2017 9th International Conference on Knowledge and Systems Engineering(KSE)

Designing an annotation scheme for summarizing Japanese judgment documents

Hiroaki Yamada*, Simone Teufel*†, Takenobu Tokunaga*
*School of Computing, Tokyo Institute of Technology, Tokyo, Japan
Email: yamada.h.ax@m.titech.ac.jp, take@c.titech.ac.jp
†Computer Laboratory, University of Cambridge, Cambridge, United Kingdom
Email: simone.teufel@cl.cam.ac.uk

***Abstract*—We propose an annotation scheme for the summarization of Japanese judgment documents. This paper reports the details of the development of our annotation scheme for this task. We also conduct a human study where we compare the annotation of independent annotators. The end goal of our work is summarization, and our categories and the link system is a consequence of this. We propose three types of generic summaries which are focused on specific legal issues relevant to a given legal case.**

I. Introduction

The task of automatic summarization has become more and more crucial for dealing with the information overload in many aspects of society. This is no different in the legal domain. The legal professions, including lawyers and judges, are at risk of becoming overwhelmed by too many documents that are relevant to their specific task.

One of the most important types of legal document in the Japanese legal system is the judgment document, a direct output from court trials. The Japanese Code of Civil Procedure demands that "the court renders its judgment based on the original judgment document." [1] During the construction and analysis of their cases, the legal professions rely heavily on judgment documents, yet they are far too long and linguistically complex to read every document in detail. Well-formed summaries of judgment documents would provide a solid solution to the problem, as they would enable a decision of which documents to read with full attention.

Manual summaries of judgment documents are expensive and time-consuming to produce and thus not universally available. There is, therefore, a significant need for the automatic and on-demand summarization of judgment documents. Our final goal is to develop methods for generating these. Our main observation is that the structure of the legal argument can guide summarization. To achieve this goal, we start by designing an annotation scheme for summarizing Japanese judgment documents. Based on an initial adaptation of an existing scheme for legal documents, we conducted an initial pilot study with two annotators. The results were encouraging but indicated that more detailed modeling of the document structure was needed. After introducing these new developments, we conducted a second study.

In this paper, we will describe the process of arriving at our annotation scheme. We will also outline how our annotation scheme can contribute to the final output, summaries of various granularities.

II. Related Work

In legal text processing, there is a tradition of using rhetorical analysis for summarization, an approach initially proposed by Teufel and Moens [2] for scientific articles. Hachey and Grover [3] were the first to apply it to legal texts in the context of English law; in their system, a sentence is labeled according to its rhetorical role in the overall judgment document. Table I shows Hachey and Grover's rhetorical labels.

There are only a few studies on the summarization of Japanese judgment documents, e.g., Banno et al. [4]. They used Support Vector Machines to extract important sentences for the summarization of Japanese Supreme Court judgments.

Several studies process legal texts from a perspective of argument mining. Mochales and Moens presented an argumentation detection algorithm using state-of-the-art machine learning techniques [5]. They report inter-annotator agreement of K=0.75 (Cohen's kappa) on the task of finding argumentation units in texts from the European Court of Human Rights. There are also studies on the relationships between arguments. Faulkner conducted an annotation of student essays concerning whether the arguments were supported and found agreement to reach K=0.70 (Cohen's kappa) [6]. Stab and Gurevych reported an inter-annotator agreement of K=0.8 (Fleiss's kappa [7]) for argumentative relations (support and attack) in essays [8] .

We combine aspects of argumentation mining and relation extraction on our annotation scheme. However, the novelty of our scheme is that it fully covers the hierarchical structure of judgment documents. Our scheme also focuses on entities called "Issue Topics," as we will explain below.

III. Scheme One and Pilot Study

In this section, we propose an adaptation of an existing annotation scheme for judgment documents to the Japanese legal system.

A. Annotation scheme

The texts we process, Japanese Civil Case judgment documents, share a common textual structure. This is a result of judges' voluntary compliance with a guideline document

978-1-5386-3576-6/17/$31.00

TABLE I
HACHEY AND GROVER'S LABELS FOR LEGAL TEXT

Label	Description
FACT	A recounting of the events or circumstances which gave rise to legal proceedings.
PROCEEDINGS	A description of legal proceedings taken in the lower courts.
BACKGROUND	A direct quotation or citation of the source of law material.
FRAMING	Part of the law lord's argumentation.
DISPOSAL	Either credits or discredits a claim or previous ruling.
TEXTUAL	A sentence which has to do with the structure of the document or with things unrelated to a case.
OTHER	A sentence which does not fit any of the above categories.

TABLE II
OUR LABELS FOR LEGAL JUDGMENT DOCUMENTS

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 supports 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.

for judgment documents of civil cases [9], in particular, the "new format," which introduced issue-focused judgment in 1990 [10].

Our target for annotation and summarization is the second section of judgment documents, "Facts and Reasons," which occupies the biggest portion of the document. It consists of the main claim of the case, the facts agreed by the interested parties in advance, the issues to be argued intensively during trial, and the judicial decision.

Hachey and Grover use the sentence as the annotation unit, and in their scheme, each sentence receives exactly one rhetorical label. We also use exclusive labeling, but our definition of the smallest annotation unit is a comma-separated (and sentence final punctuation-separated) text fragment. In Japanese, such units typically correspond to linguistic clauses. This decision was necessary because in many long sentences, there are subcomponents which play different rhetorical roles from each other. It still often happens in practice that an entire sentence is tagged with only one rhetorical role, as multiple consecutive comma-separated text units can be merged in our system if they fulfill the same rhetorical function.

The adaptations we made to Hachey and Grover's scheme are as follows: We introduced a new label called IDENTIFYING, which applies to text stating a discussion topic that marks the start of argumentation, in an attempt to capture an important aspect of the legal argumentation. We expanded the definition of Hachey and Grover's DISPOSAL, which only concerns the Law Lord's agreement or disagreement with the previous ruling, to any conclusions or arguments in judgment documents, and renamed it CONCLUSION. TEXTUAL is omitted in our annotation scheme since it is not crucial in our texts. The labels we left untouched were FACT and BACKGROUND.

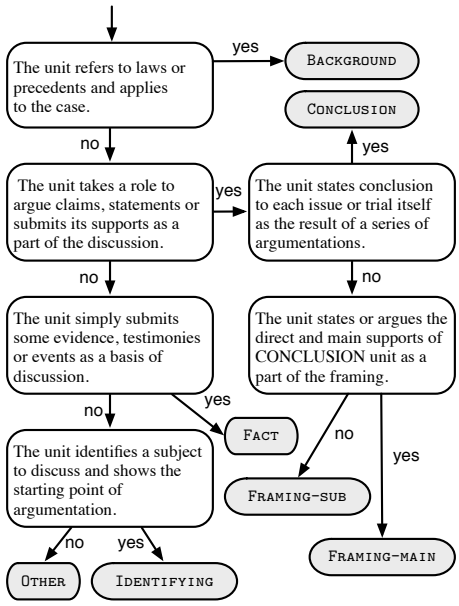

Fig. 1. Decision tree of rhetorical status

Our main modification was to split Hachey and Grover's FRAMING into FRAMING-main and FRAMING-sub. We define the higher levels of the argumentation structure as FRAMING-main and the supporting levels as FRAMING-sub. The split allows us to distinguish between different levels of importance of the supporting material covered by FRAMING. It was necessary because too much textual material would be covered by this category otherwise[1], conflicting with our goal of creating short summaries.

These adaptations resulted in the labels shown in Table II. We also supply the decision tree in Fig. 1 to facilitate our annotators' decision process.

B. Pilot Study

We conducted a pilot study to measure the inter-annotator agreement of our scheme. Two annotators participated in the experiment, the first author of this paper (who has a Bachelor of Laws degree in Japanese Law) and a Ph.D. candidate in a graduate school of Japanese Law. The legal language used in the documents is marked by extremely complex sentence structure and specialized technical terminology. The use of expert annotators is therefore essential.

We use Japanese Civil Case judgment documents written in district courts, which are available publicly from a Japanese

[1]We estimated from initial annotation that 78% of our text falls into the FRAMING role.

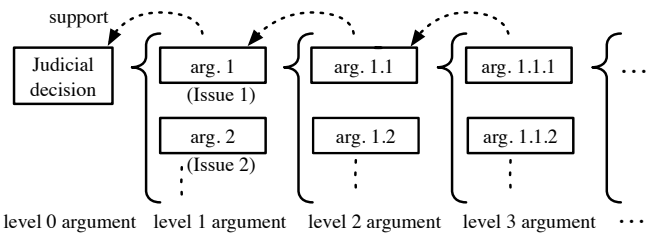

Fig. 2. Argument structure of judgment document

Court website (http://www.courts.go.jp/). In the first experiment, we use two documents, which consist of 1,517 units (30,497 characters). The annotators use the browser GUI-based annotation tool "Slate" [11].

We measure agreement using Cohen's kappa [12], a standard metric of inter-annotator agreement for categorical decisions. Kappa is a chance controlled metric that ranges between -1 and 1. The inter-annotator agreement was K=0.81 (N=1,517, n=7, k=2), where N is the number of text units, n is the number of categories and k is the number of annotators.

This level of inter-annotator agreement can be considered high; it indicates that our adaptation of Hachey and Grover's scheme for Japanese judgment documents was successful. This is not too surprising, as at least one other previous study also found good agreement in a legal system other than the English one (K=0.84 (N=16,000; n=7, k=2) for the Indian legal system [13]).

Despite this, when considering both the measured results and informal annotator feedback, we identified several points of note. Firstly, the annotators reported that the flow of the argument structure was an essential element during the annotation of rhetorical status, and formally tracking it would make their task clearer and easier. For example, in order to arrive at the correct labels, the annotators first had to find the conclusion of the overall document, second the conclusion of some important points of the judgment, which then allowed them to arrive at the FRAMING and FACT units. This feedback gave us valuable information concerning how legal documents can be best understood, and we were keen to integrate this insight into our annotation scheme.

We conclude that rhetorical status annotation alone is not sufficient to represent the information in the document. The judgment document has a complex argumentation structure, and extracting it at a more fine-grained level will enable us to generate more useful summaries, i.e., summaries focusing on the most prominent issues contained in the document. Although the rhetorical status provides useful information for the extraction and recognition of the conclusion and legal citations, in order to model the argument structure as well, we need to annotate the argumentative relations between the text units.

IV. Integrating argument structure

We revised our annotation scheme by introducing the argument structure of texts on top of rhetorical status. We observe that the original documents follow a well-defined argumentation pattern, which is illustrated in Fig. 2. A Japanese Civil Case judgment document forms one big argument, which connects the judicial decision to the plaintiff's accusation. We call this the "level 0" argument. The level 0 argument breaks down into several sub-arguments, each of which usually covers one issue topic to be challenged. We call this the "level 1" argument. Moreover, each sub-argument might itself consist of sub-arguments at lower levels (level 2,3,4...). The purpose of each sub-argument is to support the argument above it. Ultimately, at the bottom of the argument structure, facts provide the lowest level of support.

A. Issue Topic linking

The structure of the judgment document is centered around the topics of each of the strands of argumentation. 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 (Preparatory Proceedings, Japanese Code of Civil Procedure [1]). These are called the *Issue Topics*. The very plausible assumption behind this is that trials which are logically organized around Issue Topics proceed more smoothly and efficiently, particularly if the Issue Topics are well-specified.

It is our working hypothesis that Issue Topics (which correspond to the level 1 arguments in our parlance) are also extremely important in generating meaningful, coherent and useful summaries. Most legal cases consist of several Issue Topics, but 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.

We introduce a new task called *Issue Topic linking*, which defines the relation between each textual unit and its concerning Issue Topic. Annotators are instructed to indicate the Issue Topics in the text and to assign them identifiers. The annotators are asked to find the first continuous text span that explaining the Issue Topic best (defined as "in the most straightforward way"). The span must be continuous but can consist of multiple text units.

However, not all text units are related to specific Issue Topics. Some text units concern matters of the trial itself, for example, the overall conclusion or introduction. We define a special Issue Topic ID of value 0 to cover such cases. Overall, we expect this task to be relatively uncontroversial because the documents are created in a manner which is organized around Issue Topics.

B. FRAMING linking

The more detailed argument structure below level 2 can provide additional useful information for summarization. Our second new task (or annotation level) is called *FRAMING linking*. Our rhetorical status annotation partially models this

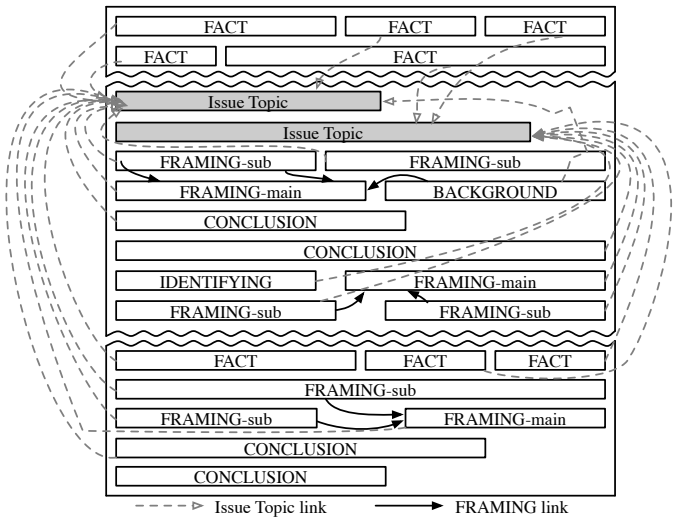

Fig. 3. Sample annotation

lower level argument structure by defining level 2 text units as FRAMING-main, and text units at level 3 and lower as FRAMING-sub. We further annotate relations between the level 2 and level 3 (and below) argument structure by introducing "support" links in the form of FRAMING linking between them.

Text units labeled with BACKGROUND and FRAMING-sub can optionally be linked to the FRAMING-main unit if the annotator considers that the BACKGROUND or FRAMING-sub units argumentatively supports the FRAMING-main. The annotators signal that there is a relation between FRAMING-sub/BACKGROUND and FRAMING-main by drawing a link in the annotation tool. The semantics of link is that the origin (BACKGROUND and FRAMING-sub) supports the destination (FRAMING-main).

Fig. 3 shows a sample of annotation with our proposed scheme. Each box corresponds to a unit with its rhetorical status.

C. Agreement metrics

1) Issue Topic identification and linking: For measuring agreement on the identification of Issue Topics, we consider two spans as agreed if more than 60% of their characters agree, assuming that those spans principally represent the same content. This criterion enables us to disregard the differences in the locations in the text and that in superficial linguistic expressions. Although we instructed the annotators to mark the first appearance of an Issue Topic when multiple spans in different locations represented the same Issue Topic, the annotators sometimes mistakenly marked the second or later spans. We initially set the threshold to 80%, but a manual inspection revealed that this resulted in many non-matching spans despite the fact that they represented the same Issue Topic. We confirmed that the 60% threshold identifies spans representing the same Issue Topic as agreed without incurring any false positives.

As the annotators may disagree on the number of Issue Topics they recognize, we report an average of two annotation metrics. $AnnotatorSet$ is a set of annotators, and i corresponds to an annotator. We first calculate an agreement score for each annotator, taking each annotator in turn as the "gold standard." In Equation (1) $a_s(i)$ is the number of spans agreed between annotator i and others, and $spans(i)$ is the number of spans annotated by annotator i. As the overall agreement score, we report the average of the two as given in (2):

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

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

$$where\ i \in AnnotatorSet.$$

For Issue Topic linking, the annotators assign an Issue Topic ID to a supporting text unit (the link *source*) to establish a link from the text unit to the Issue Topic (the link *destination*). As far as the number of sources involved in Issue Topic linking is concerned, the previous experiment showed that they are almost identical across annotators. Our overall agreement ratio is, therefore, a simple average of the two annotator's ratios. Again, we calculate agreement ratio for each annotator as in (3), and average them as defined in (4). $a_u(i)$ is the number of units agreed between annotator i and others and $units(i)$ is the number of units annotated by annotator i:

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

$$agreement_{ITL} = \frac{\sum_i agreement_{ITL}(i)}{|AnnotatorSet|}. \tag{4}$$

2) FRAMING linking: FRAMING linking is the most difficult task in our scheme. FRAMING links can hold from either BACKGROUND or FRAMING-sub to FRAMING-main, but they are optional. In contrast to Issue Topic linking, there is more possibility for disagreement (on destinations as well as sources).

We define the agreed source spans as two spans (location identity, not just textual identity), sharing more than 80% [2] of their characters. The source span agreement is calculated as the ratio of the number of agreed spans with an outgoing link, to the number of source spans with an outgoing link, as given in (5).

$$agreement_{src} = \frac{\#\ of\ agreed\ source\ spans\ with\ link}{\#\ of\ source\ spans\ with\ link}. \tag{5}$$

The FRAMING linking agreement is calculated as the ratio of the number of agreed links to the number of agreed source spans with an outgoing link, as shown in (6). We consider two links as agreed when they agree on both source and destination.

$$agreement_{fl} = \frac{\#\ of\ agreed\ links}{\#\ of\ agreed\ source\ spans\ with\ link}. \tag{6}$$

D. Experiment

In the annotation with the revised annotation scheme, in addition to the rhetorical status classification, we newly intro-

[2]The reason for setting this threshold is that we wanted to allow only short and relatively meaningless adverbial modification and such at the beginning or end of spans. As the location now has to be identical, we do not have to worry about paraphrases.

duced the following two tasks: **1. Issue Topic linking** – the Issue Topics are identified, and a unique identifier is given to each Issue Topic. All textual units labeled with a rhetorical status in the previous stage are assigned an Issue Topic ID which they support; and **2. FRAMING linking** – those textual units that support a FRAMING-main span are linked to the FRAMING-main span.

The same annotators as in the pilot study annotated eight documents, which consisted of 201,869 characters (9,879 units) in total. We provided the annotators with a guideline document of eight pages detailing the procedure. Annotators were instructed to read the target document to roughly understand its general structure and flow of discussion and to pay particular attention to Issue Topics, choosing one textual span for each Issue Topic.

In response to our earlier observations, we changed the annotation procedure by asking annotators to trace back the legal argument structure of the case during the annotation of rhetorical status. They first search for the general CONCLUSIONs of the case. They then find the CONCLUSION of each Issue Topic; next, they find the FRAMING-main which supports the CONCLUSION. Finally, they look for the FRAMING-sub elements that support the FRAMING-main. Therefore, the annotators simultaneously recover the argument structure while making decisions about the rhetorical status.

E. Result and Discussion

1) Rhetorical status classification: The inter-annotator agreement of the rhetorical classification was K=0.70 (N=9,879, n=7, k=2), noticeably lower than in our pilot experiment. This is likely to be due to variation in the data and no cause for worry, as it is still within Krippendorff's range of marginal agreement [14]. Looking at the results in detail, we observed certain systematic classification errors. Particularly, FRAMING-main and FRAMING-sub are often confused, indicating that our current annotation guidelines should be improved in this respect.

2) Issue Topic linking: The overall agreement ratio of Issue Topic Identification was $agreement_{ITI}$=0.79. A post hoc analysis showed that the two main causes of errors were discrepancies in identifying Issue Topic spans and differing opinions about how to treat compensation calculations.

In Issue Topic linking, we observe an agreement ratio of $agreement_{ITL}$=0.87. This means that the annotators had little difficulty in determining the text units supporting an Issue Topic. In combination with the result of Issue Topic identification above, capturing the argument structure at the Issue Topic level seems to be a well-defined task. This is probably a sign that the surface structure of the text sufficiently reflects the argument structure at this level in the judgment documents.

3) FRAMING linking: The agreement in source spans was measured at $agreement_{src}$=0.67 (average of 0.72 and 0.63), whereas full FRAMING linking agreement (source and destination agreement) was $agreement_{fl}$=0.66. Trying to explain the relatively low human agreement, we performed a second post-hoc analysis of the linking errors. We manually analyzed overlap of non-agreed destination spans (i.e., those which had overlap lower than 80%), in order to establish whether the overlap is meaningful (e.g. a reformulation). We found that in 48 cases out of 128 errors, there was meaningful overlap in the destination spans. This result shows the 80% threshold was too strict for the task, similar to the effect observed earlier during Issue Topic identification (Section IV-C1). If we were to consider the manually checked links with identical meaning but less than 80% character overlap as agreed, FRAMING linking agreement would rise to $agreement_{fl}$=0.79. This is an encouraging result: annotators potentially agree to a high degree on FRAMING linking, even though we cannot yet determine all of this agreement automatically.

The task we presented here captures much of the information contained in judgment documents, but due to its complexity, many aspects have to be considered to see the entire picture. Our annotation experiment showed particularly good agreement for the rhetorical status classification task, suggesting that our adaptation of Hachey and Grover's scheme to the Japanese legal system was successful. The agreement on Issue Topic linking was also high. In contrast, the FRAMING linking suffered from the difficulty of identifying destination spans in particular. By refining our guidelines, we hope to be able to improve the agreement of the FRAMING linking task. A more detailed error analysis and a discussion of metrics are reported in Yamada et al. [15].

V. Summary Design

Our aim is to reduce the cost of legal professions' investigation during their preparation for a case. In this phase of their work, the legal professions need access to the results of past trials and access to similar cases to the ones they currently work on. What would be of value in this situation is a summary which provides information of the judge's decisions in trials, including the argumentation that supports the decision. We propose to generate informative summaries of Japanese judgment documents based on an automatic annotation along the lines of this paper. We designed three basic archetypes of such summaries that could be realized with our proposed scheme:

Type A: The simplest summary consists of only the final conclusion and the major supporting argument. Determination of the rhetorical status is sufficient for this; we would simply choose only those text units labeled with CONCLUSION and FRAMING-main.

Type B: This type of summary additionally incorporates Issue Topic information, i.e. the Issue Topic text itself, and other text supporting it. This can be recovered fully from our Issue Topic links. A type B summary would thus be able to cover multiple Issue Topics mentioned in the original judgment document.

Type C: This type of summary would be structured and built like a Type B summary, but would provide readers with further information focused on a specific issue topic. In addition to the conclusions for Issue Topics, the summary would provide

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 2]: 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 3]: 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.

Fig. 4. Sample summary text (Type C, Issue Topic 2)

an Issue Topic together with supporting claims, facts and the application of the law. It would be more fine-grained than Type B because it captures all levels of the argumentation, including subordinate information of each FRAMING-main, such as BACKGROUND and FRAMING-sub. Fig. 4 shows our translation of an ideal sample of a Type C summary, which is manually extracted from an actual judgment document. The sample is focused on a specific issue topic of the case ("[Issue Topic 2]"). It consists of three parts: the description of the case, the argumentative part of Issue Topic 2 and 3 (material under the bold text headers), and the final conclusion of the case. Such a detailed and high-quality summary would not be possible without the issue topic-based argument structure captured in our scheme.

VI. Conclusion

In this paper, we described the development of our annotation scheme, which is based on our observations of the Japanese legal system. Our annotation scheme showed fair inter-annotator agreement in rhetorical status classification, Issue Topic linking, and FRAMING linking, although some problems in the FRAMING linking task remain.

The next stage of our work is to revisit the annotation guidelines and scheme considering the result of our experiments and to automate the annotation.

We will start with the automation of the tasks of rhetorical status classification and Issue Topic identification as sequential labeling tasks. We consider taking an SVM or conditional random field approach with features such as positions of units, unit length, cue phrases, legal citations, named entities and neighbor units' labels. As for linking tasks, we will classify in a pair-wise manner whether there is a link between a pair of units or not. Eventually, we will automatically generate summaries using the automatically extracted information. An Integer Linear Programming based approach seems particularly suited to combine the often conflicting demands of importance and argumentative support.

References

[1] Ministry of Justice, Japan, "Japanese code of civil procedure," *Japanese law translation*, 2012. [Online]. Available: http://www.japaneselawtranslation.go.jp/law/detail/?id=2834&vm=04&re=01&new=1

[2] S. Teufel and M. Moens, "Summarizing scientific articles: Experiments with relevance and rhetorical status," *Computational Linguistics*, vol. 28, no. 4, pp. 409–445, 2002.

[3] B. Hachey and C. Grover, "Extractive summarisation of legal texts," *Artificial Intelligence and Law*, vol. 14, no. 4, pp. 305–345, 2006.

[4] S. Banno, S. Matsubara, and M. Yoshikawa, "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*. the Association for Natural Language Processing, 2006, pp. 1075–1078.

[5] R. Mochales and M. F. Moens, "Argumentation mining," *Artificial Intelligence and Law*, vol. 19, no. 1, pp. 1–22, 2011.

[6] A. Faulkner, "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*, 2014.

[7] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[8] C. Stab and I. Gurevych, "Annotating argument components and relations in persuasive essays." in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, 2014, pp. 1501–1510.

[9] Judicial Research and Training Institute of Japan, "The guide to write civil judgements (民事判決起案の手引)." Housou-kai (法曹会), 2006.

[10] The Secretariat of Supreme Court of Japan, " 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), 1990.

[11] D. Kaplan, R. Iida, K. Nishina, and T. Tokunaga, "Slate A Tool for Creating and Maintaining Annotated Corpora," *Journal for Language Technology and Computational Linguistics*, vol. 26, no. Section 2, pp. 91–103, 2011.

[12] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 4 1960.

[13] M. Saravanan and B. Ravindran, "Identification of Rhetorical Roles for Segmentation and Summarization of a Legal Judgment," *Artificial Intelligence and Law*, vol. 18, no. 1, pp. 45–76, 2010.

[14] K. Krippendorff, *Content Analysis An Introduction to Its Methodology*, 2004.

[15] H. Yamada, S. Teufel, and T. Tokunaga, "Annotation of argument structure in japanese legal documents," in *Proceedings of the 4th Workshop on Argument Mining*. Association for Computational Linguistics, 2017, pp. 22–31.