Abu Dhabi, UAE. Association for Computational Linguistics.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 560–568, New Orleans, Louisiana. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Preethi Raghavan, Albert Lai, and Eric Fosler-Lussier. 2012. Learning to temporally order medical events in clinical text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 70–74, Jeju Island, Korea. Association for Computational Linguistics.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Kaito Sugimoto, Taichi Iki, Yuki Chida, Teruhito Kanazawa, and Akiko Aizawa. 2023. JMedRoBERTa: a japanese pre-trained language model on academic articles in medical sciences (in japanese). In *Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing*.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Marc Verhagen. 2005. Temporal closure in an annotation environment. *Language Resources and Evaluation*, 39(2/3):211–241.

Shuntaro Yada, Eiji Aramaki, Ribeka Tanaka, Fei Cheng, and Sadao Kurohashi. 2021. *Medical/Clinical Text Annotation Guidelines*.

Shuntaro Yada, Ayami Joh, Ribeka Tanaka, Fei Cheng, Eiji Aramaki, and Sadao Kurohashi. 2020. Towards a versatile medical-annotation guideline feasible without heavy medical knowledge: Starting from critical lung diseases. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4565–4572, Marseille, France. European Language Resources Association.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

13. Language Resource References

Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. 2022. Real-MedNLP: Overview of real document-based medical natural language processing task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, pages 285–296.