Annotation Study of Japanese Judgments on Tort for Legal Judgment Prediction with Rationales

Hiroaki Yamada[1], Takenobu Tokuanaga[1], Ryutaro Ohara[2]
Keisuke Takeshita[3], Mihoko Sumida[3]
Tokyo Institute of Technology[1], Nakamura, Tsunoda & Matsumoto[2], Hitotsubashi University[3]
Tokyo, Japan
yamada@c.titech.ac.jp, take@c.titech.ac.jp, r.ohara@ntmlo.com
kei.takeshita@r.hit-u.ac.jp, m.sumida@r.hit-u.ac.jp

Abstract

This paper describes a comprehensive annotation study on Japanese judgment documents in civil cases. We aim to build an annotated corpus designed for Legal Judgment Prediction (LJP), especially for torts. Our annotation scheme contains annotations of whether tort is accepted by judges as well as its corresponding rationales for explainability purpose. Our annotation scheme extracts decisions and rationales at character-level. Moreover, the scheme can capture the explicit causal relation between judge's decisions and their corresponding rationales, allowing multiple decisions in a document. To obtain high-quality annotation, we developed an annotation scheme with legal experts, and confirmed its reliability by agreement studies with Krippendorff's alpha metric. The result of the annotation study suggests the proposed annotation scheme can produce a dataset of Japanese LJP at reasonable reliability.

Keywords: Annotated corpus, Annotation scheme, Agreement study, Legal judgment prediction, Rationale extraction

1. Introduction

One of the objectives in legal information processing is to provide computational aid in the legal procedures in court cases. Legal Judgment Prediction (LJP), which predicts the outcome of a court case (Figure 1), is a crucial task to realise such a system. The automated LJP system can help not only legal professionals but also the general public who are not legal specialists. The system allows everyone to predict and foresee the outcome of litigation when involved in legal disputes. People can access the system wherever, whenever. Also, the anticipated cost of LJP will be much lower than that of human legal professionals. The system is expected to provide broader access to justice for people who have limited or no access to justice.

Although LJP has been a longstanding research topic in Artificial Intelligence, most large-scale datasets for LJP have been proposed only recently. With the increasing popularity of machine learning(ML) based approaches, various studies proposed larger datasets to train ML models. Unfortunately, the available resources are still limited to certain languages and jurisdictions. Xiao et al. (2018) proposed a dataset for Chinese Criminal cases (2.6M cases), which consist of annotations on applicable laws, charges, and prison terms. Chalkidis et al. (2019) presented 11.5K cases from the European Court of Human Rights. Their dataset was designed for violated article detection and case importance prediction. Katz et al. (2017) constructed a dataset from 28K cases of the Supreme Court of the United States. Chalkidis et al. (2021b) proposed a collection of datasets for evaluating model performance across different legal tasks including LJP in English. To train and assess LJP models, it is necessary to develop the LJP tasks and their datasets reflecting dif-
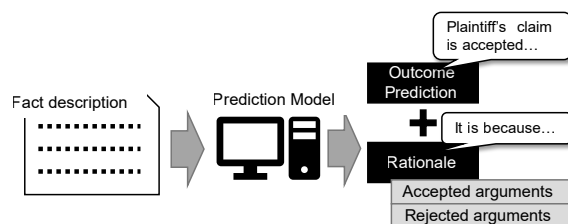

Figure 1: Legal judgment prediction with rationale extraction

ferences in jurisdictions. Here, we construct the first dataset of LJP for the Japanese judgments to provide a reliable dataset for the Japanese LJP research. As a first step, we develop an annotation scheme for the Japanese judgment in this paper.

Our primary objective is to provide an annotation scheme, which allows us to produce a reliable large-scale dataset for LJP and its rationale extraction. Our main contributions are the following. First, we introduce a novel annotation scheme designed for Japanese judgment documents. The scheme identifies the judicial decisions and their rationales. Rationales are extracted not only from facts but also from allegations and argumentations of parties involved in the proceedings. Our scheme can associate each rationale with its corresponding decision in a document with more than one issues. Hence, our scheme can provide direct causal relations between the court decisions and arguments from the parties allowing multiple court decisions in a case. Second, we conduct three annotation experiments featuring *torts*, which is an important subject in civil cases dealing with infringement of rights or legal interests that causes a plaintiff to suffer loss or harm. In this

annotation study, we use torts cases of defamation, privacy infringement and reputation injury. We describe our findings from the experiments and our revisions to improve the scheme. Third, we show that the final version of our scheme provides reliable annotations, tested on 25 documents with five annotators.

2. Background and Related Work

The approaches of LJP are roughly classified into two: symbolic systems and ML-based systems. Although symbolic systems require human experts' intervention in construction and maintenance, their behaviour is easy to understand. On the other hand, ML-based systems can automatically learn how judges make decisions from a large number of instances (e.g. judgment documents).

Recent studies of LJP actively employ ML-based models. A generic ML-based LJP model takes fact descriptions as input and predicts its outcomes or its relevant laws. Particularly, cases from the European Court of Human Rights are the popular subject for LJP (Aletras et al., 2016; Medvedeva et al., 2018; Chalkidis et al., 2019). Katz et al. (2017) proposed a dataset of cases from the Supreme Court of the United States and trained their ML models. Moreover, LJP on Chinese Criminal cases is another big venue of machine learing based LJP models (Luo et al., 2017; Zhong et al., 2018; Hu et al., 2018; Long et al., 2019; Xu et al., 2020).

On the other hand, in the Japanese LJP researches, most approaches heavily depended on symbolic systems. They infer outcomes of legal reasoning using rules and logic programming (Nitta et al., 1993b; Nitta et al., 1993a). One of the reasons why the Japanese LJP work have hardly employed ML-based approaches was the lack of a reliable dataset. Although there is a Japanese dataset for legal tasks provided for COLIEE (Rabelo et al., 2020), their dataset is designed for legal QA on the Japanese bar exam. We aim to construct a dataset with real Civil Code judgment documents to facilitate the LJP tasks. Therefore, we develop an annotation scheme to construct a dataset that facilitates the ML-based approach for the Japanese LJP.

Another reason ML approaches were not popular in the Japanese LJP was that the symbolic approaches had good compatibility with the rules of the Japanese Civil Code which define legal requirements. Proleg (Satoh et al., 2011) demonstrated the feasibility of logic programming based systems for the Japanese legal system. Proleg is a legal reasoning system based on Prolog implementing a decision-making theory used in civil litigation in Japan. However, there is still an open-ended problem, how to extract and transform natural language to logical clauses which are recognisable by reasoning engines (Navas-Loro et al., 2019). Moreover, some rules of the Japanese Civil Code do not specify the natural facts that must be proved to decide whether specific requirements for legal effects are fulfilled or not. They only provide general concepts as

requirements, and judges often need to evaluate relevant natural facts in a comprehensive manner to determine whether those requirements are satisfied or not. Such a rule is called *general clause*, and they cannot be explicitly expressed in rules or logic programming. On the other hand, ML-based approaches, which inductively learn standards of the general clause from many precedents, should perform better. Thus, constructing a dataset of Japanese judgment documents featuring *general clause* type rules at a large scale is indispensable. In this paper, we use *torts* cases to test our annotation scheme since their basic rules are provided by *general clause* rules in the Japanese Civil Code.

The success stories of deep learning methods across a broad range of tasks have called for the explainability issue (Jacovi and Goldberg, 2020) because the deep learning methods tell us little about the reasoning process leading to the output. The importance of explainability especially has to be taken more seriously in the legal domain than in other domains. Even though LJP systems are intended to be used as assistant tools for humans, they can affect people's behaviours concerning the use of the judicial system. The LJP system can indirectly influence people's social status and assets. Thus, an ideal LJP system has to explain the reason for output predictions. To this end, recent LJP studies introduced explanation tasks including court view generation (Ye et al., 2018), rationale paragraph extraction (Chalkidis et al., 2021a), and case features extraction as rationales (Ferro et al., 2019; Branting et al., 2021). Following the prior work, we include annotations of rationales similar to Chalkidis et al. (2021a), but our annotation is at character-level instead of paragraph-level. In addition, our annotation scheme can record explicit causal relations between judge's conclusions and their corresponding rationales, allowing multiple issues in a document.

3. Annotation Target

3.1. Japanese Judgment Documents

The Japanese judges are career judges, who are trained as judges right after passing the bar examination. The judges follow guidelines for writing judgment documents of civil cases. The style and structure of judgment documents are well-stabilised in the Japanese legal system (Kozuka, 2020). As a result, there is a high level of similarities in structure across judgment documents, most easily observed in a common section structure, often with similar headlines used.

This section structure is as follows: The *Main Text* is the first section covering main judgments, which render a final decision in a few sentences. The *Facts and Reasons* section takes up most of the document and is therefore the target of our annotation. *Facts and Reasons* consists of causes of action, followed by a summary of the case, facts not disputed among the parties to the proceedings, issues to be contested during the trial, and claims from the parties. The last part of *Facts*

and Reasons section contains the judicial decision in detail.

We can extract gold labels for the LJP task from the judge's reasoning and concluding sentences in the judicial decision part. At the same time, we can obtain input sentences from the neutral facts and alleged claims in the other parts. We leverage the section structure of the judgment documents to distinguish the court's decision from the other parts in designing our annotation scheme.

3.2. Subject of Annotation

We construct a dataset of judgment documents on civil cases about torts (Civil Code, Article 709)[1]. Tort is one of the common subjects in civil cases, and rules on torts are considered to be *general clause*. Under the Japanese law, tort liability is affirmed with infringement of rights or legal interests that causes a plaintiff to suffer loss or harm. Torts play an important role in disputes on the internet (e.g. defamation and privacy infringement on social media) because there is usually no contract between the parties in such a situation. Furthermore, torts in such disputes are emerging topics in the field of law since psychological and social damage online is an important issue in modern society (Sumida and Steffek, 2022). Our dataset may also provide useful material for law research.

Disputes related to the internet are often discussed in Disclosure of Identification Information of the Sender (DIIS) cases. DIIS is a mechanism provided by the Law on Limitation of Liability of Providers to enable an Internet user to demand the Internet Service Providers to disclose the sender's information (e.g. address and name) through trials.

We cover torts from DIIS cases in this case study. In addition to DIIS cases, we collected general tort cases which deal with defamation, privacy infringement and reputation injury[2]. They include the same topics as DIIS cases but their subjects are not contents on the internet but, for exmaple, contents on magazines and newspapers. Our data source of the judgment documents is a legal database "Hanreihishio[3]" provided by LIC Co., Ltd. We curate documents from the first instances of Civil Code cases. As a result, we collected 5,188 documents in total. 709 documents are from DIIS cases, and 4,478 cases are from general torts cases. Note that the database search system retrieves documents based on keywords queries so that there might be cases among the retrieved documents that in fact do not deal with the issue of tort. We, therefore, implement document screening (6.1) to exclude such

irrelevant documents.

To reduce bias in document curation, we should have collected all judgments from every court in Japan. However, not all Japanese judgments are provided in the machine-readable format from the courts. The current largest available source is the legal databases provided by publishing companies. We believe that the databases are still the best and the most reasonable data source among all possible options for now [4].

4. Concerns on sensitive data

As our target documents describe court cases, the documents can contain personal information, sensitive information of parties or legally protected information such as trade secrets. Leins et al. (2020) sheds light on potential ethical issues on constructing datasets from a sensitive data source like judgments.

In the Japanese legal system, the judgment document is an important legal document that is the direct output from court proceedings and contains the judgment, the facts and the grounds (Code of Civil Procedure, Article 252). Article 91 of Code of Civil Procedure guarantees anyone can inspect a case record, including the judgment itself. However, if documents on a case contain sensitive information about the parties' private life or technical/business information valuable for business activities, the judge can restrict access to the record upon the petition from interested parties (Code of Civil Procedure, Article 92).

In principle, the judgment documents are already in the public domain, and the parties may request to opt out of giving others access to their judgment documents. Therefore, theoretically, the sensitive secrets should not be contained in the database we use as our data source. Moreover, the publishing companies pseudonymise the documents before publishing a case in journals or databases.

Nevertheless, we still worry about the risk that some sensitive information might accidentally slip through many safety measures and the risk of dual-use. Considering the balance between the potential risks of sharing the data and the reproducibility of the study, we plan to share data only with researchers who agree with our strict terms of use.

5. Pilot Study

Chalkidis et al. (2021a) presented a task of paragraph-level rationale extraction for alleged violation prediction on cases from the European Court of Human Rights (ECtHR). Their tasks are multilabel classification to identify which European Convention on Human Rights (ECHR) articles are alleged in a case, and extraction of its rationales. They constructed a dataset of 11K ECtHR cases with annotations of alleged ECHR

---

[1]Yamamoto (2019) provides a good overview of the Japanese torts in English.

[2]Some torts cases also address other topics, including copyright and trademark infringement. In this case study, we exclude them since such cases contain a different type of legal reasoning more related to the intellectual property.

[3]https://www.hanreihisho.com/

[4]There is a project proposing to make the Japanese judgments open data. https://www.jlf.or.jp/wp-content/uploads/2021/04/pt-houkoku20210325.pdf

articles and rationale paragraphs. In their dataset, rationales are annotated by human experts only in 50 cases out of 1,000 cases from their test dataset without an agreement study, and rationales in the rest cases are automatically extracted leveraging references to facts of the cases. The references are, for example, "See paragraphs 2 and 4". They can be easily recongised by regular expressions. As the references are not available in our target documents, we need to annotate the rationales manually.

We conducted a pilot annotation on five documents to check the feasibility of manual annotation to extract rationales from the Japanese judgment documents. Also, we assessed the reliability of annotations among multiple annotators by inter-annotator agreement (IAA), which was not conducted in the previous work.

In the pilot study, we implement a simple annotation scheme to extract rationales that support court decisions on torts. This pilot version of the scheme has only one type of span, called Rationales.

The rationale spans identify important arguments (including factual allegations, legal arguments) from parties, which provide grounds for the judicial decision. In all versions of our annotation schemes, including the final version we describe later in 6, annotators are instructed to identify spans at character-level. We use a web-based annotation tool "tagtog"(Cejuela et al., 2014) in all of our annotation studies.

We randomly selected five judgment documents from our collected dataset for the pilot annotation study. We asked four experts in law to annotate the documents. The annotators are one lawyer and three professors of law. The lawyer and two of the professors are the authors of this paper. We measured the reliability of annotation with Krippendorff's α_U (Krippendorff, 1995), which was designed for unitising annotation tasks. We used an implementation provided by Meyer et al. (2014). In the rationale extraction task, we obtained $\alpha_U = 0.407$. This result indicates much lower reliability of the annotation than we expected.

To identify the source of disagreement, we manually checked the annotation and interviewed the annotators. Our findings are the followings: First, instructions to directly extract rationales cause disagreements. Although we instructed the annotators to extract only the accepted argument as rationale, some annotators mistakenly extract both accepted and rejected arguments. This is because a judgment describes accepted and rejected arguments from different parties, and both types of arguments can be relevant to decisions on torts. In the later version of our scheme (6.2.2), to make the annotation task clear, we split annotation on rationales into two tasks: extracting relevant arguments, and checking if each of the arguments is accepted or not. All extracted spans are just relevant arguments and facts. Of the spans, those annotated as accepted are the rationales. Second, rationales consist of multiple types of content: not only factual findings but also abstract norms such as tests established by the precedent. As the pilot scheme mainly considered factual allegations as rationales, the annotators got confused in annotating such abstractive arguments. In later versions of our scheme, we split rationale spans into several types (Factual Claims, Claims of Norms, and Major Claims). Third, a scheme has to express which rationale corresponds to which tortious act when there are multiple alleged actions as torts. We implement this as the task of span association (6.3).

6. Our Scheme

Given the pilot annotation study results, we have compiled our annotation scheme. The scheme consists of three stages: 1. Document screening, 2. Span extraction, and 3. Span association. This section summarises our annotation scheme and its guideline, which are written in Japanese.

6.1. Document Screening

In this paper, we focus on annotating legal argumentation on torts. We collected documents from the legal database with queries excluding documents not concerning torts. However, some irrelevant documents can still be in our documents set, and we have to filter them out manually. The document screening is a simple task to filter out such non-tort judgments. We ask annotators to read through the judicial decision part in a judgment and check if the court considers an issue of torts and makes any decisions on it. If annotators find a judgment has nothing to do with torts, they are instructed to flag the judgment and stop annotating it.

6.2. Span Extraction

Once annotators confirm that the judgment contains legal discussions on torts, they are asked to identify the text span describing the court's conclusion on torts from the judicial decision part. And then, they also extract rationales from the parts of the parties' claims and facts.

This stage consists of two tasks, identifying the court's conclusions and rationales as text spans and assigning attributes to them. There are five different types of spans for rationales and the court's conclusion.

6.2.1. Spans
We instructed annotators to extract spans according to the following definitions. The span length ranges from a single character to a single sentence. Annotators might identify no span for a type if there is no corresponding text in a document.

Court Decisions (CD): This type of span describes a judges' decision on tort. Annotators must find Court Decisions spans from the part of the judicial decision. One span identifies one tort. We ask annotators to extract the most concrete and finest-grained description if multiple texts refer to the same subject in a document.

Factual Claims (FC): This span describes important claims from parties, which are relevant to judgment on torts. Annotators must find Factual Claims spans from the parts other than the judicial decision part. The Factual Claims contain factual allegations, an assertion of opposing facts against them, and rebuttals against one's factual allegations.

Claims of Norms (NC): Claims of Norms describes abstract legal arguments regarding torts. Annotators must find Claims of Norms spans from the parts other than the judicial decision part. This type of span often consists of references to past precedents, in particular the Japanese Supreme Court judgments.

Major Claims (MC) *Removed in the final version: Major Claim spans describes important major claims from parties, which summarise and conclude based on Factual Claims. Annotators must find Major Claims spans from the parts other than the judicial decision part. The Major Claim spans are often found in the last sentence of a series of Factual Claims.

Undisputed Facts (UF): Undisputed Facts spans describes facts that play important roles in judging torts. The facts covered with these spans are undisputed by any parties. The annotators find the spans from the parts other than the judicial decision part in principle. However, annotators are allowed to annotate text describing facts in the judicial decision part only if they are indispensable in the legal reasoning on torts and if they are described as "undisputed" or "easily recognised from evidences." Undisputed Fact often identifies the subject of the plaintiff's original allegation.

6.2.2. Attributes

@Accepted Claim (@AC): Spans of FC, NC and MC have this attribute. When annotators identify these types of spans from the part of parties' claims, they are asked to check if the claim is accepted by judges in the court decision part or not. Annotators can choose either *True* or *False*.

@Who (@W): Spans of FC, NC and MC have this attribute. Parties to legal proceedings, often plaintiffs and defendants, submit their claims to the court. Annotators are instructed to identify whose claim it is. Annotators choose one from *Plaintiff*, *Defendant*, *Other*.

@Decision (@D): Only spans of CD have this attribute. Annotators interpret the identified span and annotate if the torts are affirmed by judges or not. Annotators can choose either *True* or *False*.

6.3. Associating spans

In the court case trials, we can find more than one CD span, and we have to identify which CD span non CD spans are related to. In the span association task, annotators are instructed to associate all Factual Claims, Major Claims, Claims of Norms, and Undisputed Facts spans with their corresponding CD spans. Annotators can associate a span with multiple CD spans.
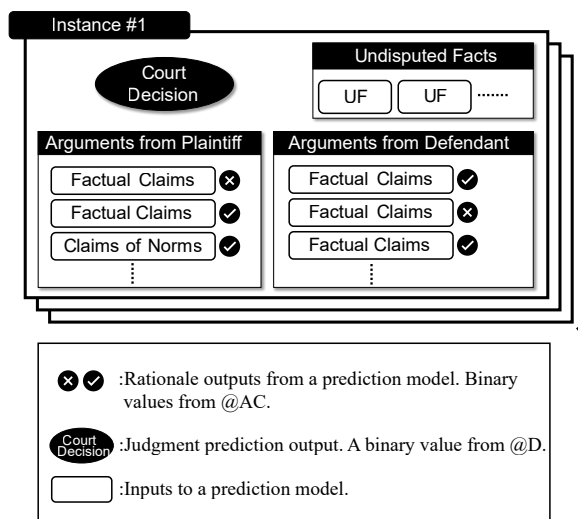

Figure 2: Product of annotation

6.4. Procedure

We instructed annotators to conduct the whole annotation process as follows. First, annotators read through a document and understand the flow of arguments. Next, annotators perform the document screening. If a document passes the screening, annotators start the annotation. In the span extraction, annotators extract CD spans and assign the @D attribute to them. Once all CD spans are extracted, annotators start to identify other types of spans that are rationales of the extracted CD spans. For each span, annotators assign necessary attributes. Finally, annotators associate every rationale span with their corresponding CD spans.

Figure 2 illustrates an annotation instance. Each instance holds a single decision on torts so that multiple instances can be produced from a document if there are multiple issues. Figure 2 introduces a simple setup for LJP. Inputs to a model are arguments from the parties and undisputed facts. The attributes of @AC and @D are the outputs of predictions for the judgment prediction and the rationale extraction, respectively.

7. Annotation Study

To check the reliability of our annotation scheme, we conducted three annotation studies. In the first annotation study (Study 1), we have employed six annotators, who are law school graduates and lawyers. The objective of Study 1 is to check the feasibility of our proposed annotation scheme and improve our annotation guidelines. The second annotation study (Study 2) involved five annotators. They are students in law who have not yet graduated from law school (most of them are undergraduates). Study 2 is designed to confirm that law students can perform our annotation. They have expert knowledge in law; however, they have less experiences in interpreting judgment documents than Study 1 annotators. As we plan to conduct our annotation at a large scale by employing many annotators,

	# of Spans (macro-avg. over rators)						misc		
	FC	NC	MC	UF	CD	Total	Rators	Docs	Avg. chars
Study 1	88.2	2.7	14.7	35.2	39.5	180.2	6	10	13507.9
Study 2	44.0	0	8.0	16.6	15.2	83.8	5	5	14890.6
Study 3	252.4	6.8	N/A	70.2	48.0	377.4	5	25	10763.6

Table 1: Statistics of annotation studies

including undergraduates, Study 2 annotation study is necessary. Table 1 summarises statistics of each annotation studies. In the third annotation study (Study 3), we confirm the reliability of our annotation scheme (identical to one presented in 6) with five annotators with various backgrounds, including three lawyers, one law school graduate and one undergraduate. This is the final step before we proceed to the production annotation.

Note that the scheme we describe in 6 is the final version after annotation studies were conducted and improvements were made. There were minor changes through the annotation studies. To the extent necessary, we identify those differences from the finalised version in the following sections.

7.1. Agreement Metrics

As our tasks except document screening are extracting different types of spans, in other words, "unitising" tasks, we decide to use Krippendorff's α_U (Krippendorff, 1995) as the main metric. As for the document screening, we use the agreement ratio and Fleiss's κ (Fleiss, 1971). We use implementations of the metrics provided by Meyer et al. (2014). Table 2 shows the result of IAA on document screening. If a document is marked as non-tort by more than half of the annotators, we exclude the document in calculating agreement in other tasks.

In span extraction and association, we calculate α_U with the offset of spans (i.e. positions of spans) and their "labels". What we use as "labels" in the calculation are different in tasks.

We use their span types (e.g. CD, FC) as labels for the spans. Table 3 shows the result of IAA on span extraction. Note that Major Claim is removed after Study 2 annotation and its IAA is not available in Study 3. The column of Overall α_U calculated over all span types, and it is not an average of a row. Other columns show α_U calculated only with its corresponding span type.

For the attributes, we regard values of attributes (e.g. True/False, Plaintiff/Defendant/Other.) as labels. We calculate the agreement metrics on each attribute type. Table 4 shows the result of IAA attributes.

In the associating spans task, we use associated CD spans for each span as labels to calculate α_U. Note that the spans are not pre-defined units like span types and attributes. Thus, CD spans are not guaranteed to be identical among annotators. To make a consistent set of CD spans, we merge CD spans from different annotators if they overlap.

	Agreement Ratio	Fleiss's Kappa
Study 1	0.83	-0.09
Study 2	1.00	N/A
Study 3	0.96	0.77

Table 2: IAA on document screening

7.2. Annotation Study 1

In Study 1, we asked six annotators to annotate ten judgment documents (Table 1). The agreement ratio of document screening is at 0.83, and Fleiss's Kappa was at -0.09 (Table 2). The reason for a high agreement ratio but a lower Fleiss's Kappa is an imbalance between non-tort and torts documents, i.e. the torts documents are dominant in the document set. Therefore, annotators rarely found non-tort documents. We did not exclude any documents in the IAA calculation according to the result of document screening.

The IAA of span extraction is at $\alpha_U = 0.427$ (Table 3). It indicates a moderate agreement, and there is much room for improvement. NC shows the lowest score when we look at IAA by categories. It is because NC spans are rare, and the annotators did not identify any spans of NC in some documents during the Study 1 annotation. If no one identifies any NC spans in a document, NC's α_U of the document falls to zero. If we define that α_U of a document is 1 when all annotators identify no NC span in the document, α_U of NC becomes 0.576.

The primary source of disagreement is UF and MC. After the Group 1 annotation, we reviewed the definitions of UF and MC and found them still ambiguous. In the Study 1 version of the annotation guideline, there were no clear guidance to find a specific type of span. The judgment documents can be segmented into three parts[5]: a part presenting judge's evaluation and conclusions, a part of facts, and a part describing claims from the parties. For example, annotators can find similar text describing facts from both the part of the fact parts and the judge's part, and they are confused in extracting UF spans. As UF spans identify only undisputed facts, annotators primarily scan and find UF spans from the fact part or the judge's part but not from the parties'

---

[5]They can be often separated using a surface structure such as headings or paragraphs, but not guaranteed. In some cases, they are concatenated and intermingled in a paragraph or even in a sentence. Thus, we need legal experts for our annotation.

	Factual Claims	Claims of Norms	Major Claims	Undisputed Facts	Court Decisions	Overall
Study 1	0.543	0.076	0.361	0.189	0.644	0.427
Study 1 (5 docs)	0.552	-0.007	0.315	0.149	0.438	0.344
Study 2	0.549	0	0.197	0.337	0.457	0.498
Study 3	0.647	0.069	N/A	0.520	0.607	0.654

Table 3: IAA on span extraction (α_U)

Target Spans:	C, NC, MJ	C, NC, MJ	CD
Attribute types:	@Accepted Claim	@Who	@Decision
Study 1	0.521	0.526	0.629
Study 1 (5 docs)	0.563	0.587	0.428
Study 2	0.605	0.516	0.438
Study 3	0.629	0.641	0.608

Table 4: IAA on attributes of spans (α_U)

	α_U
Study 1	0.301
Study 1 (5 docs)	0.209
Study 2	0.321
Study 3	0.430

Table 5: IAA on span association (α_U)

claims part. Therefore, We add further explanation of each span type explaining where to find the spans in the guideline for Study 2 annotation.

As for IAA of the attributes (Table 4), the @D attribute showed the reasonable agreement at $\alpha_U = 0.629$. The @D attribute is assigned to every CD span, and the offset of each span is identical to that of CD spans. Also, choosing the suitable attribute is considered easy since the annotators read the text of a span and check whether judges affirm the existence of torts. Once the annotators could agree to extract the same CD spans, @D could be agreed easily. Thus, the IAA score of @D should be close to that of CD spans unless the annotators disagree in the attribute assignment itself. The @W attribute showed $\alpha_U = 0.526$ and @AC attribute was at $\alpha_U = 0.521$. They are lower than @D but still moderate agreement. Similarly to @D, their annotations are dependent on their assigned spans (FC, NC, MJ). As FC is the dominant type of span, the IAA scores of @W and @AC tend to be close to the IAA score of FC. To sum up, the reliability of attributes itself does not have major flaws. Further improvement should be achieved when we improve the span extraction reliability.

The IAA scores for associating spans are at $\alpha_U = 0.301$ (Table 5). The score shows a lower value than those of other tasks. The span association task requires that CD spans has been correctly annotated. Thus, disagreements on CD spans impact rationale association. The lower α_U of the rationale association reflects its dependency on the CD span extraction. Moreover, the disagreements of the CD span extraction are amplified in associating spans. For example, suppose one of the annotators extracts an extra CD span that other annotators do not extract in a document, all the spans associated with the extra CD span cause association disagreement. In contrast, only the extra CD span is penalised in the span extraction. Therefore, we expect α_U of the association task become lower than the span extraction.

There is much room to be improved for better relia-

bility. As we discussed above, improving the reliability of annotation on CD spans should contribute to the reliability of span association. When we review the disagreement of CD spans, we observe annotators can agree on the content of CD spans but disagree with where they extract it. There can be multiple candidates for a CD span. They describe the almost same content but in different levels of abstraction. In a case where the plaintiff claimed that defendant committed torts, for example, "Defendant A's action X cannot be considered as torts", "Plaintiff's allegations are not acceptable.", and "Reject", all of these suggest judges did not find any torts. To avoid confusion, we added a guide to extract CD spans in Study 2 annotation: annotators are instructed to choose the most concrete description as CD span if multiple candidates indicate the same content (torts).

The annotators gave us feedback that fully annotated sample documents are necessary before starting annotation. In the Study 1 annotation, we only provided the annotation guideline and a few examples for each span type, but we did not provide fully annotated documents. The annotated sample documents will provide clear and instant clues to annotators. We provide the sample documents from Study 2 annotation. Another remarkable observation was that some annotators accidentally forgot to complete the document screening and the span association. We develop an automated tool to detect the missing annotation and report the status of each document online so that annotators can review their annotation by themselves. This tool was provided from Study 2 annotation.

7.3. Annotation Study 2

In Study 2, we employed five students, who had not yet graduated from law school, to annotate five documents (Table 1). None of them participated in Study 1 annotation. Considering the feedback from Study 1, we provide the annotators with fully annotated sample documents together with the guidelines. The author of the papers annotated five documents to prepare the sample

documents. Furthermore, we compiled step-by-step tutorials using the samples. In the tutorial, each annotator is first asked to annotate the documents without consulting the provided sample annotation. They can only look up the sample annotation when they are unsure of their annotation. The sample documents are given in the order of difficulties so that annotators can learn the annotation process step by step from principles to their advanced applications. The difficulties are determined according to the consensus among authors of this paper, considering levels of complication in legal reasoning, the number of tortious acts claimed, and the length of the document.

The IAA of document screening is at 1.0 according to the agreement ratio. All of the annotators agreed that all five documents deal with torts. Row of Study 2 on Table 3 shows IAA scores of span extraction. As the five documents used in Study 2 are also used in Study 1, we show IAA scores of Study 1 calculated only with the same five documents as Study 2 in "Study 1 (5 docs)" on the Table. They provide a clear comparison between Study 1 and 2 and suggest if the improvements on the guidelines and the tutorials work as expected. The IAA of span extraction is now $\alpha_U = 0.498$ (Overall) improved from 0.344 in Study 1. Undisputed Facts is remarkably improved among the span types, suggesting the revised guideline works as intended. On the other hand, Major Claim became worse in Study 2 despite the revision. Major Claims are introduced initially to identify text concluding and summarising multiple FC spans. They often contain summaries of actual claims and arguments, which are already identified by FC spans. This nature of Major Claims makes it hard to distinguish FC from Major Claims. We removed Major Claim from our scheme after Study 2 annotation since FC spans should be sufficient to provide rationales. The score of NC is 0 since all annotators agreed that there is no NC span in the documents. The reliability of attributes stays at a reasonable level as shown in Table 4. As for span association, $\alpha_U = 0.321$ is much improved from Study 1 (0.209). We can attribute this improvement to the improvement in CD spans extraction (from 0.438 to 0.457).

Even though annotators of Study 2 have less experience in interpreting judgment documents than those of Study 1, scores of IAA show reasonable reliability and are even better than Study 1. This encouraging result shows that our improved guidelines and tutorials using the annotation samples effectively train annotators.

7.4. Annotation Study 3

We improved our scheme through the two iterations of annotation studies. In addition to the changes we described above, we elaborated on what to extract for each span type for better agreement in the span extraction. In the Study 3, we assess the reliability of our final annotation scheme, including the guidelines, tutorials with samples. We ask five annotators to annotate

25 documents (Table 1). The 25 documents have no overlap with the documents used in Study 1 and 2.

We observed good agreement overall in Study 3. The agreement ratio of document screening is at 0.96, and Fleiss's $\kappa = 0.77$. They indicate stable annotation for this task. The IAA of span extraction finally achieves $\alpha_U = 0.654$ (overall). Every span type shows better α_U from Study 2 annotation. The α_U of attributes are at 0.629(@AC), 0.641(@W), 0.608(@D) showing improvement from Study 2 annotation. Span association α_U is now at 0.430 improved from Study 2 annotation. The IAA score of span association is still lower than that of span extraction. The task delivers errors from span extraction so that α_U get penalised from both the association task itself and the span extraction task as we discuss in 7.2. Although α_U of the span association was not successful as other tasks, numerical improvement of α_U through three annotation studies suggests our scheme revision has worked as intended.

8. Conclusion and Future work

Our three iterative annotation studies achieved good agreement, particularly for the span extraction and the attributes task, suggesting that our annotation scheme and training materials, including tutorials with the annotation samples, were successful. On the other hand, the span association agreement should be further improved. We will continue to improve the agreement of the span association by revising our guidelines.

In this study, we tested our scheme only on a certain type of torts cases. Although our annotation scheme is designed for general torts cases, it may require minor revision for different types of torts.

The next step of our project is deploying our annotation scheme to more legal experts and annotate judgments on torts at a larger scale. To produce a dataset capable of training and evaluating ML-based models of LJP, we aim to construct the dataset with 5,000 documents. In the production phase of annotation, we plan to provide tools to maintain the quality of annotation in addition to the tutorials and the guidelines. In the annotation studies, we prohibited the annotators from communicating with each other. In the production phase, however, online chat tools will provide a forum to exchange questions and ideas among annotators, which leads to more consistent and better annotation results. These tools should help annotators keep their annotation reliable and legitimate.

9. Bibliographical References

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., and Lampos, V. (2016). Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Comput. Sci.*, 2:e93, October.

Branting, L. K., Pfeifer, C., Brown, B., Ferro, L., Aberdeen, J., Weiss, B., Pfaff, M., and Liao, B. (2021). Scalable and explainable legal prediction. *Artificial Intelligence and Law*, 29(2):213–238, June.

Cejuela, J. M., McQuilton, P., Ponting, L., Marygold, S. J., Stefancsik, R., Millburn, G. H., Rost, B., and FlyBase Consortium. (2014). tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database*, 2014(0):bau033, April.

Chalkidis, I., Androutsopoulos, I., and Aletras, N. (2019). Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chalkidis, I., Fergadiotis, M., Tsarapatsanis, D., Aletras, N., Androutsopoulos, I., and Malakasiotis, P. (2021a). Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online, June. Association for Computational Linguistics.

Chalkidis, I., Jana, A., Hartung, D., II, M. J. B., Androutsopoulos, I., Katz, D. M., and Aletras, N. (2021b). LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. *CoRR*, abs/2110.00976.

Ferro, L., Aberdeen, J., Branting, K., Pfeifer, C., Yeh, A., and Chakraborty, A. (2019). Scalable methods for annotating Legal-Decision corpora. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 12–20, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76(5):378–382, November.

Hu, Z., Li, X., Tu, C., Liu, Z., and Sun, M. (2018). Few-Shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Jacovi, A. and Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Compu-*

tational Linguistics, pages 4198–4205, Online, July. Association for Computational Linguistics.

Katz, D. M., Bommarito, 2nd, M. J., and Blackman, J. (2017). A general approach for predicting the behavior of the supreme court of the united states. *PLoS One*, 12(4):e0174698, April.

Kozuka, S. (2020). The Style and Role of Judgments by Japanese Courts. *Zeitschrift für Japanisches Recht*, 25(49):47–75, June.

Krippendorff, K. (1995). On the reliability of unitizing continuous data. *Sociol. Methodol.*, 25:47–76.

Leins, K., Lau, J. H., and Baldwin, T. (2020). Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online, July. Association for Computational Linguistics.

Long, S., Tu, C., Liu, Z., and Sun, M. (2019). Automatic judgment prediction via legal reading comprehension. In *Chinese Computational Linguistics*, pages 558–572. Springer International Publishing.

Luo, B., Feng, Y., Xu, J., Zhang, X., and Zhao, D. (2017). Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736, Copenhagen, Denmark, September. Association for Computational Linguistics.

Medvedeva, M., Vols, M., and Wieling, M. (2018). Judicial decisions of the European Court of Human Rights: looking into the crystal ball. In *Proceedings of the Conference on Empirical Legal Studies in Europe 2018*.

Meyer, C. M., Mieskes, M., Stab, C., and Gurevych, I. (2014). DKPro agreement: An Open-Source java library for measuring Inter-Rater agreement. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 105–109, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Navas-Loro, M., Satoh, K., and Rodríguez-Doncel, V. (2019). ContractFrames: Bridging the gap between natural language and logics in contract law. In *New Frontiers in Artificial Intelligence*, pages 101–114. Springer International Publishing.

Nitta, K., Ohtake, Y., Maeda, S., Ono, M., Ohsaki, H., and Sakane, K. (1993a). HELIC-II: legal reasoning system on the parallel inference machine. *New Gen. Comput.*, 11(3-4):423–448, July.

Nitta, K., Wong, S., and Ohtake, Y. (1993b). A computational model for trial reasoning. In *Proceedings of the 4th international conference on Artificial intelligence and law*, ICAIL '93, pages 20–29, New York, NY, USA, August. Association for Computing Machinery.

Rabelo, J., Kim, M., Goebel, R., Yoshioka, M., Kano,

787

Y., and Satoh, K. (2020). COLIEE 2020: Methods for legal document retrieval and entailment. In Naoaki Okazaki, et al., editors, *New Frontiers in Artificial Intelligence - JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15-17, 2020, Revised Selected Papers*, volume 12758 of *Lecture Notes in Computer Science*, pages 196–210. Springer.

Satoh, K., Asai, K., Kogawa, T., Kubota, M., Nakamura, M., Nishigai, Y., Shirakawa, K., and Takano, C. (2011). PROLEG: An Implementation of the Presupposed Ultimate Fact Theory of Japanese Civil Code by PROLOG Technology. In *New Frontiers in Artificial Intelligence*, pages 153–164. Springer Berlin Heidelberg.

Sumida, M. and Steffek, F. (2022). What is Legal Innovation. In Mihoko Sumida et al., editors, *Rîgaru Inobêsyon Nyûmon (Introduction to Legal Innovation) (Japanese)*, page 393. KOUBUNDOU Publishers Inc.

Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Feng, Y., Han, X., Hu, Z., Wang, H., and Xu, J. (2018). CAIL2018: A Large-Scale legal dataset for judgment prediction. *arXiv*, July.

Xu, N., Wang, P., Chen, L., Pan, L., Wang, X., and Zhao, J. (2020). Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3086–3095, Online, July. Association for Computational Linguistics.

Yamamoto, K. (2019). *Basic features of Japanese tort law*. Civil law. Jan Sramek Verlag KG.

Ye, H., Jiang, X., Luo, Z., and Chao, W. (2018). Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana, June. Association for Computational Linguistics.

Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., and Sun, M. (2018). Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.

Type	Text (English versions are our translation.)
UF	被告Ｐ２は，インターネット上で発見した黒猫のイラストをダウンロードし，同イラストの頭部を切り取り，同頭部にフラダンスの衣装等を組み合わせて，被告イラスト１を作成した。
	Defendant P2 downloaded an illustration of a black cat that he found on the Internet, cut out the head part of the illustration, and made defendant's Illustration 1 by combining the head part with a hula dance outfit and other items.
FC	送信可能化されたファイルが本件各レコードの複製物であるかは客観的な証拠がなく，不知である。
	We do not know whether the files made available for transmission are copies of the records in question since there is no objective evidence.
NC	名誉感情に対する侵害を理由に不法行為が成立するのは，社会通念上許容される限度を超える侮辱行為が認められる場合に限られる（最高裁平成２２年４月１３日判決）
	A tort because of infringement of honour is established only when an insult exceeds the socially accepted limit (Supreme Court decision of April 13, 2010).
CD	本件記事１については，名誉毀損又は信用棄損による不法行為は成立しないというべきである
	Concerning the article-1, no tort for defamation or damage to reputation is established.

Table 6: Examples for each span type

Appendix: Annotation Examples

Table 6 provides examples for each span type. The examples are from different judgments.

Figure 3 shows examples of our annotation. This case involves a request to disclose the sender's information, alleging that a posting on a bulletin board system on the Internet has lowered the plaintiff's social reputation and defamed the plaintiff. Table 7 lists the corresponding annotation artifacts based on Figure 3. In the table, *Type* column indicates span types. *@W*, *@D* and *@AC* mean @Who, @Decision, and @Accepted Claims, respectively. The last column, *Assoc.* shows IDs of associated CD spans for each span. In the examples, span 1 is Undisputed Facts (UF). Spans 2, 3, 4, 5 and 6 are Factual Claims (FC) from the plaintiff. Span 7 is also an FC but from the defendant. According to the judicial decisions, FC spans of the plaintiff 2 and 3 are accepted while FC spans 4, 5 and 6 are not. The defendant's FC span 7 is accepted. All spans from 1 to 7 are associated with span 8, which is a CD span. Note that this example is one of the simplest judgments. There can be more than one CD span and much more spans from both the plaintiff and defendant in longer judgments.

Annotations on "Span associating"
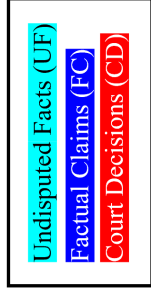
第1 請求

被告は、原告に対し、別紙発信者情報目録記載の情報を開示せよ。

第2 事案の概要

本件は、原告が、インターネット上の電子掲示板に投稿された別紙投稿目録記載の書込みにより社会的評価を低下させられた名誉権を侵害されたと主張して、特定電気通信役務提供者の損害賠償責任の制限及び発信者情報の開示を求める法律（以下「法」という。）4条1項に基づき、別紙発信者情報目録記載の情報の開示を求める事案である。

1 令和元年7月12日午後11時46分17秒、インターネット上の電子掲示板「C」中の「D」「E」に作成された「F」という標題のスレッド（以下「本件スレッド」という。）に、「X1さんは金返さないと」という書込み（別紙投稿目録記載の⑦。以下「本件投稿」という。）が、IPアドレス「○○○.○○○.○○○.○○」を経由して投稿された。

2 争点及び当事者の主張

本件の争点は、「本件投稿について権利侵害の明白性が認められるか。」であり、これに関する当事者の主張は以下のとおりである。

（1）原告の主張

ア 同定可能性

本件投稿は、一般の閲覧者の普通の注意と閲覧の仕方を基準とすると、B製作所に勤務する「X1」という人物が、特定の個人から金銭を借りた上でその返済をしていないという事実を摘示するものである。

B製作所に勤務する「X1」という姓の人物は、原告とそのいとこの2名のみである。

本件投稿の閲覧者のうち、原告を知っているが原告のいとこを知らない者は本件投稿の対象が原告と原告のいとことの間に固定可能性はある。

したがって、本件投稿の対象が「X1」という記載から原告との間に固定可能とことを争う。

（2）被告の主張

上記（1）原告の主張を、いずれも争う。

第3 当裁判所の判断

1 争点（権利侵害の明白性）について

（1）同定可能性について

前記前提事実、証拠（甲4、5）及び弁論の全趣旨によれば、本件投稿が、本件スレッド（インターネットサイト「C」中の「D」「E」上に作成された「F」と題するスレッド）に投稿された「X1さんは金返さないと」という内容のものであること、原告が岡山県倉敷市所在のB製作所に勤務する会社員であること、B製作所に勤務する「X1」という姓の人物は原告とそのいとこの2名であることが認められる。

これらの事実を踏まえると、B製作所に「X1」という姓の人物が原告のほかにもう1名いることを知っている閲覧者は、本件投稿の「X1さんが原告とそのいとこのいずれを指すのかを判断できず、また、B製作所に「X1」という姓の人が何名いるのかを知らない閲覧者は、やはり本件投稿の「X1さんが原告を指すものとは判断できないことになるのであって、一般の閲覧者の普通の注意と閲覧の仕方を基準にすると、結局、一般の閲覧者にとって本件投稿の「X1さんが原告を指すものであるとはいえないといわざるを得ない（なお、B製作所に勤務する「X1」という姓の人物が原告のみであると誤信している閲覧者が本件投稿の「X1さんが原告を指すものであるとそのような誤信を指すことはあり得るが、このような判断は誤信に基づくものであり、一般の閲覧者の普通の注意と閲覧の仕方を基準とする一般の閲覧者の普通の注意と閲覧の仕方を基準とした場合に本件投稿の「X1さんが原告を指すものであるということはできない）。

このように、本件投稿の対象が原告であるとはいえないことから、本件投稿が原告の社会的評価を低下させて原告の名誉を毀損するものである**ということはできない。**

本件投稿は、一般の閲覧者の普通の注意と閲覧の仕方を基準とすると、原告の社会的評価を低下させて原告の名誉を毀損するものであるということはできない。

B製作所に勤務する「X1」という姓の人物は、原告とそのいとこの2名のみである。

本件投稿の閲覧者のうち、原告を知っているが原告のいとこを知らない者は本件投稿の対象が原告と考えるであろう。

本件投稿の対象が原告であるとはいえないことから、本件投稿が原告の社会的評価を低下させて原告の名誉を毀損するものであるということはできない。

Figure 3: Simplified annotation samples

ID	Text (English text is our translation.)	Type	@W	@D	@AC	Assoc.
1	令和元年７月１２日午後１１時４６分１７秒，インターネット上の電子掲示板「Ｃ」中の「Ｄ」「Ｅ」に作成された「Ｆ」という標題のスレッド（以下「本件スレッド」という。）に，「Ｘ１さん金返さないと」という書込み（別紙投稿記事目録記載のもの。以下「本件投稿」という。）が，ＩＰアドレス「○○○．○○○．○○○．○○」を経由して投稿された。 At 11:46:17 PM, July 12, 2019, a posting "Mr X1, you should pay back the money" (the attached list of submitted articles) was made in the thread titled "F", which was created in "D" and "E" on the bulletin board system "C" on the Internet, via IP address ***.***.***.***.	UF	N/A	N/A	N/A	8
2	本件投稿は，一般の閲覧者の普通の注意と閲覧の仕方を基準とすると，Ｂ製作所に勤務する「Ｘ１」という人物が，特定の個人から金銭を借り入れたがその返済をしていないとの事実を摘示するものである。 This posting, based on a viewer of ordinary prudence and his way of viewing , indicate the fact that a person named "X1," who works at factory B, borrowed money from a certain individual but has not repaid it.	FC	Pl.	N/A	T	8
3	Ｂ製作所に勤務する「Ｘ１」という姓の人物は，原告とそのいとこの２名のみである。 There are only two persons with the surname "X1" who work at factory B: the plaintiff and his cousin.	FC	Pl.	N/A	T	8
4	本件投稿の閲覧者のうち，原告を知っているが原告のいとこを知らない者は本件投稿の対象が原告と考えるであろう The viewers of this posting, who know the plaintiff but do not know the plaintiff's cousin, would regard the plaintiff as the subject of the posting.	FC	Pl.	N/A	F	8
5	原告と原告のいとこを知る者が「Ｘ１」という記載から原告のことを思い浮かべることもあるはずである。 Some of people, who knows the plaintiff and the plaintiff's cousin, can recall the plaintiff from the mention of "X1".	FC	Pl.	N/A	F	8
6	本件投稿の対象と原告との間に同定可能性はある。 It is possible to identify the subject of this posting as the plaintiff.	FC	Pl.	N/A	F	8
7	上記（１）原告の主張は，いずれも争う。 We do not admit all of the above (1) allegations of the plaintiff.	FC	Def.	N/A	T	8
8	本件投稿の対象が原告であるとはいえないことから，本件投稿が原告の社会的評価を低下させて原告の名誉を毀損するものであるということはできない。 Given that we cannot find that the subject of this posting is really the plaintiff, we cannot recognize that the posting is defamatory to the plaintiff by diminishing the plaintiff's social reputation.	CD	N/A	F	N/A	N/A

Table 7: Annotation artifacts based on Figure 3