Tokyo Tech

https://www.cl.c.titech.ac.jp

Argument-based summary designs for Japanese judgment documents

Tokyo tech Tokunaga-lab

Hiroaki Yamada, Simone Teufel, Takenobu Tokunaga

https://h-yamada.jp
yamada.h.ax@m.titech.ac.jp

Goal of the project

- Summarization of Judgment documents
  - Information source: **judgment documents**
    - Long, complicated
    - Interleaved arguments
  - Target audience: people who wants to search past legal cases
    - Their task: analyze legal cases and utilize the information for their own cases/matters.
    - Information overload
  - Previous studies
    - Extractive summarization with Rhetorical status analysis approach
    - Hachey and Grover, 2006, Saravanan and Ravindran, 2010
      - Most sentences (or clauses,) are classified into “Argumentative” category in our case and it was not so helpful in constructing summary.
      - We need the relation and roles among those argumentative sentences.
  - Our idea: **Argument** focused summarization
    - Argumenta extraction is required!
    - Application is not limited to summarization
      - Highlighting the important sentences in the documents
      - Enhance precedents search engines

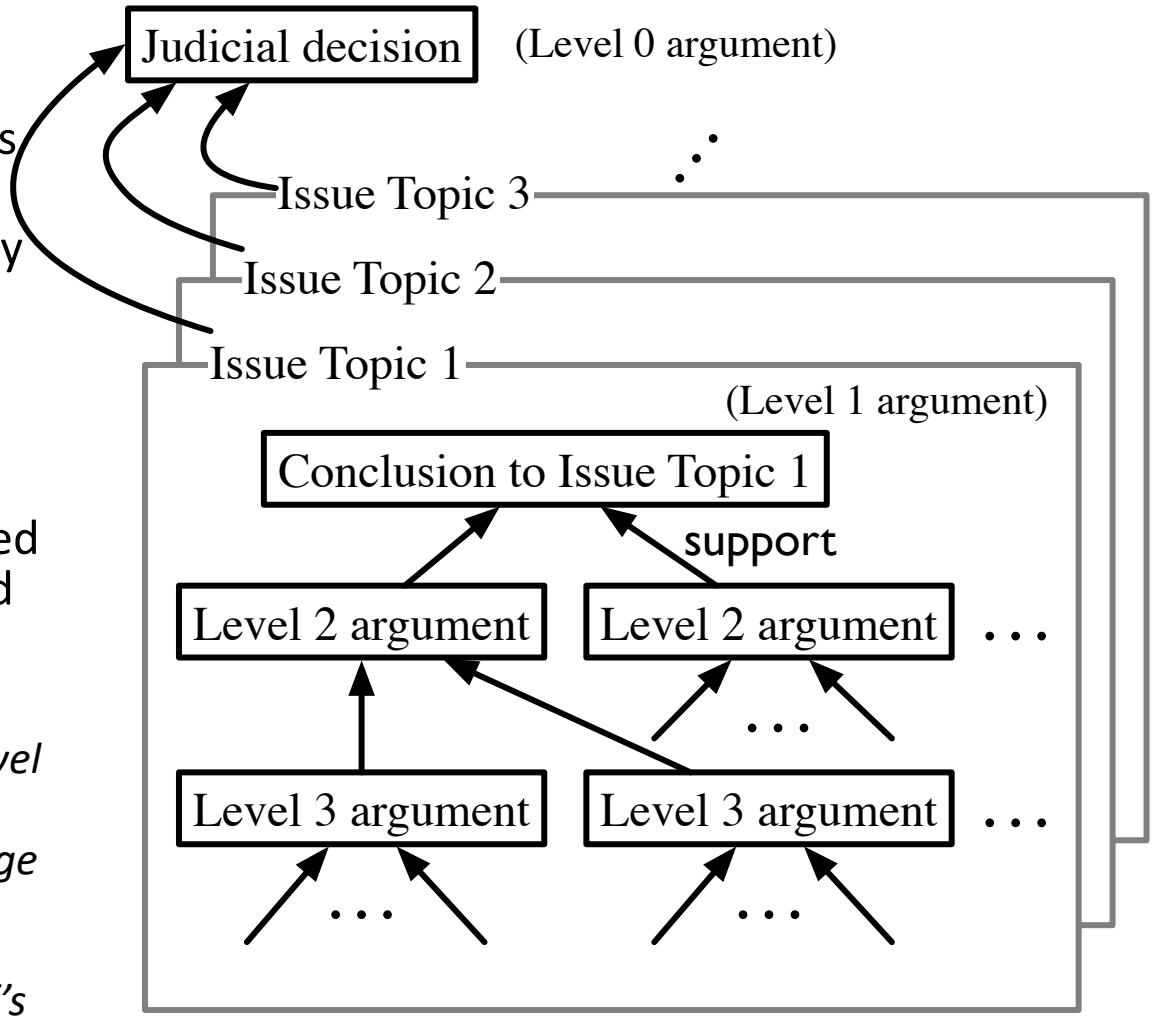
Arguments in Japanese Judgment documents

- Structure
  - Hierarchical structure which is Issue Topic centered.
  - One argument is supported by its sub arguments (e.g. Lv.2 is supported by Lv.3).
- Issue Topic:
  - Issue Topics are the main contentious items to be argued about between the interested parties.
  - Example:

    ***Case**: Road safety in a bus travel sub-contract situation*

    ***Issue Topic 1**: Details of damage incurred by plaintiff*

    ***Issue Topic 2**: Comparative negligence [degree of plaintiff's own negligence]*

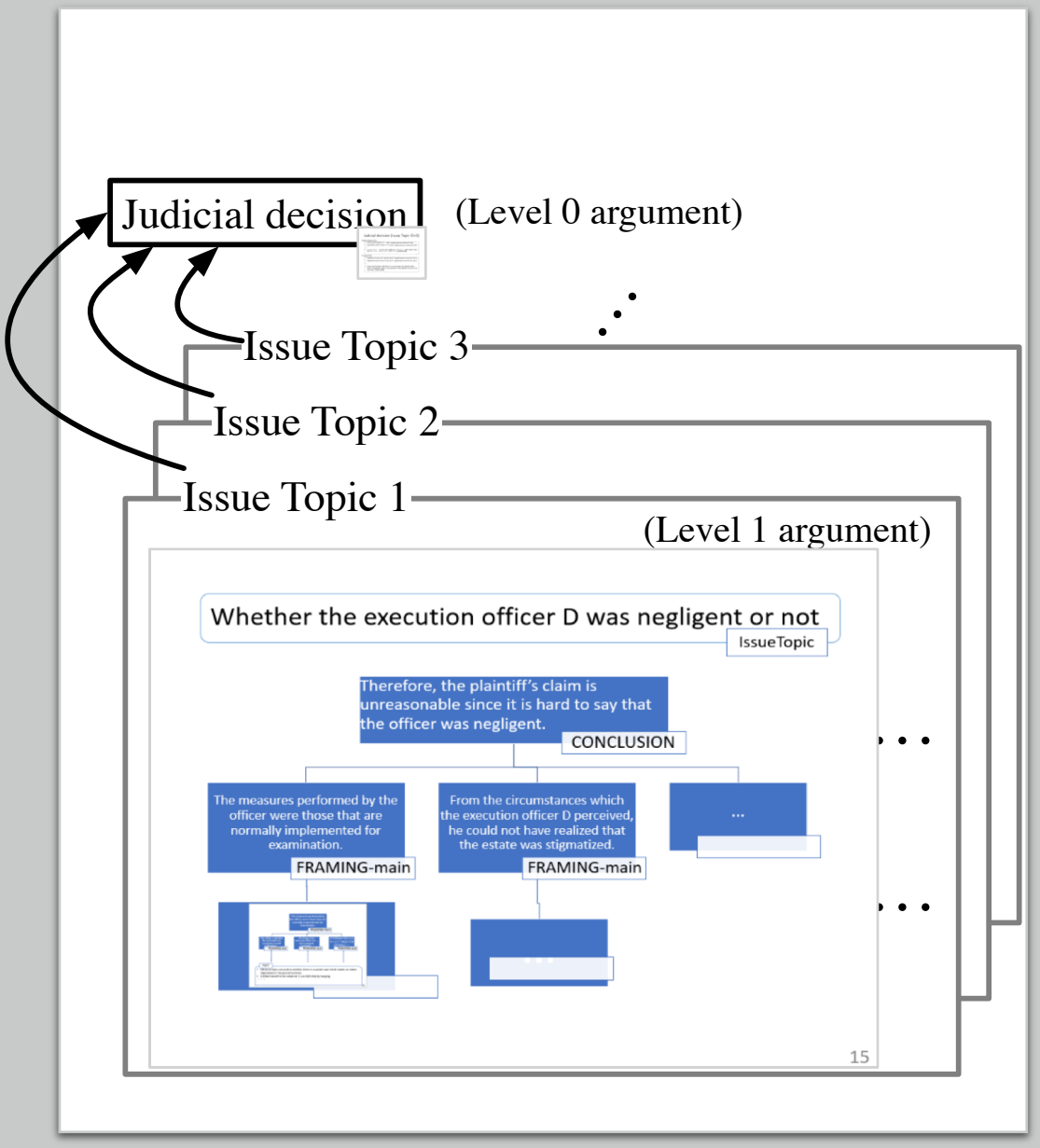


Example

http://www.courts.go.jp/app/files/hanrei_jp/301/037301_hanrei.pdf

Case: Estate compensation

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.



Judicial decision (Issue Topic ID=0)

Original Japanese Text

- 1原告の請求を棄却する。[from “judgment(main sentences)” part]
- 2訴訟費用は原告の負担とする [from “judgment(main sentences)” part]

….

- 以上のとおり，その余の点を判断するまでもなく，原告の請求は理由がないから，主文のとおり判決する。[CONCLUSION]

Translated Text

- 1 Plaintiff’s claims are rejected. [from “judgment(main sentences)” part]
- 2 Plaintiff bears all court costs. [from “judgment(main sentences)” part]

….

- 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. [CONCLUSION]

Whether the execution officer D was negligent or not

IssueTopic

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

CONCLUSION

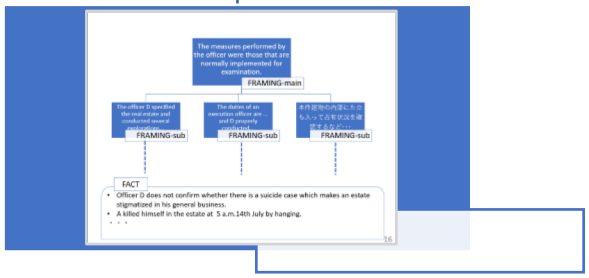
The measures performed by the officer were those that are normally implemented for examination.

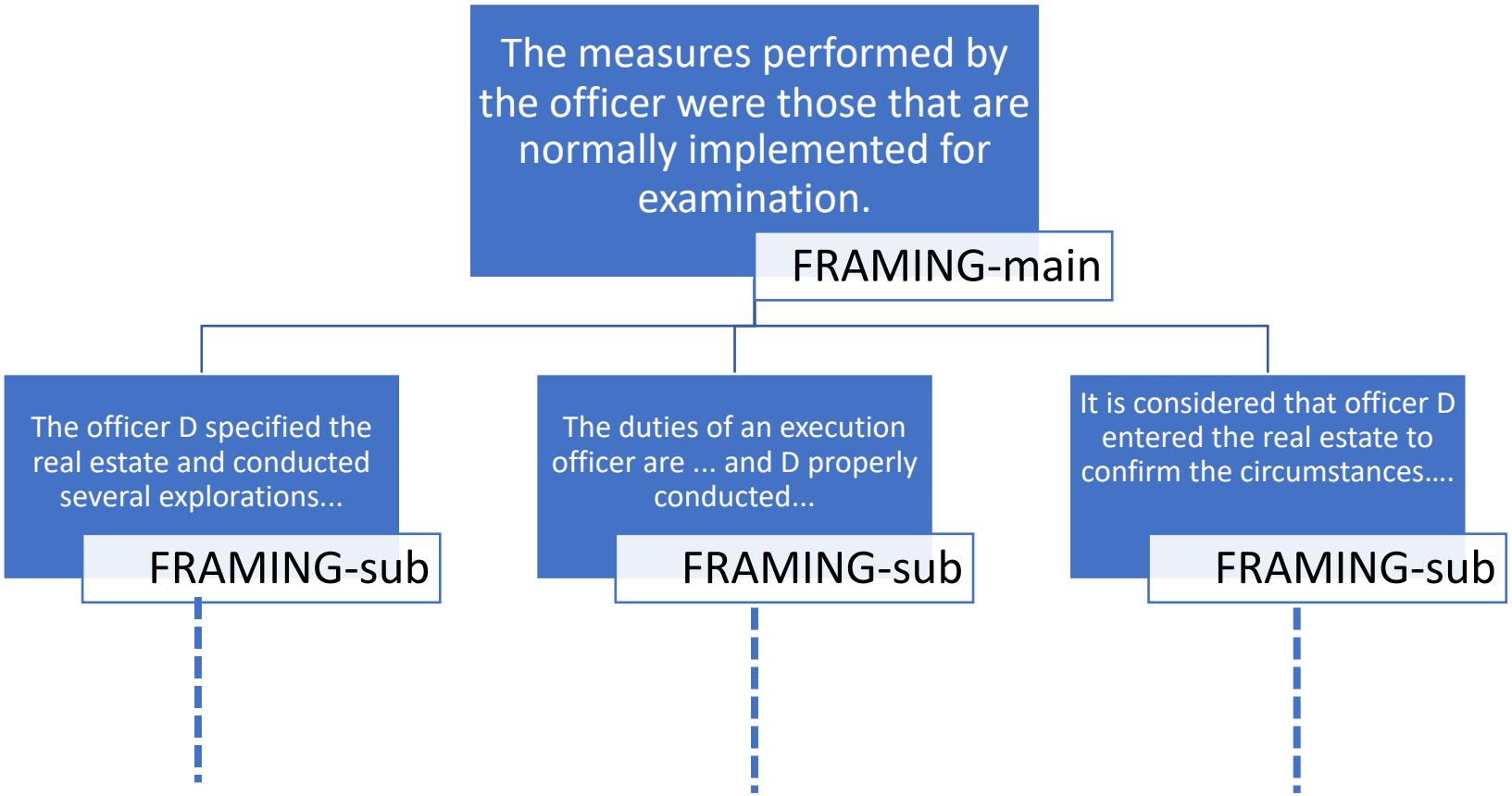
FRAMING-main

From the circumstances which the execution officer D perceived, he could not have realized that the estate was stigmatized.

FRAMING-main

…

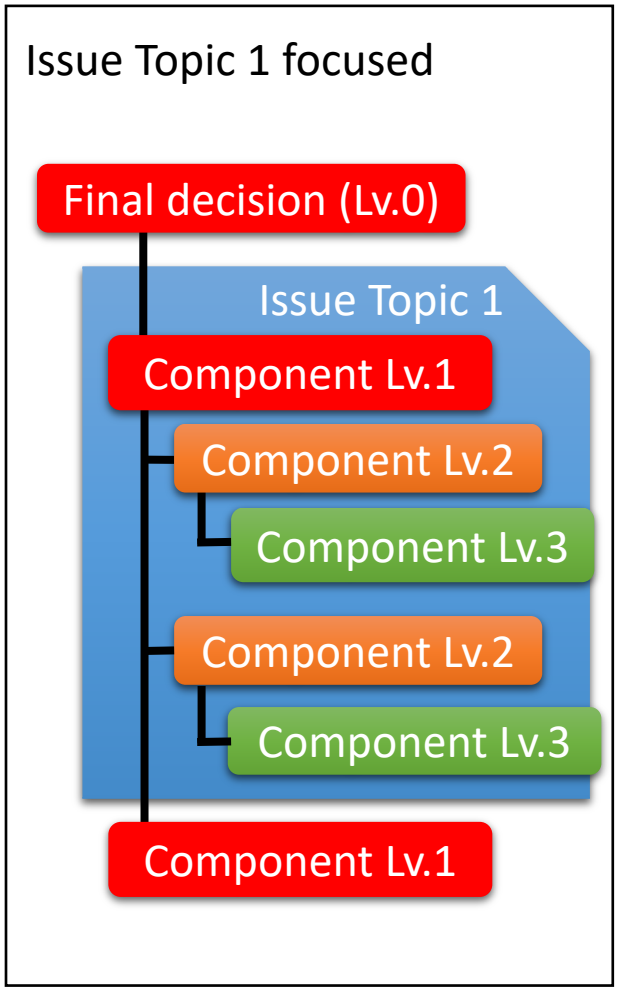




FACT

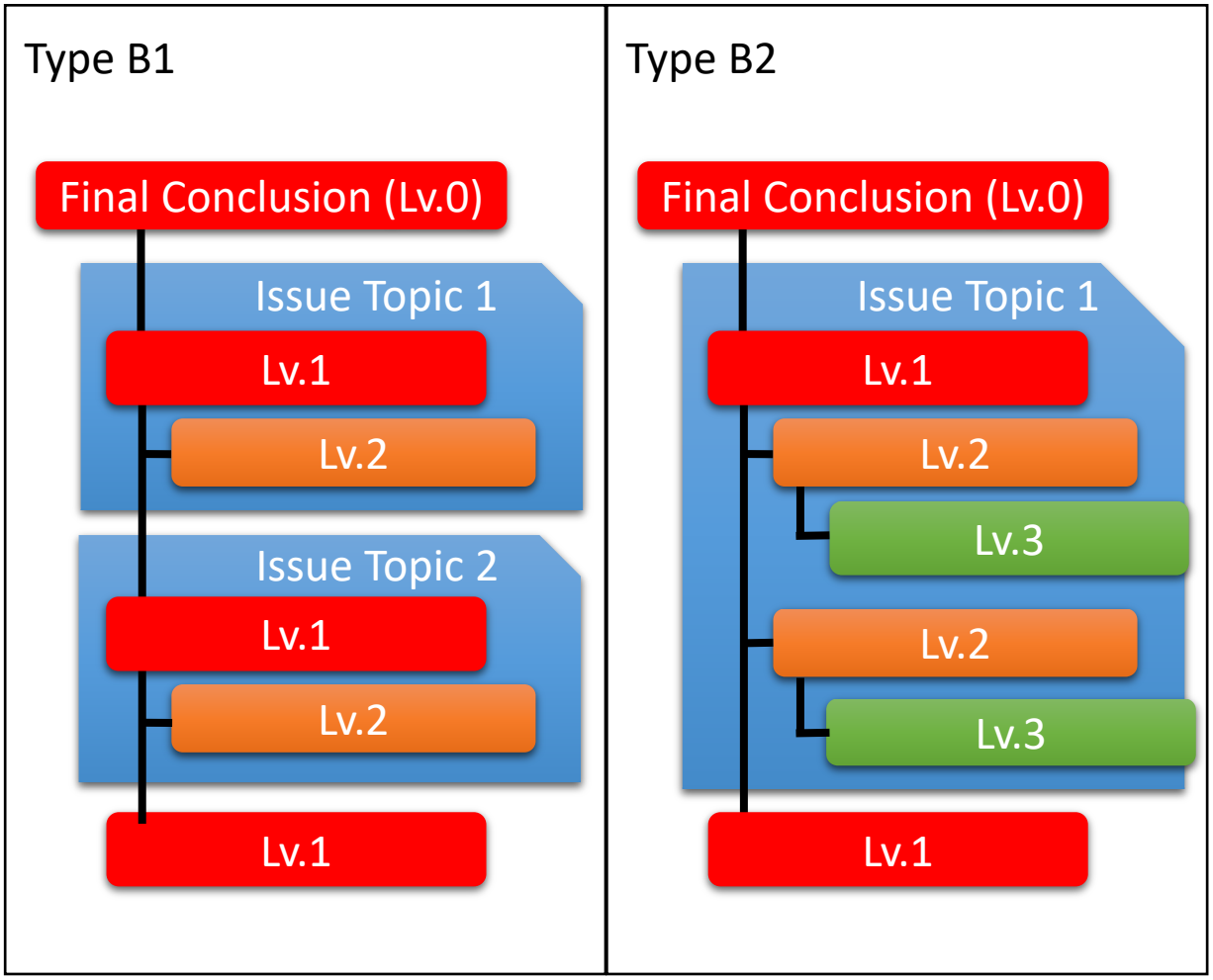
- Officer D does not confirm whether there is a suicide case which makes an estate stigmatized in his general business.
- A killed himself in the estate at 5 a.m.14th July by hanging.
- . . .

Argument-based summary



- Summary has structure
  - It presents each Issue Topic **separately**
    - One argument per Issue Topic
  - It shows **support** relationships
  - Focus on one Issue Topic
    - Selecting relevant components to the focused Issue Topic
  - More argument levels are covered

Our summary designs



Type B1:

- This type allows readers to grab the entire flows of arguments (main claims and its reasons).
- Use case / target user: people who want quickly know what is the points of the case. Ex: Civil juror, students, journalist,

Type B2:

- This type gives readers deeper arguments that are discussed in the particular Issue Topic.
- Use case / target use : People who are finding documents which let them acquire the strategy to win their own cases. Ex. Lawyers

Hiroaki Yamada, Simone Teufel and Takenobu Tokunaga. 2017. Designing an annotation scheme for summarizing Japanese judgment documents. In Proceedings of the 9th International Conference on Knowledge and Systems Engineering (KSE 2017). pages 275-280.

Type B2 summary sample

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.

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.

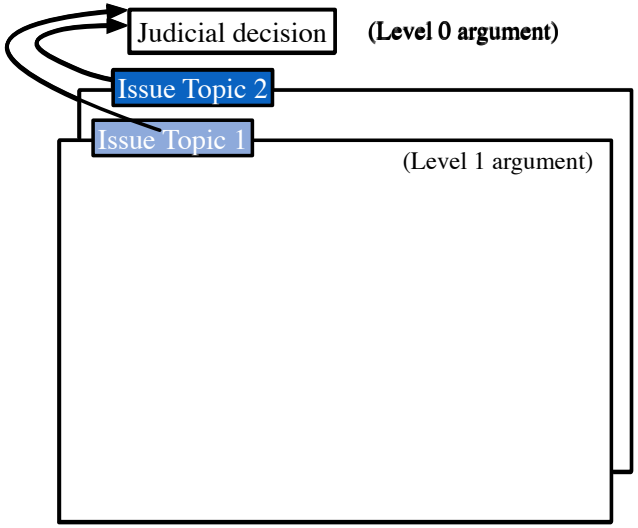
Lv0, Lv1, Lv2, Lv3, Lower

Tasks described

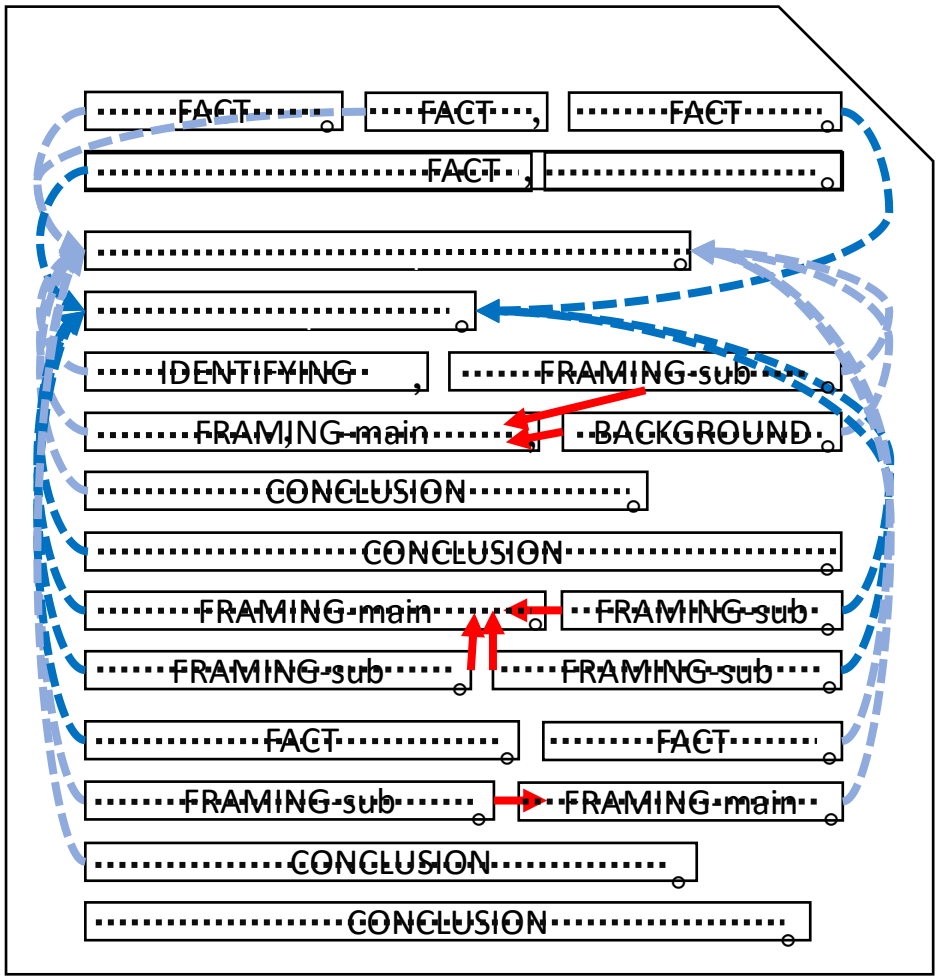
1. Rhetorical status classification
   - Assign rhetorical status to each text span.
2. Issue Topic Identification
   - Find Issue Topics.
3. Issue Topic Linking
   - Link each rhetorical unit to the Issue Topic it belongs to.
4. Argumentative relation extraction

Source	Destination
FRAMING-sub	FRAMING-main
BACKGROUND	

   - Only if argumentative support exists.



Text unit: comma-separated text piece
Span: sequence of text units



Hiroaki Yamada, Simone Teufel and Takenobu Tokunaga. 2017. Annotation of argument structure in Japanese legal documents. In Proceedings of the 4th Workshop on Argument Mining (ArgMining2017). pages 22-31.

Rhetorical status categories

Categories adaptation from Hachey and Grover (2006) for Japanese judgement documents

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

B. Hachey, C. Grover (2006): Extractive summarisation of legal texts. Artificial Intelligence and Law, 14(4).

Our corpus

- Source: http://courts.go.jp
  - Time frame: 2003/04/15 ~ 2016/12/31
  - General civil cases
  - Only documents which have manually generated summaries are extracted.
- Target documents
  - Documents exclusion criteria:
    - Exclude error documents(OCR errors)
    - Doc length<=400000
    - 150 <= summary length <= 450

Hiroaki Y, Simone T, Takenobu, T (2018) Building a corpus of legal argumentation in Japanese judgement documents: towards structure-based summarisation. Artificial Intelligence and Law, To appear

Our corpus

Overview	
# of docs	89
# of units	136972
# of sentences	37590
# of characters	2528604

Distribution of labels (units)	
FACT	43022
FRAMING-main	36648
FRAMING-sub	28857
OTHER	16816
CONCLUSION	5841
IDENTIFYING	4565
BACKGROUND	791
IssueTopic	432

Last week, we had just finished the additional annotation with 31 docs. Now 120 docs are available.

An experiment for Rhetorical status classification

by SVM, CRF, and NN.

Rhetorical status classification sample

- Assign rhetorical role to each sentence

FACT
BACKGROUND
FRAMING-main
CONCLUSION

陰影は鎖骨の幅を超えている.

……

かかる注意義務の基準となるべきものは，当時のいわゆる臨床医学の実践における医療水準であるが(最高裁昭和57年3月30日第三小法廷判決・裁民135号563項参照), 集団検診には，（1）で説示した制約・限界が内在することを照らせば，集団検診におけ胸部x線写真の読影に係る医療水準とは自ずと異なると言うべきである。

……

以上によれば，14年写真……以上ありとして指摘すべきかどうかの判断が異なりうると言わざるを得ないから，……注意義務に反するものということはできない。

Rhetorical status categories

Categories adaptation from Hachey and Grover (2006) for Japanese judgement documents

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

B. Hachey, C. Grover (2006): Extractive summarisation of legal texts. Artificial Intelligence and Law, 14(4).

Experimental setting

- Task:
  - A multi class sentence labeling task.
  - NOT multi label.
- Data:
  - All 89 documents
  - 7 classes
  - 5-fold cross validation
- Classifier:
  - Support Vector Machine with linear kernel
  - Conditional Random Field
  - Simple NN models with Char+Word embeddings

Distribution of labels (sentences)	
FACT	14816
FRAMING-main	8611
FRAMING-sub	7280
OTHER	4294
CONCLUSION	1455
IDENTIFYING	799
BACKGROUND	302
Total	37371

SVM and CRF

- Features:
  - *Bigram (the bag of lemmatized bigrams of morphemes)
  - *Sentence location
  - *Sentence length
  - Modality expressions[a]
  - Function expressions[b]
  - Cue phrases
  - Law names
- 2 models
    - Base : features only three (*)
    - All : all features

[a] 益岡隆志, 日本語モダリティ探究, くろしお出版, 2007 に基づく
[b]松吉 俊,佐藤 理史,宇津呂 武仁 日本語機能表現辞書の編纂 を利用

Models

- SVM
  - SVM-Base
  - SVM-All
- CRF
  - CRF-Base
  - CRF-All
- NN
  - Single sentence
    - Word+Char embeddings
      - Base
      - Fine-tuning (Word+Char)
      - Fine-tuning (Word+Char) + Hand crafted features
      - Fine-tuning (Word+Char) + AE
  - Inter sentence
    - [Word + Char FT(w+c)] + **LSTM-CRF**
    - [Word + Char FT(w+c)] + AE + **LSTM-CRF**

Results

- Fine-tuning(FT) showed some improvements.
- When the model uses pre-trained encoder (AE), the F-value gets better.
- The injection of hand-crafted features is not helpful.
- Inter-sentence level LSTM-CRF layer improves the results.

Model			F-value	Precision	Recall
SVM		SVM-Base	0.556	0.641	0.533
		SVM-All	0.566	0.641	0.541
CRF		CRF-Base	0.619	0.659	0.597
		CRF-All	**0.629**	0.670	0.603
NN	Single sentence	Word+Char Base	0.558	0.625	0.536
		Word+Char FT(word+char)	0.586	0.618	0.573
		[Word+Char FT(w+c)] + Hf	0.535	0.591	0.519
		[Word+Char FT(w+c)] + AE	**0.591**	0.607	0.578
	Inter-sentences	[Word + Char FT(w+c)] + LSTM-CRF	0.589	0.645	0.565
		[Word + Char FT(w+c)] + LSTM-CRF + AE	**0.612**	0.646	0.594

Tentative result

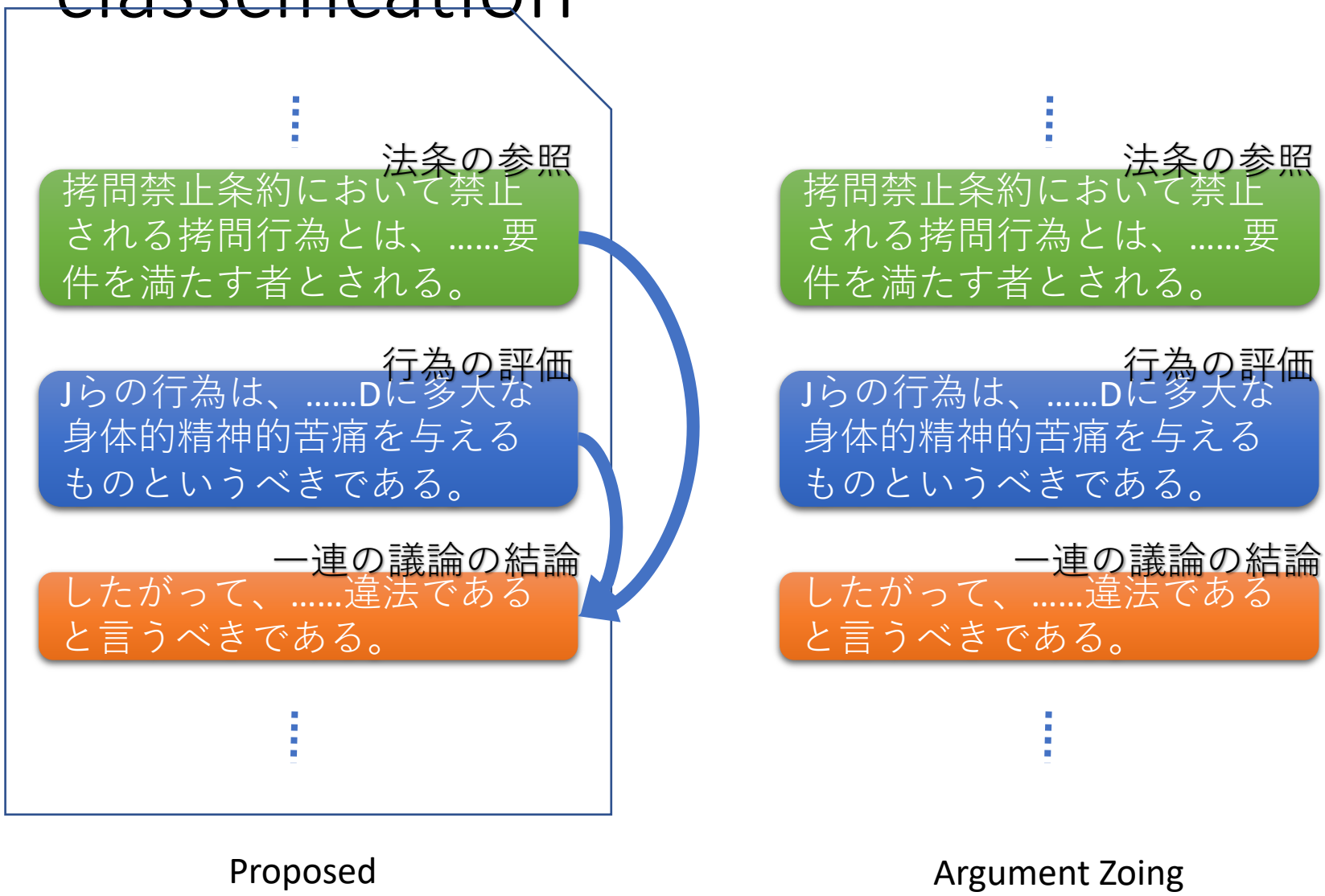
Wrap-up

- The CRF is still the best among all models.
- Fine-tuning could improve the performance.
  - Weights that are trained with unlabeled judgment documents help models.
- The model that consider inter-sentence context improve the result.
- Next, I will conduct the full experiment with additional (31 documents) and hyperparameter tuning.

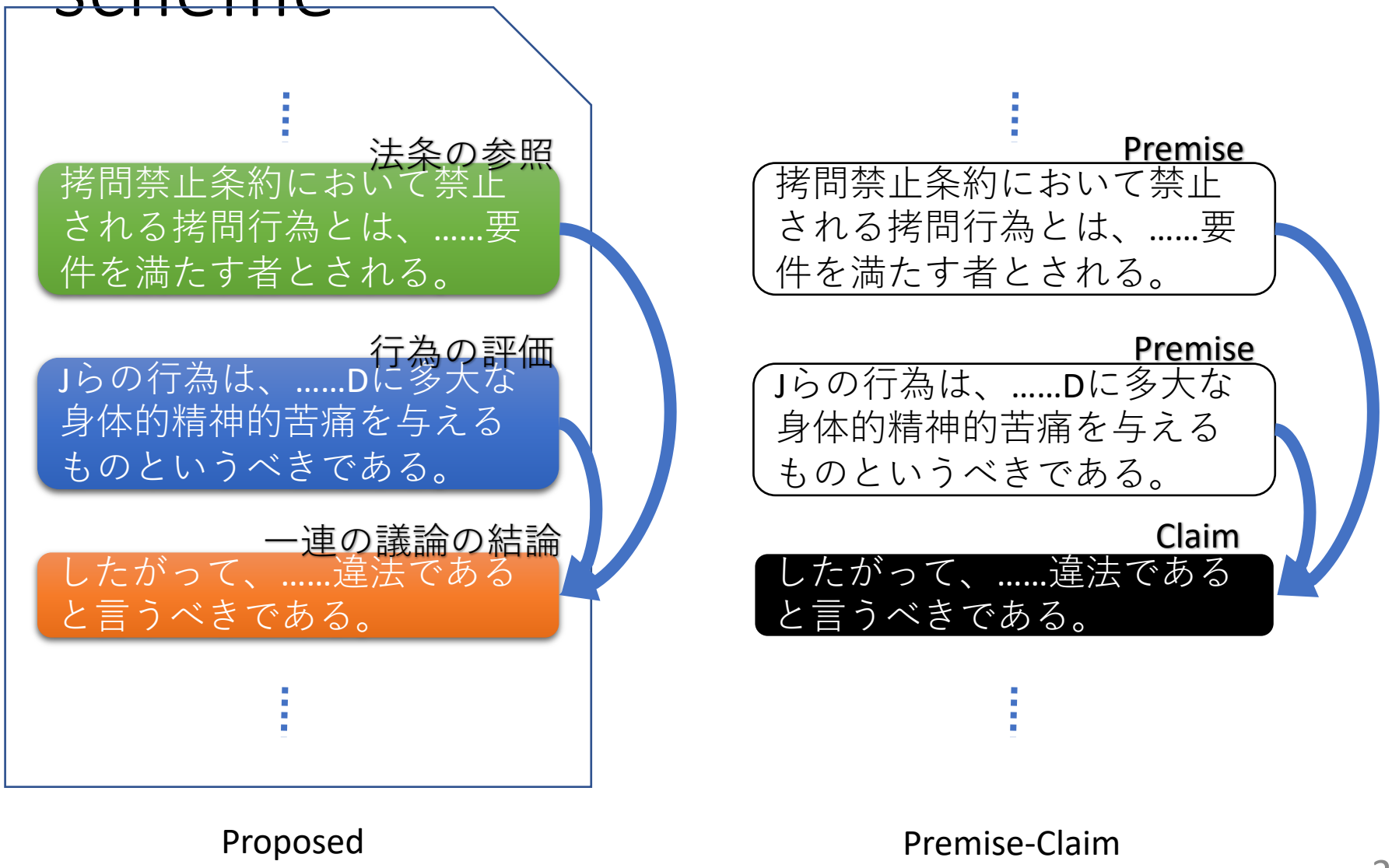
Appendix 1

Annotation scheme comparison

Proposed vs. Rhetorical status classcification



Proposed vs. Premise-Claim scheme



Appendix 2

Inter-annotator agreement related documents

Argument Zoning agreement

- Inter-annotator agreement:
  K = **0.70** (N = 9,879; n = 7, k = 2)

- Certain systematic assignment errors:
  - Most often confused: FRAMING-main and FRAMING-sub

Issue Topic Identification agreement metric

- We report agr_{ITI}, the number of spans agreeing between annotators, averaged over annotators
  - A: set of annotators.
  - a_s: number of agreeing spans (60% by characters).
  - $spans(i)$: number of spans annotated by annotator i.

$$agr_{\mathrm{ITI}}(i) = \frac{a_s}{spans(i)}$$

$$agr_{\mathrm{ITI}} = \frac{\sum_{i \in A} agr_{\mathrm{ITI}}(i)}{|A|}$$

Issue Topic Linking agreement metric

- We report agr_{ITL}, the number of units assigned to same Issue Topic, averaged over annotators
  - A: set of annotators.
  - a_u: number of units with identical Issue Topic .
  - $units(i)$: number of units annotated by annotator i.

$$agr_{\mathrm{ITL}}(i) = \frac{a_u}{units(i)}$$

$$agr_{\mathrm{ITL}} = \frac{\sum_{i \in A} agr_{\mathrm{ITL}}(i)}{|A|}$$

Issue Topic Identification/Linking results

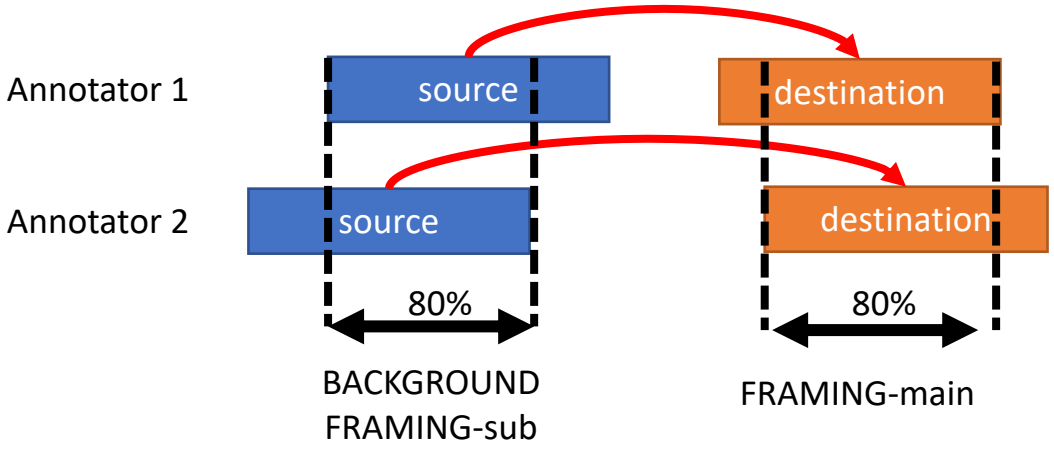
- Issue Topic Identification : $agr_{\mathrm{ITI}} = 0.79$

- Issue Topic Linking : $agr_{\mathrm{ITL}} = 0.87$

FRAMING Linking agreement metric

- Agreement on source spans
  - $agr_{src} = \frac{\#\ of\ agreed\ source\ spans\ with\ link}{\#\ of\ source\ spans\ with\ link}$
- Agreement on destination spans
  - $agr_{fl} = \frac{\#\ of\ agreed\ links}{\#\ of\ agreed\ source\ spans\ with\ link}$

Agree: “sharing more than 80% of the span in characters at the same location”



FRAMING Linking agreement result

- $agr_{src} = 0.67$
- $agr_{fl} = 0.66$

Target documents for annotation study

- “Medical negligence during a health check”
- “Threatening behavior in connection to money lending”
- “Use of restraining devices by police”
- “Fence safety and injury”
- “Mandatory retirement from private company”
- “Road safety in a bus travel sub-contract situation”
- “Railway crossing accident”
- “Withdrawal of a company’s garbage license by the city”

Rhetorical Status confusion matrix

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

FRAMING Linking Error analysis

- Error analysis
  - We manually analyzed overlap of non-agreed destination spans (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.
  - If we were to consider these error links as agreed, FRAMING linking agreement would rise to **0.788**.

  → Annotators potentially naturally and principally agree to a high degree on FRAMING Linking.

Overlap threshold 80 % or 60 %

- Issue Topic Identification (60%, character)
  - We initially set the threshold to 80%:
    - 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.
- FRAMING Linking (80%, character offset)
  - We wanted to allow only short and relatively meaningless adverbial modification.
  - As the location has to be identical, we do not have to worry about paraphrases.