Artificial Intelligence and Law
https://doi.org/10.1007/s10506-019-09242-3

Building a corpus of legal argumentation in Japanese judgement documents: towards structure-based summarisation

Hiroaki Yamada[1] · **Simone Teufel**[1,2] · **Takenobu Tokunaga**[1]

Abstract
We present an annotation scheme describing the argument structure of judgement documents, a central construct in Japanese law. To support the final goal of this work, namely summarisation aimed at the legal professions, we have designed blueprint models of summaries of various granularities, and our annotation model in turn is fitted around the information needed for the summaries. In this paper we report results of a manual annotation study, showing that the annotation is stable. The annotated corpus we created contains 89 documents (37,673 sentences; 2,528,604 characters). We also designed and implemented the first two stages of an algorithm for the automatic extraction of argument structure, and present evaluation results.

Keywords Argumentation structure · Summarisation · Manual annotation · Machine learning

1 Introduction

Information overload has become problematic in many aspects of society. This is no different in the legal domain. During the process of constructing and analysing a particular case, legal practitioners, including lawyers and judges, rely heavily on information about similar cases.

✉ Hiroaki Yamada
yamada.h.ax@m.titech.ac.jp

Simone Teufel
simone.teufel@cl.cam.ac.uk

Takenobu Tokunaga
take@c.titech.ac.jp

[1] Tokyo Institute of Technology, Tokyo, Japan

[2] University of Cambridge, Cambridge, UK

Published online: 15 February 2019

One of the most important sources of such information in the Japanese legal system is the judgement document, a direct output from court trials. But there are typically far too many such documents that are relevant. Adding to this, the documents are long and linguistically complex, so that it becomes impossible to read them all carefully.

Well-formed summaries of judgement documents would provide a solid solution to the problem, as they would facilitate the decision of which documents to be read with full attention. Of course, the highest quality summaries are those written by experts, but they are not universally available, as the manual summarisation process is time-consuming and expensive. There is therefore a significant need for the automatic on-demand summarisation of judgement documents. Our final goal is to develop methods for generating these.

Our main observation is that the structure of the legal argument can guide summarisation. In the case of our target documents, a common structure demonstrably exists, which is based around the so-called "Issue Topic", a legal concept corresponding to pre-defined main points to be discussed in a particular court case. Each Issue Topic is associated with a conclusion by the judge, and with supporting arguments for the decision. The writers of the documents, who are judges, consistently follow the principle of legal arguments, which results in a well-formed shared structure. It is this fact that we exploit in our work.

In this article, we present our new corpus of Japanese civil law judgement documents which are manually annotated with the documents' argumentative structure. The corpus contains 89 documents (37,673 sentences; 2,528,604 characters) and their summaries. We also present the corresponding annotation scheme designed for capturing this structure, and show with an agreement study that our scheme is stable. We also designed and implemented the first two stages of an algorithm for the automatic supervised extraction of argument structure as a proof of concept of our approach. Both stages rely on a mixture of n-gram, location and cue phrase features. The first stage, the identification of Issue Topic units, was evaluated at a performance of $F = 0.52$, whereas the second stage, Rhetorical Classification, performs at $F = 0.63$. These feasibility studies confirm that our scheme and the resulting corpus can be used as training material for the automatic extraction of argument structure by a supervised machine learning approach.

2 Description of judgement documents

The type and structure of legal documents in a given country are always affected by the national legal system in force. In the Japanese legal system, the judgement document is one of the most important types of legal text. Judgement documents are written by professional judges, who, after passing the bar examination, are intensively trained to write such judgements. In particular, we work with civil (as opposed to criminal) case judgement documents from courts of the first instance.

The first observation we make about these documents is that the language used is complex and often involves extremely long sentences.[1] One reason for the long sentences is the requirement on judges to define their statement precisely and strictly, which they often do by adding additional restrictive clauses to sentences.

Another observation is that understanding the language is difficult even for humans because everyday terms can take on new technical meanings in a legal context. For example, the terms "*aku-i*" and "*zen-i*", which mean "maliciousness" and "benevolence" respectively in everyday language, take on the specialised meanings of "knowing a fact" (*aku-i*) and "not knowing a fact" (*zen-i*) in law.

A third observation is that the judges seem to actively comply with a particular guideline document for writing judgement 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 judgement 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, we observed a high degree of compliance with the new format of the guidelines in recent Japanese judgement documents. As a result, there are strong similarities in argument structure across judgement documents, most easily observed in a common section structure, often with similar or identical headlines used. This section structure is as follows: The "Fact and Reasons" section takes up the biggest portion of the document and is therefore the target of our summarisation 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 section contains the judicial decision.

3 Argument structure in judgement documents

We will next describe the *argument* structure of the legal argument as opposed to the formal section structure described above. *Issue Topics*, the contentious items to be argued about in court, are the main organising principle in the logical structure of the judgement document. In the Japanese judicial system, Issue Topics are explicitly defined in so-called "preparatory proceedings", which take place ahead of each civil law trial under participation of all parties (Japanese Ministry of Justice 2012).

Most legal cases consist of several Issue Topics. What could be a possible Issue Topic depends on the case and is in principle open-ended. Examples include "whether the defendant was negligent in fulfilling their duties", "the defendant's exact actions during the crucial time frame" or "the effect of (a particular) law".

Figure 1 shows the structure of a Japanese Civil Case judgement document. The document forms one big argument, with the judicial decision (the final conclusion to

[1] The average sentence length of in our corpus of 89 documents is 44.7 words (short-unit-word, SUW), as opposed to 23.1 words (SUW) in general Japanese language [estimated from the BCCWJ (Maekawa et al. 2014)].

the plaintiff's accusation) at the root of the argument. We call the judicial decision the "level 0" argument. The level 0 argument breaks down into several sub-components (argumentation strands), each of which usually covers one Issue Topic. Each Issue Topic argumentation strand is hierarchically organised, whereby level i components might themselves consist of sub-components at lower levels $i + 1$, $i + 2$, etc.

For our analysis, the fact that the document text can be split into different argument strands according to Issue Topics is crucial: we consider each text segment as logically belonging to one Issue Topic. We could say that there is a weak form of "support" relationship between the Issue Topic (level 1) and all components at lower levels of that Issue Topic's argumentation hierarchy. Two of our annotation tasks treat phenomena directly related to Issue Topics: In "Issue Topic Identification" (cf. Sect. 5.1), Annotators identify Issue Topics and then, in "Issue Topic Linking" (cf. Sect. 5.3) classify each text piece as belonging to exactly one Issue Topic.

Within an Issue Topic tree, different levels often correspond to different rhetorical functions in the argument. For instance, lower levels tend to consist of simple supporting facts, whereas higher levels are main conclusions or high-level supporting argumentative material. Classification of the rhetorical function of a text piece is a standard task in legal text processing; our version of this task ("Rhetorical Classification", cf. Sect. 5.2) distinguishes 7 categories.

Finally, elements at lower levels of an Issue Topic tree often directly logically support those at higher levels. To capture this effect, we introduce a task called "FRAMING Linking" (shown as arrows in Fig. 1; cf. Sect. 5.4), a task which corresponds to the classification of the argumentative relation "support" in argumentation mining research.[2]

Once the four aspects of argumentation structure as described above have been recognised (either automatically or manually), we are in a position where we can create various summaries, which we will describe in the following.

4 Summary archetypes

Our summary design is based on extractive summarisation, one of the two main methods of arriving at an automatic summary. Extractive summarisation extracts phrases or sentences from the source documents in order to generate a summary, while abstractive summarisation generates a summary by compressing or manipulating an internal representation of the text in some form and then creating new text from scratch. Although abstractive summarisation can mimic human shortening or generalisation techniques, it is not always the best practical summarisation strategy. For instance, it can result in ungrammatical text, which is undesirable in our application as it might confuse or mislead the user.

[2] In the kind of legal argument treated here, "attack" relations also exist, for instance, when a plaintiff argues against a claim by the defendant, but because the orientation of the support/attack relation can be inferred by the default roles played by plaintiff and defendant, it is not necessary to annotate such "attack" relations explicitly.

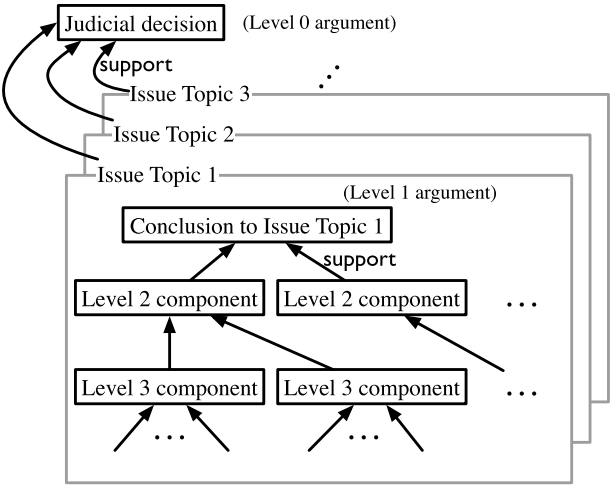

Fig. 1 Argument structure of judgement document

Some of the documents we work with already have summaries written by humans published alongside the judgement documents. However, the number of such summaries is low (on the order of few hundreds, as opposed to thousands of judgement documents), and even those summaries that do exist are not necessarily ideal by our definition. To our surprise, we found several uninformative and badly structured summaries that do not even state the main conclusion of the trial, although this is essential information for the legal professionals during their search task.

Both of these reasons speak against using supervised machine learning and training it directly on the available summaries. Instead, our system design is based on the hypothesis that Issue Topics are the lynchpin for generating coherent and meaningful summaries. In the best summaries, the logical flow is organised in such a way that the final judicial decision can be traced back through each Issue Topic's connections. We design three basic archetypes of target summaries based on this hypothesis, as illustrated in Fig. 2, and fill these archetypes with information found in the full documents as follows:

- Type A: The simplest summary consists of the judge's final decision of the case (level 0), conclusions (level 1) and some of the major supporting argumentative components for the decision (level 2; FRAMING-main[3]). Consider the example in Fig. 3, which uses material manually extracted from an actual judgement document (our translation from Japanese).
- Type B1: This type of summary additionally incorporates Issue Topic information, i.e. it states the Issue Topics themselves, and gives other components supporting them. A type B1 summary covers multiple Issue Topics, cf. Fig. 4.

[3] We will define these categories in Sect. 5.

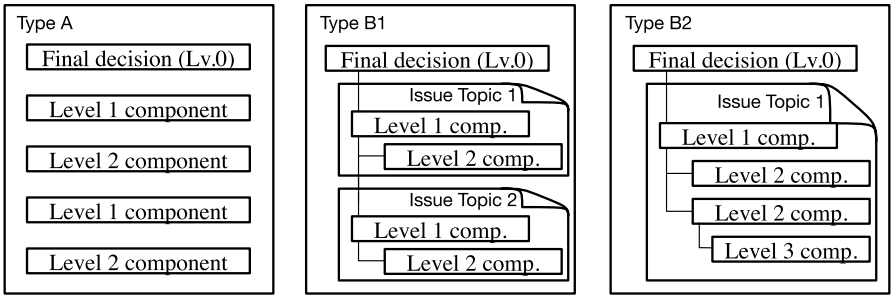

Fig. 2 Summary archetypes

The measures performed by the officer comply with the normal procedure for inspection **(Level 2: FRAMING-main)**. On the basis of his inspection, officer D could not have realised that the estate was stigmatised **(Level 2: FRAMING-main)**. 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 **(Level 2: FRAMING-main)**. Therefore, the plaintiff's claim is unreasonable since we just found that the officer was not negligent **(Level 1: CONCLUSION)**. Given what has been said above, it is not necessary to judge the other points; the plaintiff's claim is unreasonable so the judgement returns to the main text **(Level 0: CONCLUSION)**.

Fig. 3 Type A summary example

[Issue Topic 1]: Whether the execution officer D was negligent or not.

The measures performed by the officer comply with the normal procedure for inspection **(Level 2: FRAMING-main)**. On the basis of his inspection, officer D could not have realised that the estate was stigmatised **(Level 2: FRAMING-main)**. 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 **(Level 2: FRAMING-main)**. Therefore, the plaintiff's claim is unreasonable since we just found that the officer was not negligent **(Level 1: CONCLUSION)**.

[Issue Topic 2]: Whether the examination court was negligent or not.

In order for an action of the examination court to be illegal, exceptional circumstances must apply, for instance, the action must deviate to an extreme degree from the purpose of the examination system, or the action itself must be exceptionally inappropriate—a mere defect that requires redress, for instance by appeal to a higher court, is not enough **(Level 2: FRAMING-main)**. The fact that the execution court issued an Order of Investigation without including the request of investigating the cause of death of the deceased owner, cannot be regarded as an exceptional circumstance **(Level 2: FRAMING-main)**. The execution court never had an obligation to establish whether the estate was stigmatised **(Level 2: FRAMING-main)**. From this it follows that the examination court was not negligent, so the plaintiff's claim in this point is unreasonable **(Level 1: CONCLUSION)**. Given what has been said above, it is not necessary to judge the other points; the plaintiff's claim is unreasonable so the judgement returns to the main text **(Level 0: CONCLUSION)**.

Fig. 4 Type B1 summary example

– Type B2: This type of summary is similar to B1, but treats one particular Issue Topic in more depth (Issue Topic 1 in the example in Fig. 5). It additionally gives supporting argumentative components such as claims, facts and citations to laws, and thus captures all levels of the argumentation (cf. the very long Level 3 FRAMING-sub element contained in the example).

[Issue Topic 1]: Whether the execution officer D was negligent or not.

The measures performed by the officer comply with the normal procedure for inspection **(Level 2: FRAMING-main)**. On the basis of his inspection, officer D could not have realised that the estate was stigmatised **(Level 2: FRAMING-main)**. 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 **(Level 2: FRAMING-main)**.

Concerning the question whether the officer had the duty to check whether the estate was stigmatised, 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 neighbours might not have answered the officer's questions in a forthcoming manner, even if they were aware of the fact that the estate was stigmatised; there are several factors to affect the value of the estate beyond the fact that the estate was stigmatised, 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 **(Level 3: FRAMING-sub)** – 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 stigmatised **(Level 2: FRAMING-main)**. Therefore, the plaintiff's claim is unreasonable since we just found that the officer was not negligent. **(Level 1: CONCLUSION)**

Given what has been said above, it is not necessary to judge the other points; the plaintiff's claim is unreasonable so the judgement returns to the main text. (**level 0: CONCLUSION**)

Fig. 5 Type B2 summary example

In order to automatically create such summaries, Type A summaries only require that we can determine the rhetorical status (such as CONCLUSION or FRAMING-main) for each piece of text. But the more detailed and high-quality summaries of Type B1 and B2 rely on the Issue Topic-based argument structure. To provide appropriate textual material for them, we have to identify a description of the Issue Topic in question in the running text, and then connect it to larger text pieces covering the argument about this Issue Topic. For B2 summaries, we need to additionally determine the "support" links between deeper levels.

Note that this kind of summary design is very different from that of the only Japanese-language summariser for legal text currently available (Banno et al. 2006). This system, which operates on Japanese Supreme Court judgements, also relies on extractive summarisation, but uses a fixed definition of sentence importance, which is learned in a supervised manner using Support Vector Machines (SVM) and linguistic features such as morphemes, numerals, length of sentences, location, whether a sentence states *ratio decidendi* or not, and the type of conclusion (e.g., "Dismissal with prejudice" or "Reversed and remanded"). In contrast, our design relies on the different rhetorical function of text pieces.

5 Four annotation tasks

Our definition of discourse structure consists of four separate annotation tasks as follows (Anonymous, 2017).

1. *Issue Topic Identification*: find a text span that describes an Issue Topic;
2. *Rhetorical Classification*: determine the rhetorical status of each text span;
3. *Issue Topic Linking*: associate each rhetorical text span with exactly one Issue Topic;
4. *FRAMING Linking*: annotate argumentative support links between three possible rhetorical units (FRAMING-sub or BACKGROUND as source; FRAMING-main as destination).

In all annotation tasks, we use annotation units based on comma-separated units of text, which in Japanese typically correspond to linguistic clauses or phrases. In particular, we use what we call "text spans", where a text span is defined as one or more adjacent comma-separated units within one sentence (or the entire sentence if no comma is present).

We use this definition for all text spans across the tasks. However, text spans chosen during Issue Topic Identification are typically shorter (for instance, clauses and phrases which act as headlines), whereas for Rhetorical Classification, longer text spans are typically chosen by annotators. In fact, it is so common that *all* comma-separated units in a sentence share the same rhetorical role, that we considered using sentences as an alternative annotation unit, as has been done in previous work (more about this in Sect. 6.2.4). Our decision fell on text spans, however, in light of the exceptions, where different rhetorical roles share a sentence.

We will now define each task in turn.

5.1 Issue Topic Identification (ITI)

The Issue Topic is defined as the text span that describes the legal point at a question in the most straight-forward way. We instruct the annotators to find and mark the first such description in the text, under the assumption that there is only one such description (or at least that the first description is the clearest). Each Issue Topic is also assigned a unique identifier by the annotator.[4]

5.2 Rhetorical Classification (RC)

Rhetorical Classification is a commonly used approach in legal text processing for associating text pieces with their rhetorical status. Our rhetorical annotation scheme of six categories plus the OTHER category is an adaptation of Hachey and Grover (2006)'s scheme. In line with previous work, we also require classification to be exclusive, i.e., only one category can be assigned to each annotation unit.

FACT is the category used for descriptions of the facts giving rise to the case, and BACKGROUND is reserved for quotations or citations of law materials (legislation and near-precedent cases). DISPOSAL marks the final decision of the judge. IDENTIFYING is a category used for text that states discussion topics below the Issue Topic. The main argumentative material is contained in the two categories

[4] These identifiers are later normalised across annotators.

FRAMING-main and FRAMING-sub. The split corresponds to our distinction between levels 2 and 3 in the argumentation tree in Fig. 1; FRAMING-main directly supports the judge's conclusion, whereas FRAMING-sub is one of the two categories which can support FRAMING-main.

There are some text spans that aren't associated with a particular Issue Topic because they concern matters of the trial itself (such as the overall conclusion or introduction or references such as "Refer to Kou-2, pages 15, 29, 33, 169 and 220.") Material of this kind is to be annotated with category OTHER.

5.3 Issue Topic Linking (ITL)

We require that all textual material except that previously marked as OTHER is assigned to individual Issue Topics, a task called *Issue Topic Linking* in our design. Our annotators make the connection between text spans and the relevant Issue Topic explicit by marking the ID of the Issue Topic.[5]

5.4 FRAMING Linking (FL)

FRAMING links hold between BACKGROUND or FRAMING-sub (the two possible source spans), and FRAMING-main (the only possible destination span). The semantics of a FRAMING link is that the source "supports" the destination. The FRAMING-link will only be established if this support exists. Given a successful solution to FRAMING-linking, our most fine-grained summaries could display level 3 components as well as level 1 and level 2 components.

Despite the similar names, the two types of linking we define are of a different nature. Issue Topic Linking determines which Issue Topic each piece of text most strongly belongs to. It can be seen as a form of classification, where the possible classes are the Issue Topic IDs. Therefore it applies to a large number of text spans. In contrast, FRAMING Linking is much more selective; it only applies to the small number of text spans where a direct "support" relationship between levels 2 and 3 actually holds.

In both kinds of linking, annotators can disagree on what exactly the *source* of the link is, i.e., where it starts and ends. In FRAMING-linking, the annotators additionally need to delineate the text span associated with the *destination* (which is always of type FRAMING-main). This is different in the case of Issue Topic Linking, where the destinations of the links are not text spans, but IDs. We will see in Sect. 6.1.3 that the classification-like nature of Issue Topic Linking makes for easier evaluation when compared to FRAMING-linking.

Figure 6 shows part of the annotation of the text associated with the sample summaries from Sect. 4. 10 annotation units are shown; two of these describe Issue Topics 1 and 2 (Issue Topic Identification; in black). The other eight participate in Rhetorical Classification. One of these (a CONCLUSION at level 0) gives the judicial

[5] Technically, a special Issue Topic ID "0" is used for OTHER cases.

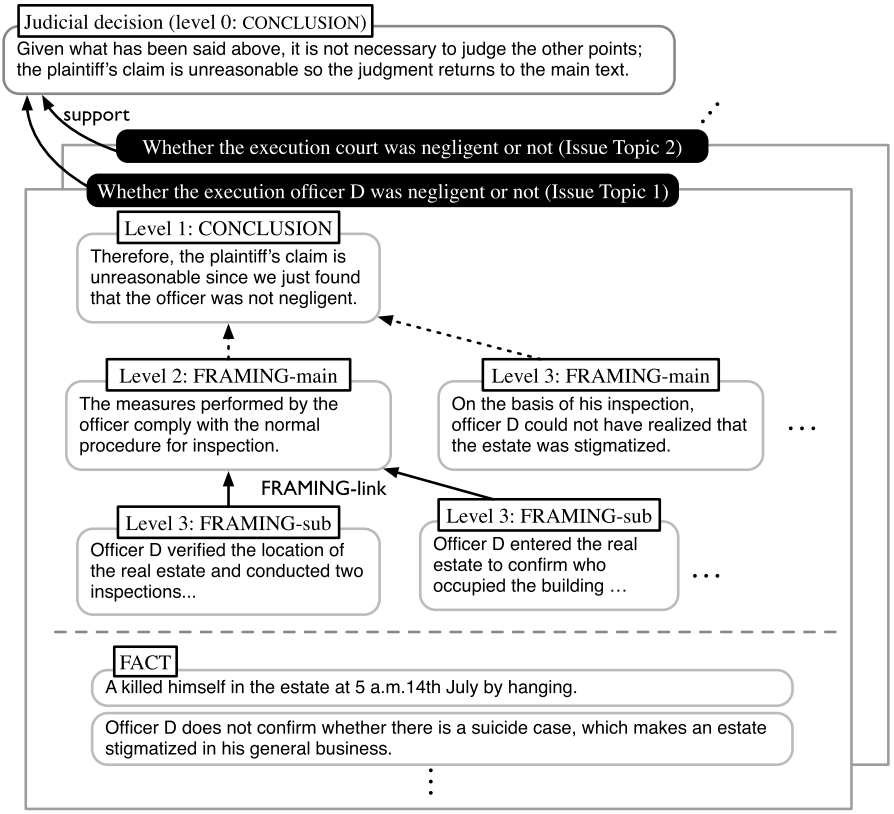

Fig. 6 Example of annotation (all 4 tasks)

decision and is thus related to the overall trial. The other 7 annotation units are associated with Issue Topic 1 (through Issue Topic Linking; here only shown by membership in the Issue Topic 1 box.). One of these seven units, also a CONCLUSION, but at level 1, gives the judge's decision on Issue Topic 1. Two others are FACTS related to Issue Topic 1. Two units are FRAMING-subs, i.e., they directly support one of the two remaining FRAMING-main. The hierarchical structure between levels 2 and 3 is directly expressed by FRAMING links (solid arrows). The other two links between level 1 and 2 (dotted arrows) exist only implicitly because the definition of FRAMING-main requires for it to support CONCLUSIONs.

6 Corpus and human annotation

The document source of our corpus is Japanese Civil Case judgement documents written in several different district courts between April 2003 to December 2016 and their summaries, downloaded from a website maintained by the court

of Japan.[6] We randomly chose 89 documents fulfilling the following selection criteria:

- The document contains at most 400,000 characters and has no encoding errors.
- The document has a published summary, which is between 150 and 450 characters long.
- The summary should have at least one sentence of general description, and one sentence of conclusion/result of the trial.

The document part of the resulting corpus contains a total of 136,972 comma-separated units (37,673 sentences). The documents (but not the summaries) were manually annotated with all four aspects of the proposed annotation scheme.

All 89 documents were annotated by a Ph.D. candidate in a graduate school of Japanese Law, who was paid for annotation. It is necessary to use expert annotators, due to the special legal language used in the documents. Out of these 89, eight documents are used for the agreement study described below. The second annotator was the first author of this paper, who holds a Bachelor of Law degree in Japanese Law.

6.1 Agreement metrics

In a task such as this, annotators can never reach 100% agreement, but if agreement is acceptably high ("stable"), it means that the guidelines describing the task are sufficiently effective. In that case, annotated materials should be usable as training for supervised learning methods even if they are created by different individuals. We therefore measure inter-annotator agreement for all four tasks.

Due to the nature of the four annotation tasks we propose here, specialised agreement metrics are necessary for all but Rhetorical Classification, which is a standard task.

6.1.1 Agreement metric for rhetorical classification

Rhetorical Classification is standardly evaluated by measuring the inter-annotator agreement (Carletta 1996), i.e., the degree to which different people agree on assigning categories, which is typically reported using Cohen's (1960) Kappa. Annotators annotate text spans, i.e., sequences of comma-separated units; note however that we have to report results in number of comma-separated units, rather than in number of text spans, as these are of variable length and start and end points might thus differ across annotators.

6.1.2 Agreement metric for issue topic identification

Descriptions of the same Issue Topic can appear in different locations, and can be expressed with superficially different linguistic expressions such as paraphrases.

[6] http://www.courts.go.jp/.

Annotators sometimes disagree which of the Issue Topic descriptions is the most appropriate.[7] As we only care to know whether annotators agree on the contents of the Issue Topics, we evaluate based on the string itself, independently of its location.

We count two spans as agreeing with each other if they overlap in more than 60% of their length in characters.[8] Per-annotator ITI agreement ($agree_{ITI}(i)$ in Eq. (1)) is then defined as the proportion of agreed spans amongst all spans identified by one annotator. Because the number of Issue Topics might differ across annotators, this agreement score is calculated for each annotator in turn, taking the other annotator as the gold standard. In Eq. (2), we then average in the obvious way.

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

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

where $spans(i)$ is the number of spans annotated by annotator $i \in AnnotatorSet$ and $a_s(i)$ is the number of spans agreed between annotator i and others.

6.1.3 Agreement metric for issue topic linking

All comma-separated units except those with the rhetorical category "OTHER" participate in Issue Topic Linking, and links go from a comma-separated unit (the source) to an Issue Topic ID (the destination). As OTHER annotation may vary across annotators, the final metric $agree_{ITL}$ in Eq. (4) is again the average per annotator of the individual numbers of accuracy given in Eq. (3):

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

$$agree_{ITL} = \frac{\sum_i agree_{ITL}(i)}{|AnnotatorSet|}, \tag{4}$$

where $units(i)$ is the number of units annotated by annotator $i \in AnnotatorSet$ and $a_u(i)$ is the number of units agreed between annotator i and others.

6.1.4 Agreement metric for FRAMING Linking

To agree on a FRAMING link, three things have to be identical: the source spans (either BACKGROUND or FRAMING-sub) must match, the destination spans

[7] Although we instructed the annotators to mark the *first* appearance of an Issue Topic when multiple spans in different locations represent the same Issue Topic, the annotators sometimes mistakenly or otherwise marked the second or later spans.

[8] We decided on this threshold by evaluating false positives and negatives with thresholds of 60%, 70% and 80%. All threshold resulted in zero false positives (i.e., all automatic matches indeed represented the same Issue Topic semantic), but we chose the 60% threshold as it naturally had the lowest false negative rate (the fewest real matches were missed).

(FRAMING-main) must match, and the link must hold between the same source and destination span.

For FRAMING Linking, two spans are defined as agreeing if they are in the identical location and share more than 80% of their characters.[9] *FRAMING source agreement* ($agree_{src}$; Eq. 5) reports the degree to which annotators agree on what the source spans for FRAMING Linking are, as the proportion of source spans of type FRAMING-sub or BACKGROUND which have an outgoing link and agree, out of the total of such spans. *FRAMING destination agreement* ($agree_{dest}$; Eq. 6) is defined as the ratio of the number of agreed links (defined as agreeing in both source and destination spans) to the number of agreed source spans with an outgoing link. FRAMING Linking consists of two stages, and if errors are made in the first stage, they will be propagated to the second stage. Equation (7) reflects this harsh reality by defining *FRAMING Linking agreement* ($agree_{fl}$), our final performance metric for FRAMING Linking, as the product of $agree_{dest}$ with $agree_{src}$.

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

$$agree_{dest} = \frac{\#\ of\ agreed\ links}{\#\ of\ agreed\ source\ spans\ with\ link} \tag{6}$$

$$agree_{fl} = agree_{src} \cdot agree_{dest} \tag{7}$$

6.1.5 Baselines for FRAMING Linking

We implemented three baselines in order to interpret our agreement results for FRAMING Linking. All three baseline models simulate only the linking step after source and destination spans have already been pre-identified, not the step of finding these spans. We give the baselines as input the source spans (those with outgoing links) and destination spans identified by one annotator. The linkings proposed by the baseline are then compared to the gold standard, which is defined as the other annotator.[10] As a consequence, we can only compare systems and baselines via FRAMING destination agreement $agree_{\text{dest}}$ (instead of the full FRAMING Linking agreement).

We created three baseline models: the "Random" model chooses one destination span for each source span randomly. The "Popularity" model chooses randomly,

[9] The reason for this stricter condition compared to ITI is that we only wanted to allow short modification (e.g., adverbials) at the beginning or end of spans. As the annotation units themselves (their locations) need to be identical, we do not have to worry about paraphrases as would be necessary in string-based comparison. Both these factors allow us to be stricter than we were for Issue Topic Identification.

[10] We considered alternative ways to suggest reasonable spans to the baseline system. Random choice of source or destination spans would result in an extremely weak baseline, as the probability of accidentally hitting a plausible source or destination span is very small. We therefore settled for an extremely strong baseline, which has access to the information of what one annotators' spans are.

but uses the observed annotation distribution of the respective other annotator. The "Nearest" model always chooses the closest following destination span (or preceding span if none exists). This is motivated by our observation that in legal arguments, the supporting material often precedes the conclusion, and is typically adjacent or at least physically close.

6.2 Agreement study

6.2.1 Materials

Eight randomly chosen judgement documents from our corpus are used for measuring inter-annotator agreement, consisting of 9,879 comma-separated units (138,482 short-unit-words). The documents are written by various judges from different courts and cover the following topics: "Medical negligence during a health check", "Threatening behaviour 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".

The annotators use 8 pages of annotation guidelines in Japanese explaining the tasks and the categories. We also supply the decision tree in Fig. 7 to facilitate the decision process during Rhetorical Classification. During the training phase, training materials separate from the documents in the agreement study were used and the paid annotator was given feedback about clear cases of wrong interpretation of the guidelines (with the first author acting as the lead annotator). Only very slight corrections to the guidelines were necessary at this stage. During the agreement study, both annotators worked entirely independently.

6.2.2 Procedure

Annotators were asked to read the entire target document to understand its general structure and flow of discussion, and to pay particular attention to Issue Topics, choosing one textual span to represent each Issue Topic (the first one, unless this first mention was not informative enough). While annotating all four tasks in order, the annotators were asked to trace back the legal argument structure of the case, first searching for the general CONCLUSIONs of the case, and then each Issue Topic's CONCLUSION; next they find the FRAMING-main which is supporting it. Finally, they look for the FRAMING-sub elements that support the FRAMING-main, expressing the "support" relationships found in the form of FRAMING links. The annotators thus recover the argument structure while making decisions about the rhetorical status at the same time.

The GUI-based annotation tool Slate (Kaplan et al. 2011) was used. Slate is a graphics-based interface that allows users to swipe in order to mark a region of text (for Issue Topic Identification), colour it by choosing a pre-defined category (for Rhetorical Classification), add properties to text spans (such as IDs for Issue Topic

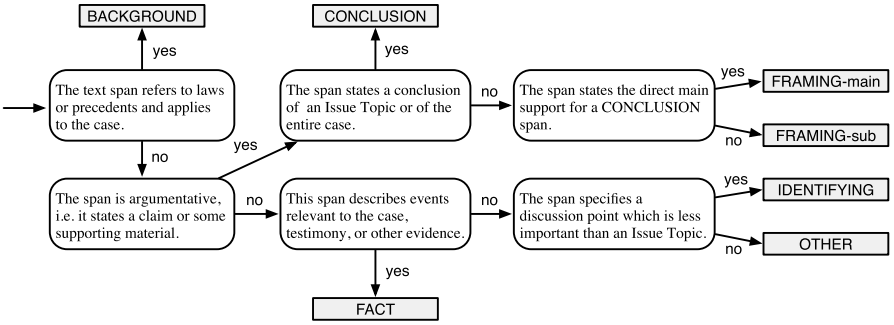

Fig. 7 Decision tree for rhetorical classification (our translation)

Linking) and add links between individual elements, which are drawn as curved lines (for FRAMING Linking). Annotators used the annotation tool offline in their own time. The annotators reported that annotation of the 8 documents took them roughly 15 h.

6.2.3 Results

Issue Topic Identification Agreement (Result) The results for the Issue Topic Identification task were measured at $agree_{\mathrm{ITI}} = 0.79$. Annotator 1 marked 24 Issue Topic spans, Annotator 2 marked 27, 20 of which were shared. One possible cause for the remaining errors was a frequently occurring item called the "compensation calculation"; annotators disagreed as to the overall importance of this item. More detailed instructions in the guidelines should help in this case. There were also some cases where annotators disagreed about whether a string containing issue-topic type material should be counted as a single Issue Topic or as two adjacent ones.

Rhetorical Classification Agreement (Result) Agreement of Rhetorical Classification was measured at K = 0.70 ($N = 9879$; $n = 7$, $k = 2$), where K is Cohen's Kappa, N is the number of text spans,[11] n is the number of categories and k is the number of annotators. There are several scales prescribing how K values should be interpreted. Out of these, Krippendorff's (2004) is the mathematically most well-founded one, but it is also strict, requiring agreement of K = 0.80 to earn the label "stable". However, even by this strict scale, our annotation would be considered "marginally stable", as it exceeds K = 0.69.

Table 1 shows the confusion matrix. Although Rhetorical Classification agreement is overall acceptable, the confusion matrix shows certain systematic assignment errors. In particular, FRAMING-main and FRAMING-sub are relatively often confused. The problem is that the categories both have a similar argumentative function, namely that of providing support for their higher-level arguments; they occur in similar locations and have similar surface characteristics such as cue phrases. They

[11] Note that N here corresponds to comma-separated units, as opposed to sentences in previous work.

Table 1 Confusion matrix for rhetorical classification (in comma-separated units)

		Annotator 2							
		IDT	CCL	FRm	FRs	BGD	FCT	OTR	Total
Annotator 1	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

Bold values indicate the numbers of units that are agreed between annotators

are thus not easily distinguished. Merging these categories would result in an overall agreement of K = 0.83 (N = 9879; n = 6, k = 2); but of course merging would defeat the final purpose of our work.

Issue Topic Linking Agreement (Result) The result for Issue Topic Linking is $agree_{\mathrm{ITL}} = 0.87$. Annotator 1 created 9336 links and Annotator 2 created 9446, out of which 8169 were shared. The annotators seem to have little trouble in determining which Issue Topic each sentence relates to. Judging by the combined results of Issue Topic Identification and Linking, the detection of Issue Topic level argument structure seems to be overall a well-defined task. This gives some credence to our working hypothesis that judgement documents are indeed closely structured around Issue Topics.

However, we noticed a phenomenon that can lead to adverse effects for the linking task. Judges sometimes reorganise Issue Topics during the trial, for instance by merging some of the smaller Issue Topics previously agreed, or by dropping Issue Topics which depend on other Issue Topics when these had collapsed during the trial. Such reorganisations can cause disagreement among annotators.

FRAMING Linking Agreement (Result) FRAMING Linking agreement was rather low at $agree_{\mathrm{fl}} = 0.44$, with source agreement $agree_{\mathrm{src}} = 0.67$ and destination agreement $agree_{\mathrm{dest}} = 0.67$. This is based on 527 source spans with links according to Annotator 1; 602 according to Annotator 2, of which 378 are agreed. The number of agreed *links* based on these spans is 250.

Since the correct identification of the text spans participating in FRAMING Linking is a precondition for FRAMING linking, measuring how well the annotators can make the 4-way distinction into FRAMING-main, FRAMING-sub, BACKGROUND and "anything else" provides an upper bound of performance that limits all further FRAMING Linking tasks. At K = 0.69 ($N = 9879$; $n = 4$, $k = 2$), this result points to the fact that low agreement for FRAMING Linking is in part a follow-on effect of disagreements in Rhetorical Classification. Figure 8 shows an example of such a disagreement: (3) and (3)′ are the same spans, but the annotators assigned different rhetorical status, FRAMING-main and FRAMING-sub, resulting in divergent FRAMING Linking structures.

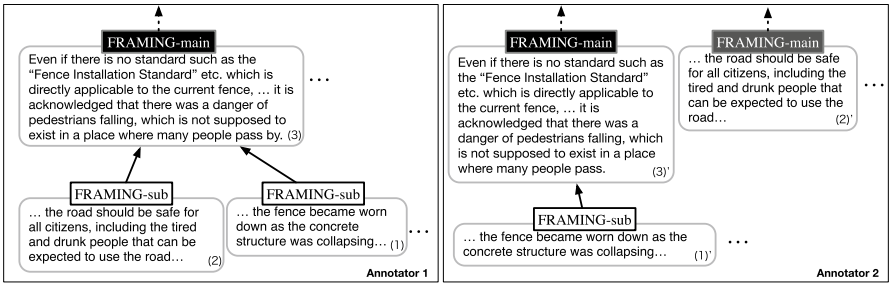

Fig. 8 Example of a disagreement in FRAMING linking (and rhetorical classification)

Next, we consider what we can learn from the baselines' performance: Random and Popularity models both perform badly at $agree_{\text{dest}} = 0.02$, whereas the Nearest model shows a rather high score ($agree_{\text{dest}} = 0.64$) when compared to the human annotators ($agree_{\text{dest}} = 0.66$). This means that the authors must have used a fixed and shared type of argumentation strategy, with components necessary for the interpretation of FRAMING Linking often found near each other.

In a further attempt to explain the relatively low FRAMING linking agreement, we performed an error analysis of the linking errors. We classified the 128 FRAMING Linking disagreements in the material, according to whether the destination spans across annotators show overlap in character position. For those 41 cases that have overlapping spans, we further examine whether the two spans share content in a meaningful manner (26 do; 15 don't). Even for those 87 spans that do not occur at the same position in the text, it is still possible that they share content, as the spans could be paraphrases of each other, so we examined these as well and found that while 65 have different meaning, 22 are actually paraphrases. If we were to consider both the 22 "reformulation" cases and the 26 "meaningful overlap" links as agreeing, the $agree_{\text{dest}}$ value would rise to 0.79. This is potentially an encouraging result as it establishes a higher upper bound on how much annotators naturally agree on FRAMING Linking.

Most errors that we categorised as "different meaning" are caused by non-agreement during the FRAMING-main identification stage. This possibly means that our definition of the FRAMING-main category in the guidelines was not yet specific enough; we will address this in the next iteration of our guidelines.

6.2.4 Discussion

It is hard to evaluate a newly defined, complex task involving the interpretation of (and judgement about) somebody else's argumentation. The annotation experiment reported here showed relatively better agreement for Rhetorical Classification, for Issue Topic Identification and for Issue Topic Linking, and relatively lower agreement for FRAMING Linking.

Table 2 Frequency of sequences of rhetorical roles in rhetorically mixed sentences (Repeated consecutive roles are mapped onto single occurrences)

Transition pattern	Freq.	Transition pattern	Freq.	Transition Pattern	Freq.
FR-s, FR-m	487	IDT, FR-s, FR-m	39	IDT, FR-m, CCL	7
FR-m, CCL	181	IDT, FR-s	10	BGD, FR-m	5
IDT, FR-m	44	FR-s, FR-m, CCL	8	FR-m, IDT	3

We now revisit the question of which annotation unit is best for Rhetorical Classification, text spans based on comma-separated units (as is the case in our existing annotation) or sentences. In the annotated material, we found that only 2.2% of all sentences (806 sentences out of 37,371[12]) contain multiple rhetorical roles. As a consequence, we decided to perform sentence-based Rhetorical Classification from now on, a decision that requires us to map annotated units onto annotated sentences.

As we are interested in argumentation patterns, it is nevertheless instructive to see which rhetorical roles are shared in those 2.2% of sentences with mixed rhetorical status. Table 2 gives the most frequent patterns along with their frequency.[13] 487 out of 806 rhetorically mixed sentences reveal the pattern "FRAMING-sub followed by FRAMING-main", and 181 rhetorically mixed sentences contain "FRAMING-main followed by CONCLUSION". In all patterns observed, a span of lower argumentation level is followed by one of a higher argumentation level. We can use this information to create a good mapping, as higher-level argumentative material tends to be overall more important for the argumentation. We therefore apply "force-right" mapping, i.e., we assign each sentence the category of its rightmost text span.

We now move to the automatic treatment of the first two tasks in our model: Issue Topic Identification and Rhetorical Classification.

7 Automatic issue topic identification

Issue Topic Identification is a crucial task as the Issue Topic is a key component of the argument structure in Japanese judgement documents. In the following experiments, we use the full corpus of 89 documents (annotated by our paid annotator).

7.1 Proposed system

We perform Issue Topic Identification as a binary classification problem on comma-separated units, using supervised machine learning with Support Vector Machines (SVM).[14] The features used for the task are the following:

[12] Issue Topics are not counted here.

[13] Our treatment of repetition means that the count for "FRAMING-main, CONCLUSION" includes occurrences of "FRAMING-main, FRAMING-main, CONCLUSION".

[14] We use Chang and Lin's (2011) implementation with a linear kernel. The SVM parameter C is set to C = 1 for all models.

Bigrams: the bag of lemmatised bigrams of morphemes in the unit, with a frequency cutoff of 20.

Unit location i/*length(D)*: the relative location of *i*-th comma-separated unit of a document *D*, where *length*(*D*) is the total number of comma-separated units *D*.

Unit length: the length of the comma-separated unit in number of characters.[15]

Keywords: keywords that are related to Issue Topic descriptions, e.g. "*souten* (Issue Topic)" and "*touhi* (propriety)". We collected 10 such keywords from two annotated judgement documents not included in the training and test data.

List marker: a binary feature that indicates whether a comma-separated unit contains a list marker (Chinese or Arabic numerals or Katakana letters, sometimes with brackets), which is a typical way to typeset Issue Topics in judgement documents. List markers are implemented as 14 regular expressions.

Nominal head: a binary feature that indicates whether the syntactic head of a comma-separated unit head is a noun or not. We use the Japanese dependency parser Cabocha (Kudo and Matsumoto 2002) to extract the head of the phrase. This feature aims to exploit the fact that many Issue Topics are stated in the form of noun phrases.

"Baseline System" refers to the system using only the first 3 features (bigram, unit location, unit length), which are very simple, while "Proposed System" refers to the system using all 6 features.

7.2 Evaluation

After performing the human annotation experiment, we had to change the definition of Issue Topics from "first or best mention" to "all mentions". We had initially expected most texts to contain only one description of each Issue Topic, but this assumption was not borne out well enough by the data. This situation is good and bad at the same time: If we accept these duplicates of Issue Topics in our annotation, we model the existing phenomena more truthfully, while increasing our chance of finding at least one version of each Issue Topic in a document, which is useful for summarisation. Also, we will end up with cleaner training data.[16]

These advantages come at a cost: re-annotation of IT duplicates is time-consuming, and we could not re-measure IAA, but we decided to take this route nevertheless. The first author reannotated first the 8 documents from the human annotation study, and subsequently the 81 documents previously annotated by the other annotator.[17] After re-annotation, the number of Issue Topics rose from 432 to 853.

Another repercussion of the changed definition of Issue Topics is that our final system now requires a method for detecting duplicates, in order to avoid summaries

[15] Average unit length is 18.5 characters, median is 14, standard deviation is 18.1.

[16] In any supervised machine learning environment, it is undesirable if entities sharing many surface features are labelled with conflicting target categories. In our case, the non-annotated duplicates would incorrectly act as negative examples for first-mention Issue Topics.

[17] The additional annotation was supported by a high-recall, low-precision automatic search, which aimed to find units sharing a high number of bigrams with units already annotated as Issue Topics.

Table 3 Issue topic identification: results (by comma-separated units)

	Precision		Recall		F-value	
	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed
Issue Topic	0.75	0.76	0.38	0.40	0.50	0.52

with multiple identical Issue Topic descriptions. Leaving this issue aside, we will treat here as correct the retrieval of all Issue Topics and their duplicates.

7.3 Results

Table 3 shows the results of the experiments, using fivefold cross-validation on all 89 documents, given in precision, recall and F-value. The proposed SVM model (F = 0.52) significantly improves over the baseline model (F = 0.50), as tested using the two-tailed McNemar test (McNemar 1947) with significance level $\alpha = 0.05$. We note that the overall F-value is affected negatively by the low recall, whereas precision is high. The difference between Baseline and Proposed system shows that the combination of the features "Keywords", "List Markers" and "Nominal Head" aid in the identification of Issue Topics. List markers, keywords and nominal head as features seem to perform as intended, but we plan to further fine-tune and supplement them with more informative features in the future.

8 Automatic rhetorical classification

8.1 Proposed systems

Mentions of Issue Topics are rare and generally occur without much dependence on the surrounding text. In contrast, Rhetorical Classification should be subject to a strong effect of the context in terms of other rhetorical roles, as the frequent sequences in Table 2 indicate. Rhetorical context effects are also predicted by our proposed argument structure and have been successfully exploited in previous work for the same task (Saravanan and Ravindran 2010). We therefore use Conditional Random Fields (CRF) (Lafferty et al. 2001),[18] in addition to an SVM system as before, which treats the task as a standard multi-class classification problem.

As features, we use variations of the three simple features developed for Issue Topic Identification. The bigram feature and unit location feature (redefined as sentence location feature) remain the same for the SVM model, but in the case of the CRF model, which does not allow for continuous values, the sentence locations are bucketed into 10 percentiles (0–10% ...90–100%). We also use a sentence length feature (calculated in characters).

[18] We used Okazaki's (2007) implementation.

Cue phrases have been found to be useful in previous legal summarisation work (Hachey and Grover 2006). We developed the following variants of cue phrase features, which are adapted to the language and law system we treat.

Modality expressions: we expect a strong connection between modality information and the rhetorical role of a target unit. For example, FRAMING-main sentences, which typically state a judge's interpretation or evaluation of facts, frequently contain so-called "truth judgement modality" (e.g, "*hazu da*" (can be expected to be) or "*beki da*" (should be). We use 8 modality features based on Masuoka's (2007) modality expression classification, namely the modalities "truth judgement" (4 features), "value judgement" (3 features), and "explanation"(1 feature).

Function expressions: Function expressions such as postpositional particles, auxiliary verbs and a special class of multi-word units give clues about the rhetorical status of a statement in the Japanese language. CONCLUSION sentences, for instance, should contain decision function expressions (such as "shall"), whereas factual sentences would normally not contain conjecture function expressions (such as "might"). We recognise the 16,801 surface expressions in the function expression dictionary by Matsuyoshi et al. (2007) and use as features the 199 semantic equivalence classes associated with them (e.g., "evidential" and "contradictory conjunction").

Cue phrases: We extracted an additional 22 phrases from a textbook used during the training of judges (Judicial Research and Training Institute of Japan 2006), and from five judgement documents not included in the training and test data.

Law names: We expect the mention of a law name to be a signal for BACKGROUND sentences, which state precedent information or give reasons related to laws. We therefore used a binary feature that indicates presence of *any* law name in the sentence. In addition, we used a list of 494 specific law names as features.

As before, the three simple features are used in the "Baseline Systems", and the classifiers using all features are referred to as the "Proposed Systems".

8.2 Evaluation

We assume in these experiments that Issue Topics have been identified by an earlier recognition stage, and will therefore exclude sentences which contain more or one units annotated as Issue Topics (there are 302 of these), both in training and testing. After performing force-right mapping as described above,[19] the distribution shown in Table 4 emerges. We can see that the proportion of OTHER is quite high at almost 40%, and that the next frequent categories are FACT and FRAMING-main at 23% and 20% each.

[19] An alternative mapping method exists: training and classification could be performed directly in comma-separated units, with results subsequently mapped to sentences. As this method resulted in inferior results, we are not reporting it further.

Table 4 Rhetorical category distribution (% out of 37,371 sentences in total)

FACT	FRAM-main	FRAM-sub	CONCL.	IDENTIF.	BACKGR.	OTHER
23.1%	19.5%	11.5%	3.9%	2.1%	0.3%	39.7%

Table 5 Rhetorical classification results (F-value; sentence-based)

Category		SVM		CRF	
		Baseline	Proposed	Baseline	Proposed
FACT	(FCT)	0.77	0.78	0.84	0.84
FRAMING-main	(FR-m)	0.62	0.61	0.59	0.60
FRAMING-sub	(FR-s)	0.28	0.32	0.49	0.49
CONCLUSION	(CCL)	0.29	0.31	0.37	0.39
IDENTIFYING	(IDT)	0.79	0.77	0.79	0.79
BACKGROUND	(BGD)	0.19	0.20	0.28	0.32
OTHER	(OTR)	0.95	0.97	0.97	0.97
Macro avg.		0.56	0.57^{+}	0.62^{+*}	0.63^{+*}

+, Means significant difference from the SVM baseline model. *, means significant difference from the SVM proposed model

8.3 Result

Across the board, the CRF performs better than the SVM, as can be seen from Table 5. In particular, the proposed CRF model (F = 0.63) significantly outperforms the proposed SVM model (F = 0.57), as well as the baseline SVM model (F = 0.56), demonstrating the importance of the rhetorical context for this task. However, the additional features modality, function expressions and cue phrases, taken together, have a limited impact on both classifiers [insignificant for the CRF at F = 0.63 vs F = 0.62, and significant but small for the SVM (F = 0.57 vs F = 0.56)]. This is somewhat disappointing, as the features were carefully constructed and represent both linguistic and legal knowledge.

We performed an ablation experiment on the Proposed CRF System to shed some light on which features contribute overall and for the recognition of specific categories; Table 6 shows both single-feature ablation (top) and leave-one-out ablation (bottom). The bigram feature is strongly dominant. This is shown in the single-feature ablation (top table, where high values mean strong features)[20]: the bigram feature on its own reaches a micro-averaged F = 0.62 on the overall task, whereas the next-best performing feature, functional expressions, only performs at F = 0.43, with cue phrases next at at F = 0.30. The least distinctive feature on its own is law names (F = 0.07). Bigram dominance is confirmed in a leave-one-out ablation study

[20] In the top part of the table, boldfacing shows the three best feature at identifying a category, unless they are too weak. In the bottom part of the table, boldfacing shows decrease when feature is left out.

Table 6 CRF model: category and feature ablation (in F-value)

Cat/feature	Bigram	Sent. len	Sent. pos	Mod	Func exp	Cue	Law	ALL
Single-feature Ablation								
FCT	**0.83**	0.49	0.40	0.44	**0.76**	**0.51**	0.38	0.84
FR-m	**0.59**	0.18	0.00	0.11	**0.53**	**0.33**	0.00	0.60
FR-s	**0.47**	0.03	0.17	**0.29**	**0.45**	0.28	0.05	0.49
CCL	**0.37**	0.00	0.00	0.00	0.02	**0.12**	0.00	0.39
IDT	**0.79**	0.00	0.00	0.00	0.08	**0.44**	0.00	0.79
BGD	**0.31**	0.00	0.00	0.00	0.02	**0.12**	0.00	0.32
OTR	**0.97**	**0.91**	0.37	0.08	**0.96**	0.29	0.03	0.97
Macro avg.	0.62	0.23	0.13	0.13	0.43	0.30	0.07	0.63
Leave-one-out Ablation								
FCT	**0.81**	0.84	0.84	0.84	0.84	0.84	0.84	0.84
FR-m	**0.57**	0.60	**0.59**	0.60	0.60	0.60	0.60	0.60
FR-s	0.50	0.49	**0.48**	0.49	0.49	0.49	**0.48**	0.49
CCL	**0.25**	**0.38**	**0.38**	0.39	**0.38**	**0.38**	0.39	0.39
IDT	**0.51**	0.79	0.79	0.79	0.80	0.79	0.79	0.79
BGD	**0.24**	0.32	0.36	**0.31**	0.33	0.33	0.34	0.32
OTR	0.96	**0.63**	0.97	0.97	0.97	0.97	0.97	0.97
Macro avg.	**0.55**	0.63	0.63	0.63	0.63	0.63	0.63	0.63

(bottom of the table, where low values indicate strong features): here, the bigram feature is the only feature that—if left out—decreases overall results, and also the single most helpful feature in the recognition of every feature except OTHER [where it is helped by the sentence length feature (F = 0.63)] and FRAMING-sub [where it is helped by the sentence position (0.48) and law name features (0.48)].

However, the total macro-F metric disguises some of the differences that matter to us: when compared to the full system, bigrams on their own are inferior at identifying every category except OTHER and IDENTIFYING. The effect of the additional features is thus to support and reinforce the bigram feature. Cue phrases are the strongest features for doing so. Their positive effect is visible particularly for BACKGROUND, CONCLUSION and IDENTIFYING (12%, 12% and 44%; top), categories where no other non-bigram feature can offer any help (except the modality feature for BACKGROUND). Categories benefiting from the function expression feature are FACT, FRAMING-main and FRAMING-sub (according to the single-feature study; top), and CONCLUSION (according to the leave-one-out ablation; bottom). The law name feature is useful for FRAMING-sub, but against expectations not for BACKGROUND. (With hindsight, we think that this may be due to the fact that repeated mentions of the law often happen in abbreviated form.)

We feel that the argumentation-based features are worth the effort overall for other reasons too. The CONCLUSION and FRAMING-sub categories are relatively sparse (as opposed to FACT and FRAMING-main), and it is well-known that purely statistical features such as the bigram feature suffer in the face of data sparseness. Because these rare categories are essential for summarisation, we appreciate the

robustness that a mix of symbolic features can offer. One of our future avenues is to acquire cue phrases and modality expressions in a robust data-driven manner.

Considering the ablation results with respect to categories, the categories which can be better distinguished than others are FACT (F = 0.84) and FRAMING-main (F = 0.60), while CONCLUSION (F = 0.39) and BACKGROUND (F = 0.32) show low performance. This can be partially explained by the low number of instances for these categories. FRAMING-sub (F = 0.49) performs worse than FRAMING-main (F = 0.60), but this time the number of instances cannot be responsible for the effect, as FRAMING-sub is relatively frequent.

Confusion between FRAMING-sub and FRAMING-main is a common theme, which we already observed during human annotation. In the summarisation stage, particularly when composing our Type B2 design summary, the confusion between FRAMING-main and FRAMING-sub will cause problems. One solution is to train a separate classifier whose task is only the distinction of FRAMING-main and FRAMING-sub sentences. Such a classifier could exploit the fact that FRAMING-main and FRAMING-sub have differences along several dimensions: FRAMING-main tend to be more general and abstract, FRAMING-sub more specific and concrete. In previous work, distinctions such as general-specific and abstract-concrete have been successfully learned from corpora, e.g. Turney et al. (2011).

9 Related work

9.1 Previous work: rhetorical structure recognition

Rhetorical Classification was originally defined for scientific articles by Teufel and Moens (2002), and later ported to the legal text domain by Hachey and Grover (2006). The sentence is chosen as the annotation unit in both works. We defined our categories based on Hachey and Grover's six categories, which are specialised to English law. Our main change, the split of their FRAMING category into FRAMING-main and FRAMING-sub, is motivated by our summary design, as it allows us to distinguish between levels 2 and 3 in the argumentative structure of judgement documents. Without the split, the argumentative text under FRAMING would cover to too much material (with nominally the same level of importance), which runs counter to the purpose of summarisation. The distinction between FRAMING-main and FRAMING-sub, despite its inherent difficulty, is therefore central to our task.

Other changes we made to Hachey and Grover's scheme are due to differences in legal systems. Hachey and Grover treat UK House of Lords[21] cases, which are by nature appeal cases. In this context their category PROCEEDINGS gave details of previous judgements in lower courts (not used by us). We also removed their category TEXTUAL, which was reserved for meta-statements about section structure.

[21] The UK House of Lords acted as the final court of appeal in the United Kingdom judicial system until the establishment of the Supreme Court in 2009.

Their category DISPOSAL (used for judges' conclusions) is similar to our CONCLUSION, but in our case this category is reserved exclusively for the conclusion of each Issue Topic. Their other two categories FACT and BACKGROUND were taken over by us as-is.

Hachey and Grover's inter-annotator agreement was K = 0.83 ($N = 1,955$, $n = 7$, $k = 2$; Cohen); you may recall that this is incidentally the same agreement as the one we reach when we merge FRAMING-main and FRAMING-sub. Hachey and Grover also tested various supervised machine learning systems trained on the human-annotated material. The best results were achieved using the classifier C4.5 (Quinlan 1993) with only the location feature (F = 0.65); the second best (F = 0.61) was achieved using an SVM with all features (location, thematic words, sentence length, quotation, entities and cue phrases). Our numerical results for RC compare favourably to these.

Saravanan and Ravindran (2010) follow the same approach, adapted to the Indian law system, and reach an inter-annotator agreement of K = 0.84 ($N = 16{,}000$; $n = 7$, $k = 2$). Using various features similar to ours a CRF classifier, their automatic results reach F = 0.82, but like Hachey and Grover's, their annotation scheme also does not make a distinction similarly difficult to our FRAMING-main versus FRAMING-sub distinction.

Compared to all previous work in Rhetorical Classification (Grover et al. 2004; Mochales and Moens 2011; Saravanan and Ravindran 2010), our model additionally offers hierarchical structuring of the argument, in the form of "support"-style links and Item Topic Identification and Linking.

9.2 Previous work: AI and law

There is a long-standing tradition in the area of AI and Laws to represent and reason about legal arguments or factors. The components of the arguments are typically represented as logical propositions. For instance, Ashley and Brüninghaus (2009) developed a system which combines case-based reasoning with information extraction from legal texts. The system automatically classifies textual descriptions of the facts of legal problems. Given a database of previously classified cases, the system can use these classification decisions to provide an evaluation and explanation about predictions about a case's outcome. The legal reasoning system achieved accuracy = 0.92 for predicting results, but the extraction stage performed only at F-value = 0.26. Satoh et al. (2011) developed a legal reasoning system called PROLEG (short for "Prolog-based Legal reasoning support system") for Japanese Civil Law cases. The system simulates JUF (Ito 2008), a reasoning strategy used by judges for decision making in civil law cases. PROLEG's simulation of JUF however requires that first the propositional content of the argument is manually extracted from natural language text.

It is well-acknowledged that there is a knowledge acquisition gap for these kind of reasoning systems: full automation is still beyond the possibility of current NLP techniques. Recent developments in closer collaboration between the argument mining community an the AI and Law community may change this in the future.

Walker et al. (2017) created a corpus of decisions adjudicating claims by US military veterans for disability compensation. By a careful analysis of the legal argumentation, they developed a representation of the legal argument which is designed for the extraction and representation of information about legal rules. Their representation consists of two parts, a set of propositional connectives such as "jointly sufficient set of necessary conditions" or "merely relevant condition". The second part concerns what they call the "semantic type" of sentences from legal texts, which roughly corresponds to our rhetorical categories. Walker et al.'s work offers several sub-divisions for FRAMING-type material, for example "policy-based-reasoning sentence or clause", "rule-based-reasoning sentence or clause" and "evidence-based-reasoning sentence or clause". Their representation allows them to express complex legal rules, but these rules are expressed simply in natural language. Automation is not envisaged.

Our work is in the spirit of these works, as it aims at modelling the underlying reasoning, but being in the tradition of NLP and argument mining, our design starts from the aim of full automatability, at the cost of being less ambitious with respect to the depth of argument treatable this way.

9.3 Previous work: argument mining

The area of argument mining is a recent research topic in natural language processing, where arguments in natural language texts are automatically analysed. Argument mining tasks generally consist of the following individual tasks: *argumentative component identification* (the task of separating argumentative text pieces from non-argumentative text pieces), *argumentative component classification* (the task of identifying the types of argumentative text pieces, for example, premises or claims) and *argumentative structure extraction* (the task of detecting argumentative links between argumentative components) (Stab and Gurevych 2017).

Some argument mining work which is specialised to legal text exists. Mochales and Moens (2011) presented an argumentation component identification and classification algorithm on legal text using machine learning techniques. In a human annotation study for argumentative component detection (into "premise" and "claim") on documents from the European Court of Human Rights (ECHR), they measured an inter-annotator agreement of $K = 0.75$ (Cohen). Their SVM model with location, length, cue phrases, articles, and tense features achieved $F = 0.71$ for claims and $F = 0.68$ for premises. They also conducted argumentation structure extraction with a manually-created context-free grammar.

Stab and Gurevych (2014a) studied argumentation in essays, using three argumentative components (Major Claim, Claim, Premise) and two relations ("support" and "attack"). They reported inter-annotator agreement of $K = 0.81$ (Fleiss's Kappa) with agreement ranging from $K = 0.83$ (Major claim) to $K = 0.66$ for Claim. Stab and Gurevych (2014b) perform automatic argumentative component classification using structural, lexical, syntactic, indicator and contextual features with an SVM, reaching $F = 0.77$. Full argumentative structure extraction is achieved in Stab and Gurevych (2017). They use an SVM with various features (structural, lexical,

syntactic, cue phrases, discourse, point-wise mutual information representing the token association between destination and source components, and shared noun between destination and source), reaching F = 0.72. The results were raised to F = 0.75 by further constraining the SVM features with an integer linear programming model (ILP).

To summarise, several existing works in the field of argument mining identify and extract argument components and links from text using various methods, some of these from legal text. Like these studies, we also perform component detection and link detection, but we treat several levels of argumentation rather than only individual "support" relationships; we cover the entire text with an analysis; and we combine argument mining analysis with Rhetorical Classification. We therefore model argument structure at a far more detailed level than earlier work. What is additionally novel in our work is that we draw a connection between argument mining and legal argumentation-based summarisation.

10 Conclusion and future work

We developed a novel annotation scheme for the annotation of the argumentative structure of Japanese judgement documents, along with an annotated corpus.[22] An important concept in the argument structure is the Issue Topic, a contentious points of the law suit. Our working hypothesis is that an Issue Topic-based argumentation structure will lead to better, more informative summaries, because most legal practitioners require information about individual Issue Topic when they perform their research.

We proposed an 4-task annotation scheme that enables us to capture the argumentative support relationship between text spans. It integrates classic Rhetorical Classification with relation-based argument mining tasks and with the new Issue Topic concept. The rhetorical status of comma-separated units plays an important role in our scheme because the argumentative "support" relationship that holds between the levels in our scheme is often realised by specialised rhetorical roles. We measured inter-annotator agreement for Issue Topic Identification at $agree_{ITI} = 0.79$, Rhetorical Classification at $K = 0.70$, Issue Topic Linking at $agree_{ITL} = 0.87$ and FRAMING Linking at $agree_{fl} = 0.44$.

One of the biggest theoretical problems we encountered was low distinguishability of FRAMING-main and FRAMING-sub categories with our current guidelines. The FRAMING categories could be sub-categorised according to argumentation types as proposed in the community of AI and Law community, for instance Walker et al.'s (2017) semantic types. We plan to rework our guidelines along those lines.

Concerning the automation of the annotation task, we conducted proof-of-concept experiments for Issue Topic Identification and Rhetorical Classification. Our SVM model for Issue Topic Identification outperformed a baseline model of bigram,

[22] Both corpus and guidelines will be made publicly available.

unit location and unit length features at F = 0.52, by using the keyword, list marker and nominal head features. For the task of Rhetorical Classification, our CRF model uses features such as bigrams, functional expressions, modality of Japanese sentences and law names, sentence location and sentence length features and achieves F = 0.63. In the future, we plan to use features more closely related to the structure of documents such as semantic similarity of a sentence with its neighbours. As for Issue Topic Identification, our next goal is the detection of duplicates (paraphrases) amongst the Issue Topic text spans.

The tasks thus far treated form the early stages of a pipelined model for the implementation of the entire automation. To make the entire system operational, we will next automate the later tasks of Issue Topic Linking and FRAMING Linking. Issue Topic Linking might profit from a topic modelling approach, using LDA (Blei et al. 2003) to define bottom-up distributions of concepts. FRAMING Linking is closely related to the extraction of supportive relations from argumentative text, although our linking defines relatively more fine-grained relations. Recent studies in the argumentation mining community apply deep learning architectures to the relation finding task (Cocarascu and Toni 2017; Potash et al. 2017), and we will investigating the potential of such methods for our data and task.

Another pressing task is the generation of the modularised summaries based on the summary designs we proposed. The main challenge with generating such a summary is keeping the output coherent. We will start with Type A summaries, which are relatively easy to build, and which don't have strong requirements on coherence, and will extend the summariser step by step towards full Issue Topic-driven summarisation.

Acknowledgement This work was partly supported by Tokyo Tech World Research Hub Initiative (WRHI) Program of Institute of Innovative Research, Tokyo Institute of Technology and the open collaborative research at National Institute of Informatics (NII) Japan (FY2018).

References

Ashley KD, Brüninghaus S (2009) Automatically classifying case texts and predicting outcomes. Artif Intell Law 17(2):125–165

Banno S, Matsubara S, Yoshikawa M (2006) Identification of Important parts in judgments based on Machine Learning. In: Proceedings of the 12th annual meeting of the association for natural language processing, the association for natural language processing, pp 1075–1078

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Carletta J (1996) Assessing agreement on classification tasks: the kappa statistic. Comput Linguist 22(2):249–254

Chang CC, Lin CJ (2011) Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):27

Cocarascu O, Toni F (2017) Identifying attack and support argumentative relations using deep learning. In: Proceedings of the 2017 conference on empirical methods in natural language processing, association for computational linguistics, pp 1374–1379

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46

Grover C, Hachey B, Hughson I (2004) The HOLJ Corpus. Supporting summarisation of legal texts. In: COLING 2004 5th international workshop on linguistically interpreted corpora, pp 47–54

Hachey B, Grover C (2006) Extractive summarisation of legal texts. Artif Intell Law 14(4):305–345

Ito S (2008) Lecture series on ultimate facts. Shojihomu **(in Japanese)**

Japanese Ministry of Justice (2012) Japanese Code of Civil Procedure Subsection 2 Preparatory Proceedings

Judicial Research and Training Institute of Japan (2006) The guide to write civil judgements, 10th edn. Housou-kai **(in Japanese)**

Kaplan D, Iida R, Nishina K, Tokunaga T (2011) Slate—a tool for creating and maintaining annotated corpora. J Lang Technol Comput Linguist 26(Section 2):91–103

Krippendorff K (2004) Content analysis: an introduction to its methodology (2nd edn). SAGE Publications, Thousand Oaks, CA

Kudo T, Matsumoto Y (2002) Japanese dependency analysis using cascaded chunking. In: CoNLL 2002: proceedings of the 6th conference on natural language learning 2002 (COLING 2002 Post-Conference Workshops), pp 63–69

Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '01, pp 282–289

Maekawa K, Yamazaki M, Ogiso T, Maruyama T, Ogura H, Kashino W, Koiso H, Yamaguchi M, Tanaka M, Den Y (2014) Balanced corpus of contemporary written Japanese. Lang Resources Eval 48(2):345–371

Masuoka T (2007) Nihongo Modariti Tankyu (Japanese Modality Investigations). Kuroshio shuppan

Matsuyoshi S, Sati S, Utsuro T (2007) A dictionary of Japanese functional expressions with hierarchical organization. J Nat Lang Process 14(5):123–146

McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12(2):153–157

Mochales R, Moens MF (2011) Argumentation mining. Artif Intell Law 19(1):1–22

Okazaki N (2007) Crfsuite: a fast implementation of conditional random fields (crfs). http://www.chokkan.org/software/crfsuite/

Potash P, Romanov A, Rumshisky A (2017) Here's my point: joint pointer architecture for argument mining. In: Proceedings of the 2017 conference on empirical methods in natural language processing, association for computational linguistics, pp 1364–1373

Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco

Saravanan M, Ravindran B (2010) Identification of rhetorical roles for segmentation and summarization of a legal judgment. Artif Intell Law 18(1):45–76

Satoh K, Asai K, Kogawa T, Kubota M, Nakamura M, Nishigai Y, Shirakawa K, Takano C (2011) Proleg: an implementation of the presupposed ultimate fact theory of Japanese civil code by prolog technology. In: Onada T, Bekki D, McCready E (eds) New frontiers in artificial intelligence. Springer, Berlin, pp 153–164

Stab C, Gurevych I (2014a) Annotating argument components and relations in persuasive essays. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics, pp 1501–1510

Stab C, Gurevych I (2014b) Identifying argumentative discourse structures in persuasive essays. In Proceedings of the 2014 conference on empirical methods in natural language processing, association for computational linguistics, pp 46–56

Stab C, Gurevych I (2017) Parsing argumentation structures in persuasive essays. Comput Linguist 43(3):619–659

Teufel S, Moens M (2002) Summarizing scientific articles: experiments with relevance and rhetorical status. Comput Linguist 28(4):409–445

The Secretariat of Supreme Court of Japan (1990) The new format of civil judgment : 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. Housou-kai

Turney PD, Neuman Y, Assaf D, Cohen Y (2011) Literal and metaphorical sense identification through concrete and abstract context. In: Proceedings of the conference on empirical methods in natural language processing, association for computational linguistics, Stroudsburg, PA, USA, EMNLP '11, pp 680–690

Walker VR, Han JH, Ni X, Yoseda K (2017) Semantic types for computational legal reasoning. In: Proceedings of the 16th edition of the international conference on artificial intelligence and law—ICAIL '17, vol 17, pp 217–226