Annotation of argument structure in Japanese legal documents

Hiroaki Yamada[†] **Simone Teufel**[‡ †] **Takenobu Tokunaga**[†]
[†]Tokyo Institute of Technology [‡]University of Cambridge
yamada.h.ax@m.titech.ac.jp simone.teufel@cl.cam.ac.uk take@c.titech.ac.jp

Abstract

We propose a method for the annotation of Japanese civil judgment documents, with the purpose of creating flexible summaries of these. The first step, described in the current paper, concerns content selection, i.e., the question of which material should be extracted initially for the summary. In particular, we utilize the hierarchical argument structure of the judgment documents. Our main contributions are a) the design of an annotation scheme that stresses the connection between legal issues (called *issue topics*) and argument structure, b) an adaptation of rhetorical status to suit the Japanese legal system and c) the definition of a linked argument structure based on legal sub-arguments. In this paper, we report agreement between two annotators on several aspects of the overall task.

1 Introduction

Automatic summarization has become increasingly crucial for dealing with the information overload in many aspects of life. This is no different in the legal arena, where lawyers and other legal practitioners are in danger of being overwhelmed by too many documents that are relevant to their specialized task. The situation is aggravated by the length and complexity of legal documents: for instance, the average sentence length in Japanese legal documents is 93.1 characters, as opposed to 47.5 characters in Japanese news text. One reason for the long sentences is the requirement on judges to define their statement precisely and strictly, which they do by adding additional restrictive clauses to sentences. As a result, it is not possible to read every document returned by a search engine. The final goal of our work is therefore to provide automatic summaries that would enable the legal professions to make fast decisions about which documents they should read in detail.

To this end, we propose an annotation scheme of legal documents based on a combination of two ideas. The first of these ideas is the observation by (Hachey and Grover, 2006) that in the legal domain, content selection of satisfactory quality can be achieved using the argumentative zoning method. The second idea is novel to our work and concerns the connection between legal argumentation and summarization. We propose to identify and annotate each relevant legal issue (called *Issue Topic*), and to provide a linked argumentation structure of sub-arguments related to each Issue Topic separately. This can provide summaries of different levels of granularity, as well as summaries of each Issue Topic in isolation. In the current paper, we describe all aspects of the annotation scheme, including an annotation study between two expert annotators.

2 The structure of judgment documents

Legal texts such as judgment documents have unique characteristics regarding their structure and contents. Japanese judgment documents are written by professional judges, who, after passing the bar examination, are intensively trained to write such judgments. This results in documents with certain unique characteristics, which are reflected in the annotation scheme proposed in this paper. The document type we consider here, the judgment document, is one of the most important types of legal text in the Japanese legal system. Judgment documents provide valuable information for the legal professions to construct or analyze their cases. They are the direct output of court trials. The Japanese Code of Civil Procedure demands that "the court renders its

Proceedings of the 4th Workshop on Argument Mining, pages 22–31

judgment based on the original judgment document" (Japanese Ministry of Justice, 2012a, Article 252). The types of documents we work with in particular are Japanese Civil (as opposed to Criminal) case judgment documents from courts of first instance.

There also exist official human summaries of judgment documents, which we can use to inform our summary design, although they were issued only for a small number of documents.

2.1 Argument structure

The legal system that is in force in a particular country strongly affects the type and structure of legal documents used, which in turn has repercussions for summarization strategies. The first source of information for our summary design is a guideline document for writing judgment documents of civil cases (Judicial Research and Training Institute of Japan, 2006). In 1990, the "new format" was proposed, based on the principle that issue-focused judgment should make the document clearer, more informative and thus more reader-friendly (The Secretariat of Supreme Court of Japan, 1990). Although both the use of the guidelines and of the "new format" is voluntary, in practice we observed a high degree of compliance with the new format of the guidelines in recent Japanese judgment documents. This is for instance evidenced in the common textual structure shared amongst the documents. The "Fact and Reasons" section takes up the biggest portion of the document and is therefore the target of our summarization task. "Facts and Reasons" consists of a claim (typically brought forward by the plaintiff), followed by a description of the case, the facts agreed among the interested parties in advance, the issues to be contested during the trial, and statements from both plaintiff and defendant. The final part is the judicial decision. The entire structure described above is often marked typographically and structurally, e.g. in headlines.

Our second source of information concerns the argument structure of the legal argument. A Japanese Civil Case judgment document forms one big argument, depicted in Fig. 1. This argument structure includes the plaintiff's statements, the defendant's statements, and the judges' statement supporting their arguments. At the top of the

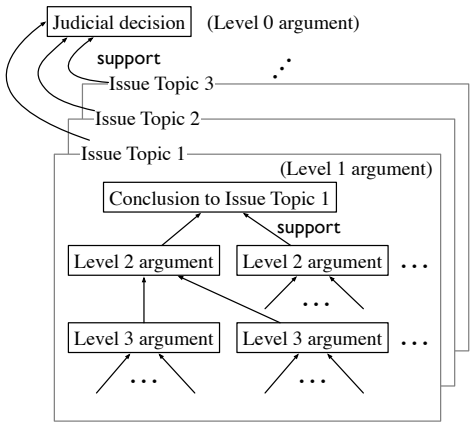

Figure 1: Argument structure of judgment document

structure, there is the root of the argument which states the judicial decision as the final conclusion to the plaintiff's accusation. We call this the level 0 argument.

The level 0 argument breaks down into several sub-arguments, which usually each cover one important topic to be debated. We call this the level 1 argument. Each level 1 argument might itself consist of sub-arguments at lower levels (levels 2, 3, 4, ...). The relation between levels is of type "support". In this kind of arguments, there are also "attack" relations. These occur, for instance, when a plaintiff argues *in favor of the negated* claim of the defendant, and vice versa. However, because these "attack" relationships follow the logic of the legal argumentation in a regular and systematic way, we decided not to explicitly model them in order to avoid redundancy in our annotation.

Our annotation scheme models this fact by calling level 2 units "FRAMING-main", units at level 3 or lower "FRAMING-sub", and by providing "support" links in the form of FRAMING linking between them. At the bottom of the argument structure, facts provide the lowest level of support, although in this work we do not model argumentation below level 3.

2.2 Issue Topics

The main organizing principle in the structure of the judgment document are the topics of each of the argumentation strands. This structure is a direct outcome of the Japanese judicial system, where most civil cases start with "preparatory proceedings". The goal of this procedure, which is carried out ahead of the trial under participation of all parties, is to define the issues to be tried

(Japanese Ministry of Justice, 2012b). These are called *Issue Topics*. Issue Topics are the main contentious items to be argued about between the interested parties. What could be a possible Issue Topics depends on the case alone; the number of issue topics is therefore unlimited. Examples include "whether the defendant X was negligent in supervising", "the defendant Y's exact actions during the crucial time frame" or "what is the effect of the law Z".

It is our working hypothesis that Issue Topics (which correspond to level 1 topics in our parlance) are extremely important in generating meaningful, coherent and useful summaries. Most legal cases consist of several Issue Topics; note that each Issue Topic is covered by its own argument subtree as depicted in Figure 1. In the best summaries the logical flow is organized in such a way that the final judicial decision can be traced back through each Issue Topic's connections. Minimally, this requires recognizing which sentence refers to which Issue Topic, i.e., linking each Issue Topic with the textual material supporting it. In what followes, we will call this task "Issue Topic linking".

2.3 Rhetorical structure

To exploit the idea that argument structure is a crucial aspect for legal summarization, we take a rhetorical status based approach. This method was originally defined for the summarization of scientific articles by Teufel and Moens (2002), but later studies such as Hachey and Grover (2006) applied the rhetorical status approach to the legal text domain for the context of English law.

Hachey and Grover defined the following seven labels: In English law, the judgment first states the facts and events, corresponding to category "FACT". "PROCEEDINGS" labels sentences which restate the details of previous judgments in lower courts. "BACKGROUND" is the label for quotations or citations of law materials which Law Lords use to discuss precedents and legislation. "FRAMING" is a rhetorical role that captures all aspects of the Law Lord's argumentation for their judgment. "DISPOSAL" is the actual decision of the lord which indicates the agreement or disagreement with a previous ruling[1]. "TEXTUAL" is used in situations where the sentence describes the structure of the document, rather than discussing content related to the case.

Hachey and Grover reported an inter-annotator agreement of K=0.83 (N=1955, n=7, k=2; where K is the kappa coefficient, N is the number of sentences, n is the number of categories and k is the number of annotators). Other researchers adopted the rhetorical status approach to their respective legal systems (Farzindar and Lapalme (2004) for the Canadian law and (Saravanan and Ravindran, 2010) for the Indian law). Farzindar and Lapalme's agreement figures are not available, but Saravanan and Ravindran's inter-annotator agreement of K=0.84 (N=16000; n=7, k=2) is remarkably similar to that of Hachey and Grover.

Our approach follows these studies in also using rhetorical structure (which we adapt to the Japanese system), but we combine this analysis with the two other levels of annotation motivated earlier (FRAMING linking and Issue Topic linking). We therefore model argument structure at a more detailed level than the earlier studies, by also considering the lower level hierarchical structure of argumentative support.

Based on the Issue Topic structure given in Figure 1, we propose to build summaries of different granularities, which have been inspired by the human summaries. Figure 2 shows an ideal sample of a structure-aware summary, using material manually extracted from an actual judgment document (our translation). The sample focuses on a specific Issue Topic of the case, namely "whether the execution officer D was negligent or not".

While full Issue Topic and FRAMING linking annotation would allow very sophisticated summaries, our fall-back strategy is to create simpler summaries using only the final conclusion and the main supporting text such as judicial decisions. For these simple summaries, performing only automatic rhetorical status classification is sufficient.

3 Annotation scheme

As discussed above, we use rhetorical status classification, which is a standard procedure in legal text processing. We also introduce two types of linking after rhetorical structure determination. These two types of linking are very different in nature and therefore need to be treated separately. The first kind of linking, Issue Topic linking, establishes a link from every single rhetorical status segment to the Issue Topic which the segment

[1] Since Hachey and Grover's target documents are from the UK House of Lords, trials are always brought forward at Courts of Appeal.

The plaintiff insists that the court executing officer was negligent in that the officer didn't notice that a person had committed suicide in the real estate when he performed an investigation of the current condition of the real estate, and also insists that the execution court was negligent in that the court failed to prescribe the matter to be examined on the examination order. As a result, the plaintiff won a successful bid for the estate with a higher price than the actual value of the estate given that the plaintiff did not have the information that the property was stigmatized. The plaintiff claims compensation for damage and delay from the defendant.

[Issue Topic]: Whether the execution officer D was negligent or not.

The measures performed by the officer were those that are normally implemented for examination. From the circumstances which the execution officer D perceived, he could not have realized that the estate was stigmatized. The officer cannot be regarded as negligent in that negligence would imply a dereliction of duty of inspection, which, given that there were sufficient checks, did not happen.

Concerning the question whether the officer had the duty to check whether the estate was stigmatized, we can observe various matters – in actuality, the person who killed himself happened to be the owner of the estate and the legal representative of the Revolving Credit Mortgage concerned, the house then became vacant and was offered for auction, but we can also observe the following: other persons but the owner himself could have committed suicide in the estate, for instance friends and family; there was a long time frame during which the suicide could have happened; the neighbors might not have answered the officer's questions in a forthcoming manner, even if they were aware of the fact that the estate was stigmatized; there are several factors to affect the value of the estate beyond the fact that the estate was stigmatized, and it is not realistic neither from a time perspective nor an economic perspective to examine all such factors specifically; and the bidders in the auction were in a position to examine the estate personally as the location of the estate was known – taking these relevant matters into consideration, it is a justified statement that the officer didn't have the duty to check in a proactive manner whether the estate was stigmatized.

Therefore, the plaintiff's claim is unreasonable since it is hard to say that the officer was negligent.

[Issue Topic]: Whether the examination court was negligent or not.

The plaintiff's claim is unreasonable for the additional reason that it is hard to say that the examination court was negligent.

Given what has been said above, it is not necessary to judge the other points;the plaintiff's claim is unreasonable so the judgment returns to the main text.

Figure 2: Sample summary/text structure

concerns. In contrast, the second kind of linking, which aims to model the "support" relationship between level 2 and level 3 shown in Figure 1, is much more selective: it only applies to those text units between which a direct "support" relationship holds. We will now introduce the following four annotation tasks: *1. Issue Topic Identification* – Issue Topic spans are marked in text, and identifiers are given to each Issue Topic; *2. Rhetorical Status Classification* – each text unit is classified into a rhetorical status; *3. Issue Topic Linking* – all rhetorical units identified in the previous stage are linked to a Issue Topic; *4. FRAMING Linking* – for those textual units participating in argumentation links between level 2 and level 3 arguments involving FRAMING-main, FRAMING-sub and BACKGROUND, additional links denoting "argumentative support" are annotated.

Label	Description
IDENTIFYING	The text unit identifies a discussion topic.
CONCLUSION	The text unit clearly states the conclusion from argumentation or discussion.
FACT	The text unit describes a fact.
BACKGROUND	The text unit gives a direct quotation or reference to law materials (law or precedent) and applies them to the present case.
FRAMING-main	The text unit consists of argumentative material that directly support a CONCLUSION unit.
FRAMING-sub	The text unit consists of argumentative material that indirectly supports a CONCLUSION unit or that directly supports a FRAMING-main unit.
OTHER	The text unit does not satisfy any of the requirements above.

Table 1: Our Rhetorical Annotation Scheme for Japanese Legal judgment documents

Earlier studies (Hachey and Grover, 2006; Saravanan and Ravindran, 2010) chose the sentence as the annotation unit and defined exclusive labeling, i.e., only one category can be assigned to each annotation unit. We also use exclusive labeling, but our definition of the smallest annotation unit is a comma-separated text unit. In Japanese, such units typically correspond to linguistic clauses. This decision was necessitated by the complexity and length of the legal language we observe, where parts of a single long sentence can fulfill different rhetorical functions. While annotators are free to label each comma-separated unit in a sentence separately, they are of course also free to label the entire sentence if they wish.

3.1 Issue Topic Identification

Our annotators are instructed to indicate the spans of each Issue Topic in the text, and to assign a

unique identifier to each[2]. Annotators are asked to find the span that best ("in the most straightforward way") describes the Issue Topic. We expect this task to be relatively uncontroversial, because the documents are produced in such a way that it should be clear what the Issue Topics are (cf. our discussion in section 2.2).

3.2 Rhetorical Status Classification

Our annotation scheme (Table 1) is an adaptation of Hachey and Grover's scheme; we introduce six labels for rhetorical status and an "OTHER" label. Two of the labels are retained from Hachey and Grover: the submission of fact (FACT) and the citation of law materials (BACKGROUND). DISPOSAL is redefined as CONCLUSION, in order to capture the conclusion of each Issue Topic. IDENTIFYING is a label for text that states the discussion topic. TEXTUAL is dropped in our annotation scheme since this label is not relevant to our summary design.

Our main modification is to split Hachey and Grover's FRAMING category into FRAMING-main and FRAMING-sub, in order to express the hierarchical argumentation structure discussed in section 2.1. Apart from the fact that the split allows us to recover the argument structure, we found that without the split, the argumentative text under FRAMING would cover too much text in our documents. Since our objective is to generate summaries, having too much material in any extract-worthy rhetorical role is undesirable.

We also introduce a slightly unusual annotation procedure where annotators are asked to trace back the legal argument structure of the case. They first search for the general CONCLUSIONs of the case. They then find each Issue Topic's CONCLUSION; next they find the FRAMING-main which is supporting. Finally, they look for the FRAMING-sub elements that support the FRAMING-main. They will then express the "support" relationship as FRAMING links (as described in section 3.4). Therefore, the annotators simultaneously recover the argument structure while making decisions about the rhetorical status.

3.3 Issue Topic Linking

Issue Topic linking concerns the relation between each textual unit and its corresponding Issue Topic. Every unit is assigned to the single Issue Topic ID the annotators recognize it as most related to. But not all textual material concerns specific Issue Topics; some text pieces such as the introduction and the final conclusion are predominantly concerned with the overall trial and judicial decision. We define a special Issue Topic ID of value 0 to cover such cases.

3.4 FRAMING Linking

Units labeled with BACKGROUND and FRAMING-sub can be linked to FRAMING-main, if the annotator perceives that the BACKGROUND or FRAMING-sub material argumentatively supports the statements in FRAMING-main. The semantics of a link is that the origin (BACKGROUND and FRAMING-sub) supports the destination (FRAMING-main).

4 Agreement metrics

Due to the nature of the four annotation tasks we propose here, different agreement metrics are necessary. While rhetorical status classification can be evaluated by Cohen's Kappa (Cohen, 1960), all other tasks require specialized metrics.

4.1 Issue Topic Identification

We perform a comparison of the *textual* material annotated as Issue Topic rather than the exact *location* of the material in the texts, as we only care to know whether annotators agree about the semantic content of the Issue Topics. We count two spans as agreed if more than 60% of their characters agree. The reason why we introduce the 60% overlap rule is that annotators may differ slightly in how they annotate a source span even if the principal meaning is the same. This difference often concerns short and relatively meaningless adverbial modification and such at the beginning or end of meaningful units. We manually confirmed that no false positives occurred by setting this threshold.

However, as annotators may disagree on the *number* of Issue Topic they recognize in a text, we first calculate an agreement ratio for each annotator by equation (1) and average them to give the overall agreement metric as in equation (2), where $a_s(i)$ is the number of spans agreed between annotator i and others and $spans(i)$ is the number of spans annotated by annotator i:

$$agreement_{\text{ITI}}(i) = \frac{a_s(i)}{spans(i)}, \quad (1)$$

[2]We later normalize the identifiers for comparison across annotators.

$$agreement_{\text{ITI}} = \frac{\sum_i agreement_{\text{ITI}}(i)}{|AnnotatorSet|}, \quad (2)$$

$$where\ i \in AnnotatorSet.$$

4.2 Issue Topic Linking

In the case of Issue Topic linking, the destinations of each link are fixed, as each Issue Topic is uniquely identified by its ID. As far as the sources of the links are concerned, their numbers across annotators should be almost the same. They only differ if the annotators marked different units as "OTHER" during rhetorical status classification, as all units except those marked "OTHER" are linked to an Issue Topic by definition. We therefore report an average agreement ratio as in equation (4) over each annotator agreement given by equation (3), where $a_u(i)$ is the number of units agreed between annotator i and others and $units(i)$ is the number of units annotated by annotator i:

$$agreement_{\text{ITL}}(i) = \frac{a_u(i)}{units(i)}, \quad (3)$$

$$agreement_{\text{ITL}} = \frac{\sum_i agreement_{\text{ITL}}(i)}{|AnnotatorSet|}, \quad (4)$$

$$where\ i \in AnnotatorSet.$$

4.3 FRAMING Linking

FRAMING linking is the most difficult task in our scheme to evaluate. FRAMING links hold between either BACKGROUND or FRAMING text units as the source, and FRAMING-main text units as the destination. The FRAMING linking task therefore consists of three parts, the identification of source spans, the identification of destination spans, and the determination of the most appropriate destination for each source span (linking).

The destinations are not uniquely identified in terms of an ID (as was the case with Issue Topic linking), but are variable, as the annotators mark them explicitly in the text using a span.[3]

First, we measure how well human annotators can distinguish all categories that participate in FRAMING linking (CONCLUSION, FRAMING-main, FRAMING-sub, BACKGROUND and "anything else"). The degree to which this subdivision is successful can be expressed by Kappa. This number gives an upper bound of performance that limits all further FRAMING linking tasks.

We also measure agreement on *source* spans for FRAMING linking. We define agreement as "sharing more than 80% of the span in characters", and report the number of spans where this is so, over the entire number of spans, as $agreement_{\text{src}}$, defined in equation (5). We only count those spans that are labeled as FRAMING-sub or BACKGROUND and are linked to somewhere[4].

$$agreement_{\text{src}} = \frac{\#\ of\ agreed\ source\ spans\ with\ link}{\#\ of\ source\ spans\ with\ link} \quad (5)$$

Finally, we measure agreement on *destination* spans that are linked to from source spans. Destination agreement $agreement_{\text{fl}}$ is the number of agreed[5] source spans which also have agreed destination spans, over all agreed source spans:

$$agreement_{\text{fl}} = \frac{\#\ of\ agreed\ links}{\#\ of\ agreed\ source\ spans\ with\ link}. \quad (6)$$

4.4 Baselines for FRAMING Linking

We implemented strong baselines in order to interpret our results for FRAMING linking. All baseline models output the linking result for each annotator, given as input the respective other annotator's source spans that have a link, and their destination spans. There is no other well-defined way to give any system options to choose from, and without these options, the baseline is unable to operate at all. In our setup, the baseline models simply simulate the last linking step between pre-identified source and destination spans, by one plausible model (namely, the other annotator).

Our three baseline models are called Random, Popularity and Nearest. The Random model chooses one destination span for each source span randomly. The Popularity model operates random by observed distribution of destinations. The distribution is calculated using the other annotator's data. The Nearest model always chooses the closest following destination candidate span available. This is motivated by our observation that in legal arguments, the supporting material often precedes the conclusion, and is typically adjacent or physically close.

[3] Note that the *source* spans are always variable, both in FRAMING linking and in Issue Topic linking.

[4] Although logically all such spans should be linked to somewhere, we observed some cases where annotators mistakenly forgot to link.

[5] For the definition of agreement on destination spans, the 80% rule is applied again.

5 Annotation experiment

Given the intrinsic subjectivity of text-interpretative tasks, and the high complexity of this particular annotation task, achieving acceptable agreement on our annotation is crucial for being able to use the annotated material for the supervised machine learning experiments we are planning for automating the annotation. We therefore perform an experiment to measure the reproducibility of our scheme, consisting of the 4 tasks as described above.

5.1 Annotation procedure

Two annotators were used for the experiment, the first author of this paper (who has a bachelor of Laws degree in Japanese Law), and a PhD candidate in a graduate school of Japanese Law. It is necessary to use expert annotators, due to the special legal language used in the documents. Legal domain texts can be hard to understand for lay persons, because terms have technical meanings which differ from the meaning of the terms when used in everyday language. For example, the terms "悪意 (*aku-i*)" and "善意 (*zen-i*)" (which mean "maliciousness" and "benevolentness" respectively in everyday language), have a different meaning in a legal context, where "悪意" means knowing a fact, and "善意" means not knowing a fact.

The annotators used the GUI-based annotation tool Slate (Kaplan et al., 2011). We gave the annotators a guideline document of 8 pages detailing the procedure. In it, we instructed the annotators to read the target document to understand its general structure and flow of discussion roughly and to pay particular attention to Issue Topics, choosing one textual unit for each Issue Topic. They perform the tasks in the following order: (1) Issue Topic Identification, (2) Rhetorical Status Classification, (3) FRAMING linking and (4) Issue Topic linking. As mentioned earlier, tasks (2) and (3) logically should be performed simultaneously since the process which identifies FRAMING-main, FRAMING-sub and BACKGROUND is closely connected to the FRAMING linking task. The annotators were instructed to perform a final check to verify that each Issue Topic has at least one rhetorical status. As Issue Topics tend to have at least one unit for every rhetorical status, this check often detects slips of the attention.

	An. 1	**An. 2**
Issue Topic spans	24	27
Agreed spans (overlap)	20	
$agreement_{\text{ITI}}(i)$ (overlap)	0.833	0.741
$agreement_{\text{ITI}}$ (overlap)	**0.787**	

Table 2: Issue Topic Identification Results (in spans)

Annotator 1 \ Annotator 2	IDT	CCL	FRm	FRs	BGD	FCT	OTR	Total
IDT	171	13	4	19	0	0	3	210
CCL	0	299	142	45	0	6	4	496
FRm	0	89	1187	812	12	13	27	2140
FRs	24	15	229	2327	23	108	12	2738
BGD	3	0	11	21	150	37	1	223
FCT	12	12	52	218	0	3197	18	3509
OTR	26	7	27	9	0	99	395	563
Total	236	435	1652	3451	185	3460	460	9879

Table 3: Confusion Matrix for Rhet. Status (units)

We used Japanese Civil Case judgment documents written in several district courts, which are available publicly from a Japanese Court website (http://www.courts.go.jp/). We annotated 8 Japanese civil case judgment documents, which consist of 9,879 comma-separated units and 201,869 characters in total. The documents are written by various judges from several courts and cover the following themes: "Medical negligence during a health check", "Threatening behavior in connection to money lending", "Use of restraining devices by police", "Fence safety and injury", "Mandatory retirement from private company", "Road safety in a bus travel sub-contract situation", "Railway crossing accident", and "Withdrawal of a company's garbage license by the city".

5.2 Results

5.2.1 Issue Topic Identification

The results for the Issue Topic identification task are given in Table 2. The overall agreement ratio observed is 0.787. An error analysis showed that the two main remaining problems were due to the splitting of an Issue Topic by one annotator and not by the other, and a different interpretation of whether compensation calculations should be annotated or not.

5.2.2 Rhetorical Status agreement

Agreement of rhetorical classification was measured at K=0.70 (N=9879; n=7, k=2; Cohen). Note that the number of units N (entities assessed) is the number of comma- (or sentence-final punc-

	An. 1	An. 2
Annotated units	9336	9446
Agreed units	8169	
$agreement_{\text{ITL}}(i)$	0.875	0.865
$agreement_{\text{ITL}}$	**0.870**	

Table 4: Issue Topic Linking Results (in units)

	An. 1	An. 2
# of source spans(FRs or BGD)	544	666
# of source spans with links	527	602
# of agreed source spans with link	378 (67.26%)	
# of agreed links	250	
$agreement_{\text{fl}}$	**0.661**	

Table 5: FRAMING Linking Results

Baseline Model	$agreement_{\text{fl}}$
Random	0.016
Popularity	0.024
Nearest	0.644

Table 6: FRAMING Linking Baselines

tuation) separated text pieces, as opposed to sentences in previous work. Table 3 gives the confusion matrix for the Rhetorical Status Classification task. Although the Kappa value indicates good agreement for rhetorical classification overall, the confusion matrix shows certain systematic assignment errors. In particular, FRAMING-main and FRAMING-sub are relatively often confused, indicating that our current annotation guidelines should be improved in this respect.

5.2.3 Issue Topic Linking agreement

The result for Issue Topic linking is shown in Table 4. At 0.870, the agreement ratio indicates good agreement. The annotators seem to have little trouble in determining which Issue Topic each sentence relates to. This is probably due to the fact that the judgment documents are closely structured around Issue Topics, as per our working hypothesis. Annotators often arrive the same interpretation because the argument is logically structured and the components necessary for interpretation can be found nearby, as the strong performance of the Nearest baseline demonstrates. However, we also noticed an adverse effect concerning Issue Topics. Judges sometimes reorganize the Issue Topics that were previously defined, for instance, by merging smaller Issue Topics, or in the case of dependencies between Issue Topics, by dropping dependent Issue Topics when the Issue Topics they depend on have collapsed during the trial. Such reorganizations can cause disagreement among annotators.

In sum, the detection of Issue Topic level argument structure seems to be overall a well-defined task, judging by the combined results of Issue Topic Identification and Linking.

5.2.4 FRAMING Linking agreement

Agreement of rhetorical status classification of text units involved in FRAMING linking was measured at K=0.69 (N=9879; n=4, k=2) and source agreement is given in Table 5. The baseline results are given in Table 6. The Nearest baseline model shows a rather high score ($agreement_{\text{fl}}$=0.644) when compared to the human annotators ($agreement_{\text{fl}}$=0.661). The distance between source spans and destination spans clearly influences the FRAMING linking task, showing a strong preference by the judges for a regular argumentation structure. We also observe that the distances involving FRAMING links are shorter than those for Issue Topics.

Trying to explain the relatively low human agreement, we performed an error analysis of the linking errors, classifying the 128 errors made during FRAMING linking. We distinguish destination spans that show character position overlap across annotators, from those that do not. For those that have overlapping spans, we check whether this corresponds to shared content in a meaningful manner. Even for those spans that are not shared in terms of character positions, content could still be shared, as the spans could be paraphrases of each other, so we check this as well. We found that 26 error links had meaningful overlap and 22 error links were reformulations. If we were to consider "reformulation" and "meaningful overlap" links as agreed, the $agreement_{\text{fl}}$ value would rise to 0.788. This is potentially an encouraging result for an upper bound on how much annotators naturally agree on FRAMING linking.

Most errors that we categorized as "different meaning" are caused by non-agreement during the FRAMING-main identification stage. From this result, we conclude that improving the instructions for the identification of FRAMING-main is vital for the second phase of our annotation work. However, an interesting result is that even if annotators disagree on FRAMING-main identification, the

non-agreed links still share linking structure. We observe that often the same set of source spans are linked to some destination span, although the destination itself is different across annotators. Our agreement metrics are thus underestimating the degree of shared linkage structure.

6 Related Work

There is little work on the summarization of Japanese judgment documents, Banno et al. (2006) amongst them. They used Support Vector Machines (Joachims, 1999) to extract important sentences for the summarization of Japanese Supreme Court judgments.

Several past studies share our interest in capturing the argumentation with rhetorical schemes.

Mochales and Moens (2011) presented an argumentation detection algorithm using state-of-the-art machine learning techniques. They annotated documents from the European Court of Human Rights (ECHR) and the Araucaria corpus for argumentation detection, achieving inter-annotator agreement of K=0.75 in Cohen's Kappa (ECHR). On a genre other than legal text, Faulkner (2014) annotated student essays using three tags ("for", "against" and "neither"), reaching inter-annotator agreement of K=0.70 (Cohen). As far as the rhetorical status classification part of our scheme is concerned, the closest approach to ours is Al Khatib et al. (2016), but they do not employ any explicit links, and they work on a different genre (news editorials).

A task related to our linking steps is the determination of relations between argument components. Stab and Gurevych (2014) annotated argumentative relations (support and attack) in essays; they reported inter-annotator agreement of K=0.81 (Fleiss) for both support and attack. Hua and Wang (2017) proposed an annotation scheme for labeling sentence-level supporting arguments relations with four types (STUDY, FACTUAL, OPINION, REASONING). Their results for argument type classification are as follows: K=0.61 for STUDY, K=0.75 for FACTUAL, K=0.71 for OPINION, and K=0.29 for REASONING.

However, these two relation-based studies discover only one aspect of argument structure, whereas our combination of linking tasks and a rhetorical status classification task means that we address the global hierarchical argument structure of a text.

There has also been some recent work on agreement metrics for argument relations. As far as agreement on detection of argumentative components is concerned, Kirschner et al. (2015) point out that measures such as kappa and F1 score may cause some inappropriate penalty for slight differences of annotation between annotators, and proposed a graph-based metric based on pair-wise comparison of predefined argument components. This particular metric, while addressing some of the problems of kappa and F1, is not directly applicable to our annotation where annotators can freely chose the beginnings and ends of spans. Duthie et al. (2007) introduce CASS, a further, very recent adaptation of the metric by Kirschner et al. that can deal with disagreement in segmentation. However, the only available implementation is based on the AIF format.

7 Discussion and future work

It is hard to evaluate a newly defined, complex task involving argumentation and judgment. The task we presented here captures much of the information contained in legal judgment documents, but due to its inherent complexity, many different aspects have to be considered to see the entire picture. Our annotation experiment showed particularly good agreement for the rhetorical status labeling task, suggesting that our adaptation to the Japanese legal system was successful. The agreement on Issue Topic Identification and linking was also high. In contrast, the FRAMING linking, which annotators disagreed on to a higher degree than in the other tasks, suffered from the difficulty of identifying destination spans in particular. We can improve the agreement of the FRAMING linking task by refining our guidelines. Moreover, in order to achieve our final goal of building a flexible legal summarizer, we plan to analyze the relationship between human generated summaries and annotated documents on rhetorical status and links.

The next stage of our work is to increase the amount of annotation material for the automatic annotation of judgment documents with the proposed scheme. We will automate the annotation for the rhetorical status classification task with supervised machine learning and extend the automation step by step to linking tasks, based on the result of the rhetorical status classification.

References

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Shinji Banno, Shigeki Mtsubara, and Masatoshi Yoshikawa. 2006. Identification of Important parts in judgments based on Machine Learning (機械学習に基づく判決文の重要箇所特定). In *Proceedings of the 12th Annual Meeting of the Association for Natural Language Processing*, pages 1075–1078. the Association for Natural Language Processing.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Rory Duthie, John Lawrence, Katarzyna Budzynska, and Chris Reed. 2007. The CASS Technique for Evaluating the Performance of Argument Mining. In *Proceedings of the 3rd Workshop on Argument Mining*, pages 40–49, Berlin, Germany. Association for Computational Linguistics.

Atefeh Farzindar and Guy Lapalme. 2004. LetSum, an automatic Legal Text Summarizing system. *Jurix*, pages 11–18.

Adam Faulkner. 2014. Automated Classification of Argument Stance in Student Essays: A Linguistically Motivated Approach with an Application for Supporting Argument Summarization. *All Graduate Works by Year: Dissertations, Theses, and Capstone Projects*.

Ben Hachey and Claire Grover. 2006. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345.

Xinyu Hua and Lu Wang. 2017. Understanding and detecting supporting arguments of diverse types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada. Association for Computational Linguistics.

Japanese Ministry of Justice. 2012a. Japanese code of civil procedure (article 252).

Japanese Ministry of Justice. 2012b. Japanese Code of Civil Procedure Subsection 2 Preparatory Proceedings.

Thorsten Joachims. 1999. Making large scale SVM learning practical. In *Advances in kernel methods: support vector learning*.

Judicial Research and Training Institute of Japan. 2006. *The guide to write civil judgements (*民事判決起案の手引*)*, 10th edition. Housou-kai (法曹会).

Dain Kaplan, Ryu Iida, Kikuko Nishina, and Takenobu Tokunaga. 2011. Slate A Tool for Creating and Maintaining Annotated Corpora. *Jlcl*, 26(Section 2):91–103.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications. pages 1–11.

Raquel Mochales and Marie Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

M. Saravanan and B. Ravindran. 2010. Identification of Rhetorical Roles for Segmentation and Summarization of a Legal Judgment. *Artificial Intelligence and Law*, 18(1):45–76.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 1501–1510.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

The Secretariat of Supreme Court of Japan. 1990. *The new format of Civil judgements : The group suggestion from the improving civil judgments committee of Tokyo High/District Court and the improving civil judgments committee of Osaka High/District Court (*民事判決書の新しい様式について：東京高等・地方裁判所民事判決書改善委員会, 大阪高等・地方裁判所民事判決書改善委員会の共同提言*)*. Housou-kai (法曹会).