Toward Automatic Delegation Extraction in Japanese Law

Tsuyoshi Fujita, Yuya Sawada, Yusuke Sakai, Taro Watanabe
Nara Institute of Science and Technology (NAIST)
fujita.tsuyoshi.fy4@naist.ac.jp
{yuya.sawada.sr7, sakai.yusuke.sr9, taro}@is.naist.jp

Abstract

The legal systems have a hierarchical structure, and a higher-level law often authorizes a lower-level law to implement detailed provisions, which is called *delegation*. When interpreting legal texts with delegation, readers must repeatedly consult the lower-level laws that stipulate the detailed provisions, imposing a substantial workload. Therefore, it is necessary to develop a system that enables readers to instantly refer to relevant laws in delegation. However, manually annotating delegation is difficult because it requires extensive legal expertise, careful reading of numerous legal texts, and continuous adaptation to newly enacted laws. In this study, we focus on Japanese law and develop a two-stage pipeline system for automatic delegation annotation. First, we extract keywords that indicate delegation using a named entity recognition approach. Second, we identify the delegated provision corresponding to each keyword as an entity disambiguation task. In our experiments, the proposed system demonstrates sufficient performance to assist manual annotation in practice.

1 Introduction

The legal systems in many jurisdictions have a hierarchical structure, and a higher-level law often authorizes a lower-level law to implement detailed provisions, which is called *delegation* (Yoshida, 2012; Del Monte and Mańko, 2021; Whittington and Iuliano, 2017; Sim et al., 2024). Figure 1 shows an example of delegation in Japanese law, where Article 24-3 of the Long-Term Care Insurance Act states that *"other necessary matters pertaining to a Designated and Entrusted Juridical Person for Prefectural Affairs are prescribed by a Cabinet Order."* The detailed provisions regarding the *"Designated and Entrusted Juridical Person for Prefectural Affairs"* are delegated to Article 11-9 of the Order for Enforcement of the Long-Term Care Insurance Act, which is a lower-level Cabinet Order. Understanding such delegation is essential for interpreting how multiple laws work in coordination. However, reading legal texts with numerous delegation clauses requires identifying the delegated laws and repeatedly consulting them, which imposes a substantial burden on legal practitioners. Thus, support systems that automatically identify and visualize delegation have long been developed by legal information providers for commercial and governmental use.

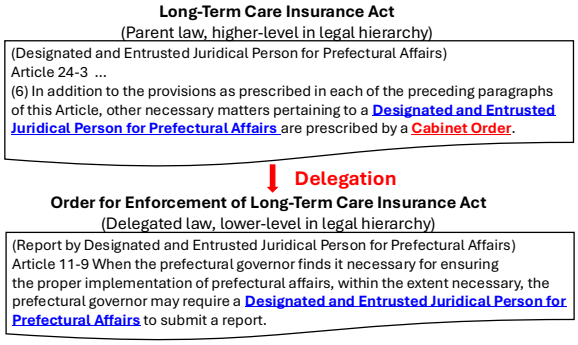

Figure 1: Example of delegation in Japanese laws. The original Japanese text is provided in Appendix I.

Effective deployment of such a system requires the system providers to annotate delegation across the entire body of laws, and updates must be made every time new laws are enacted or existing laws are amended. Over 6,000 new or amended laws are issued annually on average in Japan[1], for instance, and manually annotating the delegation for each revision requires significant costs and efforts.

In this study, we focus on Japanese law and build a two-stage pipeline system (Pozzi et al., 2023; Kannan Ravi et al., 2021; van Hulst et al., 2020; Sawada et al., 2024) to automatically identify del-

[1]Based on the legal information database D1-Law.com (https://www.daiichihoki.co.jp/d1-law/) provided by DAI-ICHI HOKI CO., LTD., we computed the average number of newly enacted laws and amended laws over five years from January 1, 2020, to December 31, 2024.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 5: Industry Track, pages 688–710

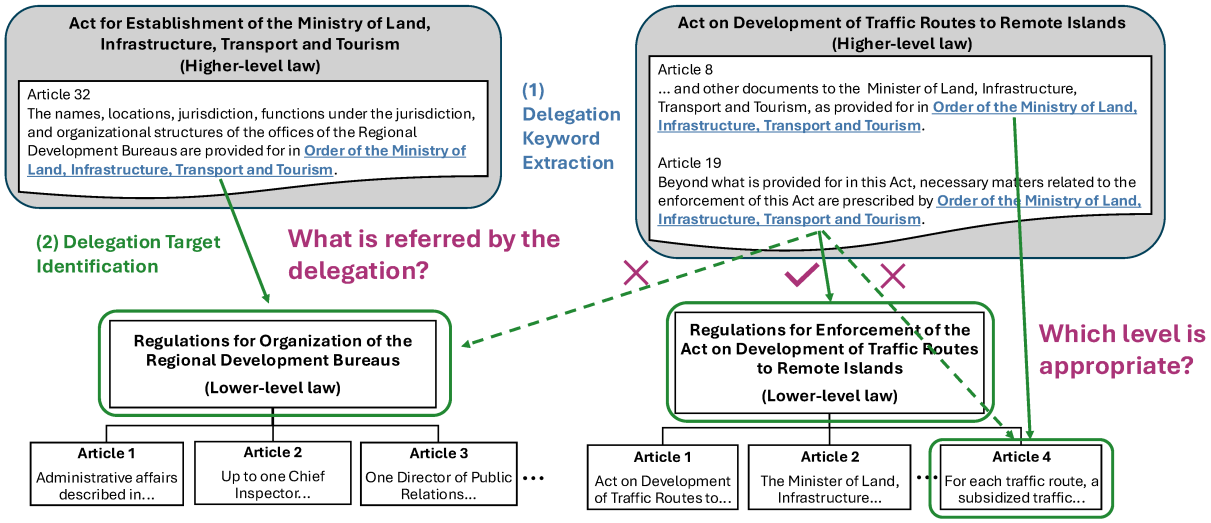

Figure 2: Overview of the delegation extraction task. First, we extract delegation keywords from a provision, such as *"Order of the Ministry of Land, Infrastructure, Transport and Tourism"* (**delegation keyword extraction**). Because multiple instances may appear as a delegation keyword, we then identify the appropriate Order issued by the Ministry from among many candidates, such as the *Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands*. In addition, since a delegation keyword may refer either to an entire law or to a specific provision, we also determine the appropriate granularity, i.e., Article 4 (**delegation target identification**).

egation. We formulate the extraction of keywords that signify delegation (*delegation keyword extraction*) as a Named Entity Recognition (NER) task, and the identification of the corresponding delegated provisions (*delegation target identification*) as an Entity Disambiguation (ED) task.

Experimental results using language models with diverse architectures show that the delegation keyword extraction achieves approximately 95 points in precision, recall, and F1, and the delegation target identification achieves a Recall@10 above 90. These results indicate that the system is sufficiently effective for a semi-automatic annotation workflow, where the pipeline presents a ranked list of candidate delegated provisions to annotators at legal information providers for selecting the correct one. This reduces the number of provisions that annotators must examine and decreases the annotation workload.

2 Delegation in Japanese Law

Figure 2 illustrates an overview of our delegation extraction task. When presented with a provision in a higher-level law as input[2], we first extract the delegation keywords appearing in the provision. Then, using these extracted keywords, we identify the specific provision in a lower-level law to which the higher-level provision delegates authority. In general, when a delegation source law is enacted, the corresponding delegation target law has not yet been promulgated. As a result, delegation keywords do not refer to the titles of specific laws. Instead, they appear as expressions indicating the type of the target laws, e.g., *"Cabinet Order"* or *"Order of the Ministry of Land, Infrastructure, Transport and Tourism"*, or the existence of delegated matters, e.g., *"as prescribed by the Minister of Economy, Trade and Industry"*. Therefore, the task involves first extracting these delegation keywords from the higher-level provision (**delegation keyword extraction**), and then identifying the specific lower-level provision to which they refer (**delegation target identification**).

A major challenge in delegation extraction is the high degree of ambiguity in delegation keywords. Because these expressions do not specify the exact title of the delegation target law, e.g., *"Order of the Ministry of Land, Infrastructure, Transport and Tourism"*, the same keyword may refer to different laws depending on the context. Hence, the system must select the correct delegation target law from many candidates, considering the content of the delegation source law and its provisions.

Furthermore, Japanese laws exhibit a hierarchi-

[2]Henceforth, we refer to the higher-level law that delegates authority as *delegation source law* and its relevant provision as *delegation source provision*. We also refer to the lower-level law to which the authority is delegated as *delegation target law* and its relevant provision as *delegation target provision*.

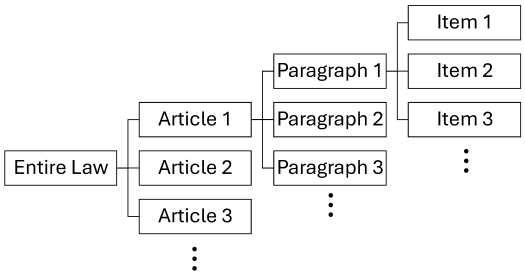

Figure 3: Hierarchical structure within a law, in which an upper-level element, e.g., Article 1, may contain multiple lower-level elements, e.g., Paragraph 2.

cal internal structure consisting of provisions at different levels of granularity, including articles, paragraphs, and items, as shown in Figure 3. Consequently, delegation target provisions may also vary in their granularity. Correctly determining the appropriate level for a delegation target requires not only resolving cross-law references but also interpreting the scope and abstraction level of the delegated matter. This requirement constitutes another distinctive challenge in delegation extraction. Appendix A provides further details and examples illustrating these two sources of difficulty drawn from Japanese laws. Appendix H compares the delegation extraction task with prior work.

3 Dataset and Task Definition

We construct a *delegation extraction dataset* that involves two subtasks, delegation keyword extraction and delegation target identification, based on a manually annotated legal provision data provided by DAI-ICHI HOKI CO., LTD. The provision data covers provisions from all laws enacted by the National Diet as well as orders issued by national administrative bodies. Each annotated provision contains information on the positions of delegation keywords and, where applicable, the corresponding delegation target provisions at various levels of granularity. From this data, we extract the subset of provisions that include both delegation keywords and target provisions, and we construct a delegation extraction dataset consisting of 69,730 sentences obtained by splitting 19,907 provisions at sentence boundaries. Using this dataset, we decompose and formalize the problem into the two subtasks. Appendix B provides detailed explanations, statistics, and examples of the dataset.

Delegation Keyword Extraction The dataset comprises 20,723 delegation keywords, which appear in 20,386 out of 69,730 sentences. Using the positional information of these delegation keywords within sentences, we formulate this subtask as an NER task that may contain zero or more named entities. We evaluate the model based on the exact matches between the gold and predicted keyword spans, using precision, recall, and F1 score.

Delegation Target Identification This task involves identifying the correct delegation target from all candidate provisions, analogous to entity disambiguation. Each delegation keyword in the dataset is annotated with a label indicating the delegation target provision at a specific level of granularity, i.e., the *delegation target label*, such as entire law, article, paragraph, or item. We utilize five levels of granularity when constructing the *provision database*, including entire laws, articles, paragraphs, items, and supplementary provisions, as these levels appear at least once as delegation targets in the dataset. The provision database contains approximately 2.28 million provisions, each consisting of the law title, article number, and text body. We treat all provisions in this database as candidate delegation targets and align their provision IDs with the delegation target labels in the delegation extraction dataset. To evaluate model performance in the semi-automatic annotation workflow described in Section 1, we use Recall@k (R@k), which measures whether the correct delegation target is included in the top-k retrieved candidates, and Mean Reciprocal Rank (MRR), which measures the reciprocal of the rank assigned to the correct target. We report results at four levels of granularity: entire law, article, paragraph, and item. Evaluation details are provided in Appendix C.

4 Method

As shown in Figure 2, we extract delegation by a pipeline system that consists of a delegation keyword extraction and a target identification module.

4.1 Delegation Keyword Extraction

We build keyword extraction models based on three approaches: sequence labeling (Devlin et al., 2019; Li et al., 2021; Lai et al., 2022) and span classification (Yamada et al., 2020; Fu et al., 2021) based on encoder-only language models, and the more recent generation-based method (Zhou et al., 2024; Sainz et al., 2024) based on decoder-only models. We evaluate them and use the best-performing model as the keyword extraction module.

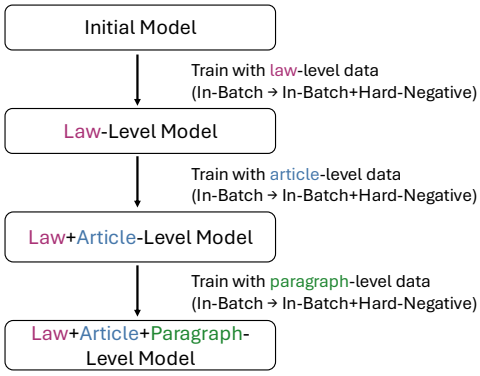

Figure 4: Overview of the granularity-aware training.

The sequence labeling model predicts BIO tags, while the span classification model identifies spans of up to 16 tokens corresponding to delegation keywords. However, in Japanese, both approaches suffer from tokenization mismatches, as the lack of whitespace often leads to boundary errors. Therefore, we first split each sentence at potential keyword boundaries using a dictionary and then tokenize each segment separately before feeding it into the model. The generation-based model directly generates delegation keywords. We compare English and Japanese prompts to examine whether a model benefits more from the dominant pre-training language or from prompts that align more naturally with Japanese legal texts. Appendix D.1 provides further details.

4.2 Delegation Target Identification

We build a model that searches for the delegation target provision in the provision database based on a delegation keyword using an entity retrieval framework. It is designed to incorporate the hierarchical structure of laws. Appendix D.2 provides further details of the model definition, model training, and inference.

Entity Retrieval A mention referring to a particular entity is treated as a query, and the system retrieves the corresponding entity from a knowledge base. Some studies employ pretrained text embedding models (Wang et al., 2024a) or dual encoders (Gillick et al., 2018) to convert both the mention and the entity into vector representations, enabling vector similarity search between them (Orlando et al., 2024; Nakatani et al., 2025; Gillick et al., 2019; Wu et al., 2020). In this study, we treat the delegation keyword as the query and the provision database as the target knowledge base, performing vector similarity search between keyword representations and candidate provision representations. We construct two retrieval models: a text embedding model and a dual encoder model. We compare their performance to analyze the characteristics of each approach and identify the method most suitable for the delegation extraction task.

Training We introduce the *granularity-aware training* strategy illustrated in Figure 4. We gradually change the granularity of both the delegation target labels in the delegation extraction dataset and the candidate provisions in the provision database, moving from the law-level to the article-level, and then to the paragraph-level. The model is trained in three stages, and we apply both `in-batch` training and `in-batch+hard-negative` training at each stage. For each input, `in-batch` uses the gold delegation targets of the other samples within the same minibatch as negative examples. In `in-batch+hard-negative`, we also include hard negatives, which are highly similar but incorrect provisions retrieved by the model using `in-batch` training. Notably, as shown in Figure 3, legal documents have a hierarchical structure comprising multiple levels of granularity, including entire laws, articles, and paragraphs. For tasks with such hierarchical label structures, coarse-to-fine training, where models are first fine-tuned on coarse-grained labels and then tuned on fine-grained ones, improves label prediction performance (Stretcu et al., 2020; Sadat and Caragea, 2022; Banerjee et al., 2019). Therefore, we progressively shift the training data from coarser to finer granularity to capture both the global relationships among laws and the fine-grained relationships among provisions at the article and paragraph levels.

5 Experimental Setup

We conduct five-fold cross-validation for both the delegation keyword extraction module and the delegation target identification module using the dataset described in Section 3. For evaluation, 20% of the dataset is held out as the test set, while the remaining 80% is further divided into training and development sets with a 9:1 ratio. We report the mean score and standard deviation across the five folds. We also evaluate the overall performance of the pipeline system that integrates these two modules.

Delegation Keyword Extraction We compare the performance of three model types: a span classification model based on LUKE (Yamada

et al., 2020), a sequence labeling model based on BERT (Devlin et al., 2019), and generative extraction models based on Llama 3.1 (Grattafiori et al., 2024) and Llama 3.1 Swallow (Fujii et al., 2024). For LUKE and BERT, we further compare models that incorporate sentence pre-segmentation to handle cases where token boundaries do not align with keyword boundaries (denoted as $\text{LUKE}_{\text{split}}$ and $\text{BERT}_{\text{split}}$), against models trained without pre-segmentation ($\text{LUKE}_{\text{w/o split}}$ and $\text{BERT}_{\text{w/o split}}$). Llama 3.1 is evaluated using the English prompt. For Swallow, we report results for both the English prompt ($\text{Swallow}_{\text{en}}$) and the Japanese prompt ($\text{Swallow}_{\text{ja}}$). All models are fine-tuned by each training data. Appendix E.1 provides the complete training details, e.g., hyperparameters.

Delegation Target Identification We use the multilingual E5 (Wang et al., 2024b) as the text embedding model and Japanese Tohoku-BERT as components of the dual encoder. We train both models using the granularity-aware training strategy described in Section 4.2 to construct delegation target identification models (E5_{step} and $\text{BERT}_{\text{step}}$). For these two models, we evaluate performance under two configurations during the final stage of training with item-level data: one using only `in-batch` training, and the other using both `in-batch` and `in-batch+hard-negative` training. As baselines, we also report the performance of models trained on all data without modifying either the training data or the provision database ($\text{E5}_{\text{w/o step}}$ and $\text{BERT}_{\text{w/o step}}$). For these models, we similarly evaluate performance under the same two settings: `in-batch` only, and `in-batch` plus `in-batch+hard-negative`. Appendix E.2 provides further training details and information.

Pipeline System We integrate both modules and use the keyword positions predicted by the keyword extraction module to construct the input for the delegation target identification module. The delegation target retrieval model then performs inference, and a prediction is counted correct only when the results of both the delegation keyword extraction and the delegation target identification are correct. Based on this criterion, we compute precision, recall, and F1 score to evaluate the overall performance of the pipeline system (Ayoola et al., 2022). For each module, we use the model that achieved the highest performance in its respective standalone experiments.

Model	Precision	Recall	F1
$\text{LUKE}_{\text{w/o split}}$	**97.6** (±0.211)	89.4 (±0.368)	93.3 (±0.266)
$\text{LUKE}_{\text{split}}$	95.7 (±0.464)	94.7 (±0.263)	**95.2** (±0.330)
$\text{BERT}_{\text{w/o split}}$	<u>96.9</u> (±0.514)	89.3 (±0.231)	92.9 (±0.317)
$\text{BERT}_{\text{split}}$	94.2 (±0.589)	94.5 (±0.276)	<u>94.4</u> (±0.407)
Llama 3.1	93.2 (±0.347)	94.6 (±0.240)	93.9 (±0.169)
$\text{Swallow}_{\text{en}}$	93.4 (±0.435)	**94.9** (±0.363)	94.1 (±0.309)
$\text{Swallow}_{\text{ja}}$	93.5 (±0.434)	<u>94.8</u> (±0.345)	94.1 (±0.303)

Table 1: Results for delegation keyword extraction

6 Results and Discussions

We will focus on summarizing important aspects here, and defer more discussions to Appendix F.

Delegation Keyword Extraction As Table 1 shows, all models achieved F1 scores above 93, indicating high overall performance on the delegation keyword extraction task. $\text{LUKE}_{\text{split}}$ achieved the best performance with the F1 score of 95.2. Among the model types, the encoder-only models, $\text{LUKE}_{\text{split}}$ and $\text{BERT}_{\text{split}}$, outperformed the decoder-based models, Llama 3.1, $\text{Swallow}_{\text{en}}$, and $\text{Swallow}_{\text{ja}}$. Moreover, the comparable performance of $\text{Swallow}_{\text{en}}$ and $\text{Swallow}_{\text{ja}}$ indicates that the difference between English and Japanese prompts has little effect in this task. This suggests that while decoder-only models designed for text generation can achieve reasonable performance through fine-tuning, encoder-only models are more suitable for this task due to their ability to efficiently capture bidirectional contextual information.

Delegation Target Identification Table 2 shows the results for the delegation target identification task. For each model, we focus on the better-performing variant between those trained with and without `in-batch+hard-negative` at the final training stage, and derive the following observations. First, both E5_{step} and $\text{BERT}_{\text{step}}$ outperform $\text{E5}_{\text{w/o step}}$ and $\text{BERT}_{\text{w/o step}}$, respectively, achieving improvements of about 20–30 points in R@1 across all evaluation granularities. In terms of MRR, E5_{step} and $\text{BERT}_{\text{step}}$ improve by about 10–15 points at the law level and by 20–30 points at finer granularities. These results indicate that the granularity-aware training strategy effectively enhances model performance. Second, both E5_{step} and $\text{BERT}_{\text{step}}$ achieve R@10 scores above 90 even at the most fine-grained "item" level. This suggests that the models provide sufficient performance for annotation support scenarios, where candidate provisions are presented to annotators, who then se-

Model	Eval Granularity	R@1	R@5	R@10	R@50	R@100	MRR
$E5_{w/o\ step}$	Entire Law	69.4 (±1.80)	85.5 (±0.685)	90.0 (±0.319)	95.9 (±0.0997)	97.2 (±0.116)	76.3 (±1.24)
		69.5 (±3.03)	85.1 (±1.60)	89.1 (±1.40)	94.7 (±0.614)	96.2 (±0.307)	76.4 (±2.15)
	Article	50.0 (±1.85)	70.0 (±1.39)	77.3 (±1.18)	88.9 (±0.585)	92.1 (±0.457)	58.8 (±1.49)
		43.8 (±7.02)	63.4 (±4.91)	70.8 (±3.95)	83.8 (±1.49)	87.9 (±1.03)	52.7 (±5.95)
	Paragraph	22.4 (±3.79)	62.6 (±2.76)	72.9 (±1.97)	87.2 (±0.745)	90.9 (±0.506)	40.6 (±3.25)
		22.8 (±7.08)	56.3 (±8.56)	65.4 (±7.07)	80.6 (±3.44)	85.4 (±2.18)	38.3 (±7.48)
	Item	21.8 (±3.91)	62.0 (±2.87)	72.4 (±2.08)	87.0 (±0.744)	90.8 (±0.533)	40.0 (±3.38)
		22.4 (±7.24)	55.7 (±8.90)	64.9 (±7.38)	80.3 (±3.58)	85.2 (±2.27)	37.9 (±7.68)
$E5_{step}$	Entire Law	93.2 (±0.347)	96.8 (±0.355)	97.6 (±0.300)	98.9 (±0.190)	99.2 (±0.152)	94.7 (±0.311)
		92.6 (±0.892)	96.7 (±0.416)	97.7 (±0.351)	98.9 (±0.282)	99.2 (±0.206)	94.4 (±0.654)
	Article	79.2 (±0.929)	90.5 (±0.873)	93.3 (±0.731)	96.9 (±0.486)	97.9 (±0.254)	84.1 (±0.886)
		78.9 (±1.84)	90.1 (±1.41)	92.9 (±1.09)	96.6 (±0.840)	97.5 (±0.610)	83.9 (±1.59)
	Paragraph	51.0 (±6.99)	88.4 (±1.02)	92.1 (±0.883)	96.5 (±0.554)	97.5 (±0.342)	68.3 (±3.75)
		56.3 (±6.64)	87.4 (±2.47)	91.4 (±1.75)	96.0 (±1.07)	97.1 (±0.882)	70.6 (±4.41)
	Item	50.8 (±7.00)	88.1 (±0.986)	91.9 (±0.847)	96.4 (±0.554)	97.5 (±0.357)	68.0 (±3.74)
		56.0 (±6.70)	87.1 (±2.56)	91.2 (±1.85)	95.9 (±1.09)	97.0 (±0.940)	70.3 (±4.46)
$BERT_{w/o\ step}$	Entire Law	72.0 (±1.04)	85.2 (±0.815)	89.4 (±0.512)	95.4 (±0.350)	96.9 (±0.274)	77.8 (±0.852)
		75.0 (±1.65)	87.4 (±0.903)	91.0 (±0.422)	96.0 (±0.194)	97.1 (±0.225)	80.5 (±1.27)
	Article	48.5 (±0.951)	67.2 (±1.07)	74.8 (±0.973)	87.6 (±0.913)	91.0 (±0.607)	57.0 (±0.889)
		50.7 (±2.11)	69.2 (±1.80)	76.3 (±1.21)	87.6 (±0.706)	91.0 (±0.585)	59.0 (±1.91)
	Paragraph	19.5 (±2.58)	58.0 (±1.99)	69.0 (±1.40)	85.3 (±1.00)	89.4 (±0.701)	37.0 (±2.25)
		25.9 (±2.89)	62.1 (±2.54)	71.7 (±1.52)	85.3 (±0.775)	89.3 (±0.675)	42.5 (±2.52)
	Item	18.7 (±2.59)	57.2 (±1.99)	68.4 (±1.43)	85.0 (±1.00)	89.2 (±0.685)	36.2 (±2.27)
		25.4 (±2.84)	61.6 (±2.50)	71.2 (±1.56)	84.9 (±0.775)	89.0 (±0.681)	42.0 (±2.50)
$BERT_{step}$	Entire Law	89.9 (±1.35)	94.6 (±0.854)	96.0 (±0.670)	98.3 (±0.328)	98.8 (±0.163)	92.0 (±1.02)
		91.6 (±1.78)	96.2 (±0.746)	97.5 (±0.365)	98.9 (±0.162)	99.3 (±0.138)	93.6 (±1.26)
	Article	70.8 (±4.21)	84.7 (±2.48)	88.9 (±1.75)	95.1 (±0.725)	96.6 (±0.464)	76.8 (±3.46)
		78.3 (±3.86)	90.0 (±1.97)	92.9 (±1.20)	97.0 (±0.355)	97.9 (±0.224)	83.4 (±3.06)
	Paragraph	44.5 (±5.73)	80.9 (±3.63)	86.7 (±2.40)	94.4 (±0.898)	96.2 (±0.624)	61.0 (±4.86)
		55.6 (±5.43)	87.5 (±2.58)	91.5 (±1.52)	96.4 (±0.526)	97.5 (±0.330)	70.1 (±4.23)
	Item	44.0 (±5.78)	80.5 (±3.73)	86.4 (±2.48)	94.3 (±0.961)	96.1 (±0.604)	60.6 (±4.93)
		55.3 (±5.45)	87.2 (±2.61)	91.3 (±1.57)	96.3 (±0.537)	97.5 (±0.354)	69.8 (±4.25)

Table 2: Results of the delegation target identification task. For $E5_{step}$ and $BERT_{step}$, the upper rows show results with `in-batch` only at the last step of the granularity-aware training, while the lower rows show results with both `in-batch` and `in-batch+hard-negative`. For $E5_{w/o\ step}$ and $BERT_{w/o\ step}$, the upper rows show results using `in-batch` once, and the lower rows show results using one round each of `in-batch` and `in-batch+hard-negative`.

Granularity	Precision	Recall	F1
Entire Law	89.6 (±0.234)	88.6 (±0.345)	89.1 (±0.244)
Article	76.7 (±0.888)	75.9 (±0.972)	76.3 (±0.920)
Paragraph	54.0 (±6.90)	53.3 (±6.72)	53.6 (±6.81)
Item	53.6 (±6.97)	53.0 (±6.79)	53.3 (±6.88)

Table 3: Results for the pipeline system combining $LUKE_{split}$ and $E5_{step}$

lect the correct delegation target provision. By presenting highly ranked candidate provisions, the models can help reduce the number of provisions annotators need to inspect, thereby lowering the annotation burden.

Pipeline System Table 3 shows the results of the pipeline system combining $LUKE_{split}$ and $E5_{step}$, which achieved the highest performance in the above experiments. For $E5_{step}$, we use the better-performing variants at each granularity, trained with or without `in-batch+hard-negative` at the final step of the granularity-aware training. The pipeline system achieves a precision, recall, and F1 score of around 90% at the law level, indicating that it can reasonably to handle the high ambiguity of delegation keywords. However, its performance drops substantially at finer granularities, i.e., article, paragraph, and item levels, revealing remaining challenges in selecting the appropriate granularity of delegation target provisions in accordance with the abstraction level of delegated matters.

7 Case Studies

To further examine the effect of the granularity-aware training described in Section 4.2, we analyze cases where $E5_{w/o\ step}$ failed to identify the correct delegation target provision while $E5_{step}$ succeeded. For both models, we use variants trained with `in-batch+hard-negative` training, which

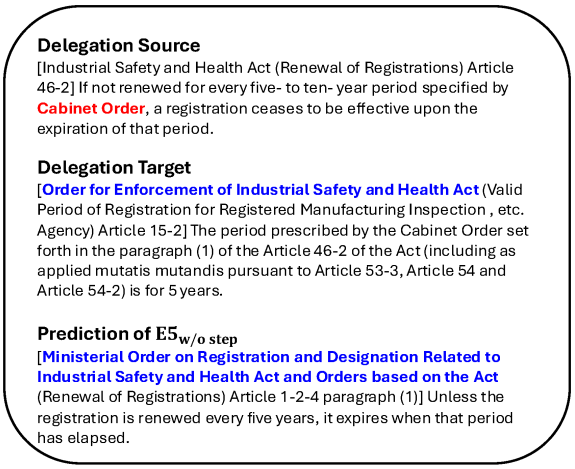

Figure 5: A case in which $E5_{\text{w/o step}}$ predicted an incorrect provision whose content is highly similar to the delegation source provision, whereas $E5_{\text{step}}$ correctly identified the delegation target provision.

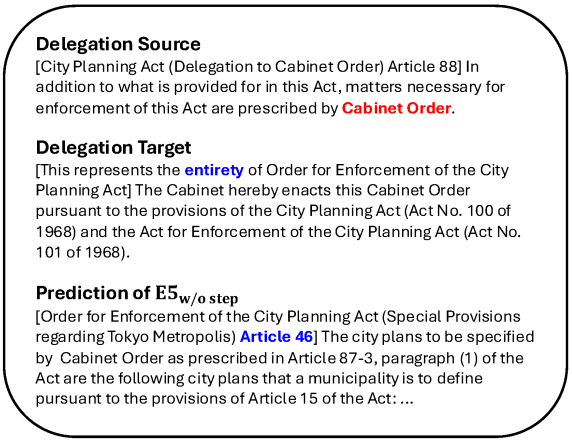

Figure 6: A case in which $E5_{\text{w/o step}}$ predicted an provision at an incorrect granularity, whereas $E5_{\text{step}}$ correctly identified the delegation target provision.

achieved better performance at the finest "item" level, compared to the models constructed without `in-batch+hard-negative` training.

In the example shown in Figure 5, the delegation keyword is "Cabinet Order" and the correct delegation target provision is Article 15-2 of the *Order for Enforcement of the Industrial Safety and Health Act*. Here, this *Order* is itself a Cabinet Order, and the reference expression "Article 46-2 of the Act" in the target provision indicates the source provision, which together allow the correct target to be inferred. While $E5_{\text{step}}$ correctly identified the target, $E5_{\text{w/o step}}$ instead predicted Article 1-2-4, paragraph (1) of the *Ministerial Order on Registration and Designation Related to Industrial Safety and Health Act and Orders based on the Act*, whose heading "Renewal of Registrations" and description "Unless the registration is renewed every five years, it expires" are semantically similar to the source provision. This observation suggests that the granularity-aware training in Section 4.2 enables the model to perform inference not only based on textual similarity, but also by exploiting cues such as law types and reference expressions. This capability likely contributes to handling the substantial ambiguity inherent in delegation keywords in the delegation extraction task.

Figure 6 shows an example where the delegation keyword is "Cabinet Order," and the correct delegation target is the entirety of the *Order for Enforcement of the City Planning Act*. The enacting clause of this *Order* identifies itself as a Cabinet Order established based on the *City Planning Act*, which is the delegation source law. This allows us to infer that this *Order* is the delegation target law. Moreover, from the general provision in the source article, "matters necessary for enforcement of this Act are prescribed by Cabinet Order," one can infer that the delegation target is not a specific element within the *Order* but rather the *Order* in its entirety. While $E5_{\text{step}}$ correctly identified the delegation target, $E5_{\text{w/o step}}$ incorrectly predicted Article 46 of the *Order for Enforcement of the City Planning Act*. Article 46 regulates only a specific matter, i.e., city planning in "Tokyo Metropolis," and is thus not appropriate as a delegated target. As discussed in Section 2, the delegation extraction task has the difficulty of selecting the appropriate granularity of the delegation target. The example in Figure 6 suggests that the granularity-aware training in Section 4.2 helps the model to handle relatively coarse-grained decisions such as choosing between "entire law" and "article".

8 Conclusion

This study introduces the task of automatically extracting delegation between legal provisions in Japanese laws. We developed a two-stage pipeline system comprising the delegation keyword extraction module and the delegation target identification module. Our experiments demonstrated the effectiveness of the granularity-aware training strategy, where we gradually refine the granularity of the training data from the law-level to the article-level and then to the paragraph-level. The resulting performance is sufficient to support human annotation in our actual operational setting.

Limitations

Task Setting and Practical Objectives Our current aim is to support human annotators in practical workflows, rather than to achieve fully automatic delegation extraction. Accordingly, our system is designed to reduce annotation effort by retrieving high-quality candidate provisions that can be efficiently verified and corrected by experts. Given this practical requirement, the experimental design emphasizes realistic annotation-support scenarios rather than end-to-end autonomous extraction. Within this scope, we have confirmed that the proposed system meets the performance needs identified in our operational context. Fully automatic delegation extraction, as well as extensions to more complex settings, e.g., multi-hop delegation across multiple laws, are left for future work.

Data Availability and Practical Relevance Our dataset contains proprietary annotations that cannot be publicly released due to contractual and operational constraints. While this limitation affects public availability, the dataset was constructed specifically for real-world deployment within a realistic annotation workflow. Therefore, the findings presented here reflect the characteristics and requirements of actual delegation extraction tasks in Japanese laws, and we believe they offer direct value for applied legal NLP research.

Dataset Scope We primarily focus on Japanese legislation. Although the core task of extracting delegation clauses is not unique to Japan, legal drafting conventions and terminology vary considerably across jurisdictions, such as the EU (Del Monte and Mańko, 2021), the UK (Sim et al., 2024), and the United States (Whittington and Iuliano, 2017). Nevertheless, focusing on a single, well-defined jurisdiction enables us to clarify fundamental challenges in delegation extraction and to demonstrate the feasibility and utility of our framework in a practical setting. Therefore, investigating cross-jurisdictional extensions and evaluation on other legislative corpora is left for future work.

Model Training As discussed in Section 6, our system shows degraded performance when identifying delegation targets at finer-grained levels of legal structure, such as paragraphs and items. However, even at these finer-grained levels, our system achieves performance sufficient for semi-automatic annotation support scenarios. Looking ahead, a key direction for future work is to incorporate the hierarchical structure of legislative texts directly into the model architecture or loss function design. Exploring these measures, in addition to utilizing our granularity-aware training strategy, would enable the system to better capture relationships across different levels of legal granularity and enhance prediction accuracy at finer-grained levels.

Multiple Delegation Targets We regard a prediction as correct if the model identifies at least one of the true delegation targets for a given delegation keyword, even when multiple targets exist. However, multi-target cases are rare. Our evaluation setting simplifies the task compared to real legal analysis, where it is sometimes necessary to identify all corresponding delegation targets to fully understand the scope and effect of a provision. Nonetheless, this formulation captures an essential aspect of the delegation extraction problem: successfully retrieving at least one valid target already reduces the effort required for delegation annotation. Moving forward, a possible extension is to enhance the delegation target identification module to handle multi-target scenarios, e.g., by adding a binary classifier that determines whether each retrieved candidate is a valid delegation target. This would enable more comprehensive extraction and improve the utility in practical annotation settings.

Ethical Considerations

Licenses Our dataset was constructed from the legal texts publicly available in the e-Gov Legislation Search (e-Gov 法令検索)[3] and the proprietary annotation data provided by DAI-ICHI HOKI CO., LTD. The translations of legal texts used in this paper are prepared by the authors with reference to the Japanese Law Translation Database System[4]. Both the e-Gov Legislation Search and the Japanese Law Translation Database System provide data under terms of use compatible with the Creative Commons Attribution 4.0 International License (CC BY), which permits the use of the data in this study. In addition, DAI-ICHI HOKI CO., LTD. permitted the authors to use its annotation data.

Harmful Content The data used in this study are publicly available legal texts and proprietary annotations of delegation on them, both of which are free of harmful content.

[3] https://laws.e-gov.go.jp/

[4] https://www.japaneselawtranslation.go.jp/

Acknowledgements

In this study, we used the proprietary annotation data of delegation in Japanese law, provided by DAI-ICHI HOKI CO., LTD. We thank the anonymous reviewers for their valuable comments and suggestions.

References

Nida Ahmed, Seemab Latif, Rabia Irfan, Adnan Ul–Hasan, and Faisal Shafait. 2022. Comparison of transformer models for information extraction from court room records in pakistan. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 01–06.

Mousumi Akter, Erion Çano, Erik Weber, Dennis Dobler, and Ivan Habernal. 2025. A comprehensive survey on legal summarization: Challenges and future directions. *Preprint*, arXiv:2501.17830.

Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2025. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *Preprint*, arXiv:2410.21306.

Ting Wai Terence Au, Vasileios Lampos, and Ingemar Cox. 2022. E-NER — an annotated named entity recognition corpus of legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 246–255, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. 2019. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6295–6300, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting contract elements. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, page 19–28, New York, NY, USA. Association for Computing Machinery.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, 11:102050–102071.

Micaela Del Monte and Rafał Mańko. 2021. Understanding delegated and implementing acts. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/690709/EPRS_BRI(2021)690709_EN.pdf. Accessed: 2025-11-15.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maria Duarte, Pedro A. Santos, João Dias, and Jorge Baptista. 2022. Semantic norm recognition and its application to portuguese law. *Preprint*, arXiv:2203.05425.

Yi Feng, Chuanyi Li, and Vincent Ng. 2024. Legal case retrieval: A survey of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6472–6485, Bangkok, Thailand. Association for Computational Linguistics.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. SpanNER: Named entity re-/recognition as span prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7183–7195, Online. Association for Computational Linguistics.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

Akshita Gheewala, Chris Turner, and Jean-Rémi de Maistre. 2019. Automatic extraction of legal citations using natural language processing. In *Proceedings of the 15th International Conference on Web Information Systems and Technologies*, WEBIST 2019, page 202–209, Setubal, PRT. SCITEPRESS - Science and Technology Publications, Lda.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *Preprint*, arXiv:1811.08008.

Ingo Glaser, Bernhard Waltl, and Florian Matthes. 2018. Named entity recognition, extraction, and linking

in german legal contracts. In *Jusletter IT*, February. Editions Weblaw. Publisher Copyright: (c) 2018 Editions Weblaw. All rights reserved.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. Overview of benchmark datasets and methods for the legal information extraction/entailment competition (coliee) 2024. In *New Frontiers in Artificial Intelligence*, pages 109–124, Singapore. Springer Nature Singapore.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ben Hagag, Gil Gil Semo, Dor Bernsohn, Liav Harpaz, Pashootan Vaezipoor, Rohit Saha, Kyryl Truskovskyi, and Gerasimos Spanakis. 2024. LegalLens shared task 2024: Legal violation identification in unstructured text. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 361–370, Miami, FL, USA. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in Indian court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 184–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Manoj Prabhakar Kannan Ravi, Kuldeep Singh, Isaiah Onando Mulang', Saeedeh Shekarpour, Johannes Hoffart, and Jens Lehmann. 2021. CHOLAN: A modular approach for neural entity linking on Wikipedia and Wikidata. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 504–514, Online. Association for Computational Linguistics.

Takahiro Komamizu, Katsuhiko Toyama, Nobuo Kawaguchi, and Tomoya Sano. 2022. Towards mobility-related law search by utilizing relationship between laws. *The Japanese Society for Artificial Intelligence Technical Report, Type 2*, 2022(SWO-057):04. (In Japanese).

Peichao Lai, Feiyang Ye, Lin Zhang, Zhiwei Chen, Yanggeng Fu, Yingjie Wu, and Yilei Wang. 2022. PCBERT: Parent and child BERT for Chinese few-shot NER. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2199–2209, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xiaonan Li, Yunfan Shao, Tianxiang Sun, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2021. Accelerating BERT inference for sequence labeling via early-exit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 189–199, Online. Association for Computational Linguistics.

Antoine Louis and Gerasimos Spanakis. 2022. A statutory article retrieval dataset in French. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6789–6803, Dublin, Ireland. Association for Computational Linguistics.

Jorge Martinez-Gil. 2023. A survey on legal question–answering systems. *Computer Science Review*, 48:100552.

Hibiki Nakatani, Hiroki Teranishi, Shohei Higashiyama, Yuya Sawada, Hiroki Ouchi, and Taro Watanabe. 2025. A text embedding model with contrastive example mining for point-of-interest geocoding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7279–7291, Abu Dhabi, UAE. Association for Computational Linguistics.

Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. ReLiK: Retrieve and LinK, fast and accurate entity linking and relation extraction on an academic budget. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14114–14132, Bangkok, Thailand. Association for Computational Linguistics.

Vasile Pais, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Riccardo Pozzi, Riccardo Rubini, Christian Bernasconi, and Matteo Palmonari. 2023. Named entity recognition and linking for entity extraction from italian civil judgements. In *AIxIA 2023 – Advances in Artificial Intelligence*, pages 187–201, Cham. Springer Nature Switzerland.

Damith Premasiri, Tharindu Ranasinghe, Ruslan Mitkov, Mo El-Haj, and Ingo Frommholz. 2025. Survey on legal information extraction: current status and open challenges. *Knowledge and Information Systems*, 67:11287–11358.

Mobashir Sadat and Cornelia Caragea. 2022. Hierarchical multi-label classification of scientific documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8923–8937, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*.

Yuya Sawada, Yuichiro Yasui, Hiroki Ouchi, Taro Watanabe, Masayuki Ishii, Shotaro Ishihara, Takeshi Yamada, and Hiroyuki Shindo. 2024. Construction and analysis of similarity-based nikkei company id linking system. *Journal of Natural Language Processing*, 31(3):1330–1355. (In Japanese).

Shahmin Sharafat, Zara Nasar, and Syed Waqar Jaffry. 2019. Data mining for smart legal systems. *Computers & Electrical Engineering*, 78:328–342.

Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*, 13:18253–18276.

Duncan Sim, Richard Whitaker, and Graeme Cowie. 2024. Delegated powers and framework legislation. `https://researchbriefings.files.parliament.uk/documents/CBP-10046/CBP-10046.pdf`. Accessed: 2025-11-15.

Otilia Stretcu, Emmanouil Antonios Platanios, Tom Mitchell, and Barnabás Póczos. 2020. Coarse-to-Fine Curriculum Learning for Classification. In *International Conference on Learning Representations (ICLR) Workshop on Bridging AI and Cognitive Science (BAICS)*.

Weihang Su, Yiran Hu, Anzhe Xie, Qingyao Ai, Quezi Bing, Ning Zheng, Yun Liu, Weixing Shen, and Yiqun Liu. 2024. STARD: A Chinese statute retrieval dataset derived from real-life queries by non-professionals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10658–10671, Miami, Florida, USA. Association for Computational Linguistics.

Rohit Upadhya and Santosh T.y.s.s. 2025. LexCLiPR: Cross-lingual paragraph retrieval from legal judgments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13971–13993, Vienna, Austria. Association for Computational Linguistics.

Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. Rel: An entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2197–2200, New York, NY, USA. Association for Computing Machinery.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Keith E. Whittington and Jason Iuliano. 2017. The myth of the nondelegation doctrine. *University of Pennsylvania Law Review*, 165(2):379–.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Toshihiro Yoshida. 2012. Series: Introduction to legal research for r&d and business part 2: National legal system. *Journal of Information Processing and Management*, 55(8):591–595. (In Japanese).

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted distillation from large language models for open named entity recognition. In *The Twelfth International Conference on Learning Representations*.

A Difficulties in the Delegation Extraction Task

The examples in this appendix provide concrete instantiations of the two main sources of difficulty discussed in Section 2: (1) ambiguity arising from delegation keywords that do not specify the exact target law, and (2) variation in the appropriate granularity of the delegation target provision.

A.1 Ambiguity of Delegation Keyword

As shown in Figure 2, both Article 32 of the *Act for Establishment of the Ministry of Land, Infrastructure, Transport and Tourism* and Article 19 of the *Act on Development of Traffic Routes to Remote Islands* contain the delegation keyword "Order of the Ministry of Land, Infrastructure, Transport and Tourism." In the former case, the delegation target law is the *Regulations for Organization of the Regional Development Bureaus*, which specify the responsibilities and internal structure of the Regional Development Bureaus. In the latter case, the delegation target law is the *Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands*, which prescribes various matters related to the enforcement of the *Act on Development of Traffic Routes to Remote Islands*. Although the same surface form "Order of the Ministry of Land, Infrastructure, Transport and Tourism" appears in both examples, the appropriate delegation target law must be selected from many Orders issued by the Ministry, based on the content of the delegation source law and provision[5].

A.2 Variation in the Delegation Target Granularity

Article 19 of the *Act on Development of Traffic Routes to Remote Islands* mentioned in Appendix A.1 delegates general matters concerning the enforcement of the *Act* to the entirety of the *Regulations for Enforcement of the Act*. In contrast, Article 8 of the same *Act* delegates authority not to the entirety of the *Regulations*, but specifically to Article 4 of the *Regulations*, as shown in Figure 2. Article 8 of the *Act* stipulates that "Order of the Ministry of Land, Infrastructure, Transport and Tourism," which is the delegation keyword, prescribes submission of a traffic route profit and loss statement. This is why the delegation target provision is Article 4 of the *Regulations* that describes the details of the submission, instead of its entirety. Accurately identifying the proper granularity of the delegation target provision requires interpreting both the scope and the level of abstraction of the matter being delegated.

B Detailed Dataset Information

Table 4 shows the frequency and examples of keywords in the keyword extraction dataset. The delegation keywords include not only single nouns such as "Cabinet Order" and "Order of the Ministry of Land, Infrastructure, Transport and Tourism," but also enumerations of multiple nouns, such as "Order of the Ministry of Internal Affairs and Communications / Order of the Ministry of Finance" and expressions involving verbs, such as "prescribed by the Minister of Health, Labor and Welfare."

Table 5 presents the distribution of delegation target labels and the provision database by granularity. While most delegation target labels correspond to entire laws, articles, or paragraphs, the provision database also contains approximately 800,000 provisions at other granularities, namely items and entire supplementary provision sections[6], which are less frequently observed as delegation target labels.

Figure 7 and Table 6 show examples in the keyword extraction dataset and the provision database, respectively. The example in Figure 7 is annotated with the delegation keyword span "[258, 313]," which identifies the keyword "Order of the Ministry of Land, Infrastructure, Transport and Tourism," as well as the delegation target provision ID "12." The delegation target provision can be identified from the provision database by looking up this ID.

[5]The dataset used in this study (described in Section 3) contains 330 distinct delegation target laws referred to by the keyword "Order of the Ministry of Land, Infrastructure, Transport and Tourism".

[6]Japanese laws consist of a main provision section and a supplementary provision section. The main provision section contains substantive provisions of the law, while the supplementary provision section includes ancillary regulations, such as the effective date of the law. Although both the main provision sections and supplementary provision sections are composed of elements such as articles, paragraphs, and items, in this study we do not distinguish whether a given element belongs to the main provision sections or to the supplementary provision sections.

Frequency	Unique Keywords	Total Occurrences	Example Keywords
101–	22 (4.9%)	18,576 (90.0%)	Cabinet Order, Order of the Ministry of Land, Infrastructure, Transport and Tourism
11–100	46 (10.3%)	1,289 (6.2%)	prescribed by the Minister of Health, Labor and Welfare
2–10	154 (34.5%)	634 (3.1%)	Order of the Ministry of Internal Affairs and Communications / Order of the Ministry of Finance
1	224 (50.2%)	224 (1.1%)	specified separately by the Minister of Finance

Table 4: Frequency and examples of delegation keywords in the delegation extraction dataset

Granularity	Delegation target labels	Provision database
Entire law	5,169	29,788
Article	16,923	418,855
Paragraph	12,906	1,020,253
Item	25	599,426
Supp	12	208,252
Total	35,035	2,276,574

Table 5: Breakdown of delegation target labels and the provision database by granularity. The counts of delegation target labels are shown in total occurrences. "Supp" refers to the granularity of the entire supplementary provisions section.

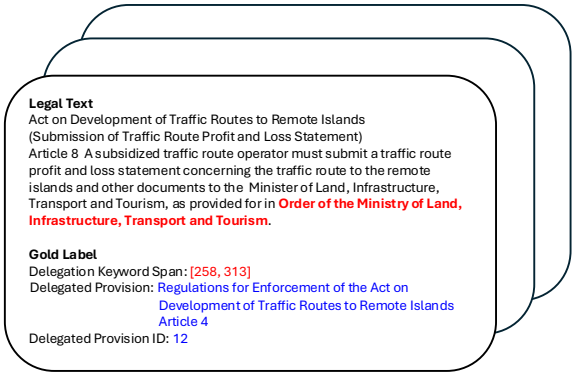

Figure 7: Example of the keyword extraction dataset

ID	Title of Law	Article Number	Text
1	Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands	Entirety	[This represents the entirety of Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands] Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands are hereby established as follows.
2	Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands	Article 1	[Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands (Application for Traffic Route Subsidy) Article 1] Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands…
3	Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands	Article 1 Paragraph (1)	[Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands (Application for Traffic Route Subsidy) Article 1 paragraph (1)] Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands…
4	Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands	Article 1 Paragraph (2)	[Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands (Application for Traffic Route Subsidy) Article 1 paragraph (2)] The written application prescribed in the preceding paragraph is to be…
⋮			
12	Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands	Article 4	[Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands (Submission of Traffic Route Profit and Loss Statement) Article 4] For each traffic route…

Table 6: Example entries in the provision database

C Evaluation Details

In the evaluation of the delegation target identification module, we construct the input for the model by utilizing the positions of delegation keywords, which serve as the gold labels in the delegation keyword extraction task. Based on these inputs, the model retrieves the top-k ($k = 1, 5, 10, 50, 100$) candidate provisions, from which we calculate R@k and MRR to assess model performance. R@k represents the proportion of instances in which the correct delegation target provision appears among the top-k candidates, and is defined as follows:

$$\mathrm{R@}k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{r_i \leq k\} \tag{1}$$

where N denotes the total number of input instances, r_i is the rank of the correct delegation target for instance i, and $\mathbb{1}\cdot$ is an indicator function that returns 1 if the condition holds and 0 otherwise. Meanwhile, MRR measures how highly the correct delegation target appears in the similarity ranking, and is defined as:

$$\mathrm{MRR} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{r_i} \tag{2}$$

The evaluation is conducted at four levels of granularity: entire law, article, paragraph, and item. For each granularity, when the gold delegation target provision is annotated at a finer level, a prediction is considered correct if the predicted delegation target matches the gold delegation target up to the evaluated granularity. For instance, in the article-level evaluation, if the gold delegation target is specified at a finer level, such as paragraph or item, the prediction is considered correct as long as it matches the gold target up to the article level. For example, if the correct target provision is Article 12, paragraph 1 of the *Regulations*

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 08 Oct 2025

A virtual assistant answers questions from a user based on the provided text.<|eot_id|><|start_header_id|>user<|end_header_id|>

Find all the keywords associated with legal delegation in the following text of Japanese laws and regulations. The output should be in a list of tuples of the following format: [("keyword 1", "DELEGATION"), ...].
Text: {text}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 8: English prompt used in the generation-based extraction model.

for Enforcement of Navigation Aids Act (see Figure 15 in Appendix G.2), then predicting Article 12 of the same *Regulations* is also treated as correct. Conversely, when the gold target is annotated at a coarser granularity than the evaluation level, a prediction is counted as correct only when it exactly matches the gold target. For example, in the article-level evaluation, if the gold target is annotated at a coarser level such as entire law, only predictions of the entire law are considered correct; predictions of individual articles within that law are counted as incorrect[7]. When multiple delegation targets are associated with a single instance, the prediction is regarded as correct if any of them are successfully retrieved.

D Methodology Details

D.1 Delegation Keyword Extraction

D.1.1 Pre-Segmentation of Input Texts

To handle the misalignment between token boundaries and delegation keyword boundaries caused by tokenization errors, we employ the pre-segmentation of input texts at potential keyword boundaries. Concretely, we first identify strings in input texts that exactly match delegation keywords appearing in the training and development sets, which together cover 80% of the dataset. For each input sentence, we scan for these exact matches and split the input text into segments at their boundaries. Each resulting segment is tokenized independently using the tokenizer of the model. Finally, the tokenized segments are concatenated to form the input sequence.

[7]Although the dataset includes provisions at the entire supplementary provision sections level, which lies between entire law and article in granularity, we exclude this level from evaluation for simplicity.

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

バーチャルアシスタントは、提供されたテキストに基づいてユーザーの質問に答えます。
<|eot_id|><|begin_of_text|><|start_header_id|>user<|end_header_id|>

以下の日本の法令文から、法令間の委任関係を示すキーワードを見つけてください。 出力は以下の形式のタプルのリストにしてください： [("keyword 1", "DELEGATION"), ...]
Text: {text}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Figure 9: Japanese prompt used in the generation-based extraction model.

D.1.2 Prompts for Generation-Based Models

Figures 8 and 9 show the English and Japanese prompts we use for generation-based delegation keyword extraction models.

D.2 Delegation Target Identification

D.2.1 Model Definition

The delegation target identification model computes the similarity between a delegation keyword and a candidate provision using the inner product of their vector representations:

$$score(x, y) = \boldsymbol{h}_x^\top \boldsymbol{h}_y \tag{3}$$

where $\boldsymbol{h}_x$ and $\boldsymbol{h}_y$ denote the vector representations of the keyword description x and the candidate provision description y, respectively. Each vector is obtained by encoding the token sequences corresponding to x and y as follows:

$$\boldsymbol{h}_x = red(E_1(x)), \tag{4}$$
$$\boldsymbol{h}_y = red(E_2(y)) \tag{5}$$

Here, E_1 and E_2 are encoders. When using a text embedding model, a single encoder is employed ($E_1 = E_2$). The function $red(\cdot)$ converts the encoder output into $\boldsymbol{h}_x$ and $\boldsymbol{h}_y$. For the text embedding model, it is defined as average pooling over the final-layer token representations (Wang et al., 2024a), while for the dual encoder architecture, the final-layer [CLS] token output is used (Humeau et al., 2020; Wu et al., 2020). The maximum token length is set to 128, and tokens beyond this limit are truncated.

The input description x of a delegation keyword is constructed as:

$$x = \texttt{header}\ \texttt{ctxt}_l\ [\mathrm{M_s}]\ \texttt{keyword}\ [\mathrm{M_e}]\ \texttt{ctxt}_r \tag{6}$$

where `keyword` denotes the delegation keyword, and $\texttt{ctxt}_l$ and $\texttt{ctxt}_r$ represents its left and right

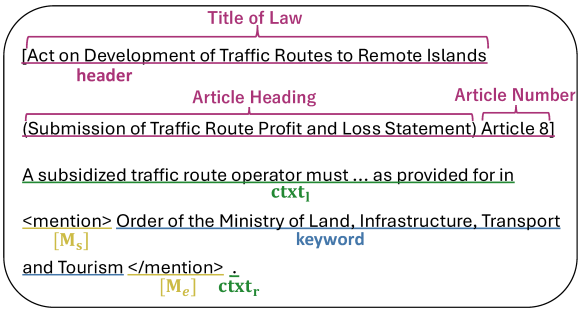

Figure 10: Example of delegation keyword description x

Figure 11: Example of candidate provision description y

contexts, respectively. $[\mathrm{M_s}]$ and $[\mathrm{M_e}]$ are special tokens that mark the start and end of the keyword span (Wu et al., 2020). `header` is a concatenation of the law title, article heading, and article number, enclosed in brackets. Because law titles and article numbers may be referenced in target provisions, and article headings often summarize the content of the provision, this information can be useful for delegation target identification. For instance, for Article 8 of the *Act on Development of Traffic Routes to Remote Islands*, the `header` is "[Act on Development of Traffic Routes to Remote Islands (Submission of Traffic Route Profit and Loss Statement) Article 8]," and it is followed by the delegation keyword "Order of the Ministry of Land, Infrastructure, Transport and Tourism" and its surrounding context (see Figure 10). When using a text embedding model, the prefix "query:" is added to the beginning of x.

Each candidate provision description y from the provision database $\mathcal{Y}$ is constructed as:

$$y = \texttt{header text} \tag{7}$$

where `text` represents the provision text, and `header` is constructed in the same way as for x. When the candidate refers to an entire law, the header is a concatenation of the string "This represents the entirety of" and the law title. In this case, when the law contains an enacting clause, it is used as `text`. If no such clause exists, `text` remains empty. Since enacting clauses often contain references to higher-level laws that serve as legal bases of the law, they may provide useful clues for delegation target identification (Komamizu et al., 2022). For example, in the case of the entirety of the *Regulations for Organization of the Regional Development Bureaus*, the `header` is "[This represents the entirety of Regulations for Organization of the Regional Development Bureaus]," and the `text` consists of its enacting clause (see Figure 11). When using a text embedding model, "passage:" is added at the beginning of y, whereas for a dual encoder model, a special token is inserted between `header` and `text` (Wu et al., 2020).

D.2.2 In-Batch training and In-Batch+Hard-Negative Training

During training, we compute the probability that y in the provision database $\mathcal{Y}$ is the delegation target provision for x, denoted as $P(y \mid x)$, based on their similarity scores. The model parameters are optimized to maximize the similarity between x and its correct target provision. To reduce computation cost, $P(y|x)$ is approximated as:

$$P(y \mid x) \simeq \frac{\exp(score(x, y))}{\sum_{y' \in \mathcal{Y}_C} \exp(score(x, y'))} \tag{8}$$

where y' denotes a provision in the candidate set $\mathcal{Y}_C$, which consists of delegation target provisions in the current minibatch $\mathcal{Y}_B \subset \mathcal{Y}$ and hard negative samples $\mathcal{Y}_{hard} \subset \mathcal{Y}$, i.e., $\mathcal{Y}_C = \mathcal{Y}_B \cup \mathcal{Y}_{hard}$. The hard negative samples are constructed by first training a model using only the candidate provisions in the minibatch (`in-batch` training with $\mathcal{Y}_C = \mathcal{Y}_B$), then retrieving the top 10 provisions in $\mathcal{Y} \backslash \{y\}$ that have the highest similarity to x (Gillick et al., 2019). Finally, using parameters from the `in-batch` training as initialization, we fine-tune the model on $\mathcal{Y}_C$ to minimize the following loss (`in-batch+hard-negative` training):

$$\mathcal{L} = -score(x, y) + \log(\sum_{y' \in \mathcal{Y}_C} \exp(score(x, y'))). \tag{9}$$

D.2.3 Granularity-Aware Training

As described in Section 4.2, we employ the granularity-aware training strategy during the

model training. In this strategy, we gradually change the set of provisions $\mathcal{Y}$ from the law-level to the paragraph-level, and perform model training described in Appendix D.2.2 in three stages.

In the first stage of training, the model is trained on law-level data, where both the gold labels y in the training data and the provision database $\mathcal{Y}$ are converted to the law-level granularity. Concretely, if a training instance specifies a delegation target provision at the article or paragraph level, the corresponding label is replaced with the provision ID representing the entire law. For example, if Article 4 of the *Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands* is designated as the delegation target, it is replaced with the entirety of the *Regulations*. In addition, all provisions finer than the law-level are removed from the provision database $\mathcal{Y}$. Using these law-level training data and database, we perform both `in-batch` training and `in-batch+hard-negative` training for the retrieval model.

Next, we successively train the model using article-level data and paragraph-level data, where the granularity of the data is refined from articles to paragraphs. As in the law-level stage, we convert the target granularity of the training data, and remove the provisions that are finer than the target granularity from the provision database[8]. During each training stage, the model parameters are initialized with those obtained from the previous stage.

D.2.4 Inference

At inference time, the model predicts the delegation target provision $\hat{y}$ for an input provision x as the candidate provision $y \in \mathcal{Y}$ with the highest similarity score, as defined in Equation 10. The representation of candidate provisions is precomputed and cached, and the nearest neighbor for an input provision x is searched using Faiss (Johnson et al., 2021).

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} score(x, y) \tag{10}$$

In nearest neighbor search, we exclude candidate provisions belonging to the same law as the input provision. Although these provisions often receive high similarity scores due to their identical or similar content to the input, they cannot be assigned as the delegation target provision, as delegation occurs only from higher-level to lower-level laws. To ensure valid predictions, we exclude the provisions that belong to the same law as the input provision, and identify the delegation target provision as the one with the highest similarity among all remaining candidate provisions.

[8]Due to the data specification used in this study, some "articles" in the law are stored in the provision database not as articles themselves but as the first paragraph of the corresponding article. For these "articles," we concatenate all paragraphs belonging to the same article to newly construct an "article," and then remove finer-grained elements. In total, 42,613 new articles were created following this procedure.

Models	Hugging Face ID
Delegation Keyword Extraction	
LUKE	studio-ousia/luke-japanese-large-lite
BERT	tohoku-nlp/bert-large-japanese-v2
Llama3.1	meta-llama/Llama-3.1-8B-Instruct
Swallow	tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5
Delegation Target Identification	
E5	intfloat/multilingual-e5-base
BERT	tohoku-nlp/bert-large-japanese-v2

Table 7: List of the models we used in this study and their corresponding Hugging Face IDs.

E Detailed Experimental Setup

We use the Hugging Face Transformers (Wolf et al., 2020) implementation when we train models. Table 7 shows the list of model names and their Hugging Face IDs. For delegation target identification, we use the Base model of E5 and the Large model of BERT to keep the model sizes approximately comparable. We train and evaluate all the models using a single NVIDIA A6000 GPU with 48GB of memory or a single NVIDIA A100 GPU with 80GB of memory.

E.1 Delegation Keyword Extraction

We fine-tune all the models on our training set, with the generation-based models using LoRA (Hu et al., 2022). Table 8 shows the models and the hyperparameters used in training. Note that for training Llama 3.1 Swallow, the same hyperparameters as Llama 3.1 are used.

E.2 Delegation Target Identification

For delegation target identification, we use E5 and BERT as encoder models. Both models are trained using the hyperparameters listed in Table 9. Regarding the number of epochs, we first perform `in-batch` training followed by `in-batch+hard-negative` training for four epochs for each step.

	LUKE	BERT	Llama 3.1
batch size (train)	8	10	10
batch size (eval)	8	20	4
learning rate	1e−5	1e−5	1e−5
epochs	5	20	5
optimizer	AdamW	AdamW	AdamW
eps	1e−6	1e−6	1e−6
scheduler	linear	linear	cosine
warmup ratio	0.06	0.06	0.1
weight decay	0.01	0.01	0.1
max grad norm	0.00	0.00	0.3
β	[0.9, 0.98]	[0.9, 0.98]	[0.9, 0.98]

Table 8: Hyperparameters used in training delegation keyword extraction models

batch size (train)	16
batch size (eval)	16
learning rate	1e−5
epochs	4
optimizer	AdamW
eps	1e−6
scheduler	linear
warmup ratio	0.06
weight decay	0.01
max grad norm	0.00
β	[0.9, 0.98]

Table 9: Hyperparameters used in training delegation target identification models

F Additional Discussions

F.1 Delegation Keyword Extraction

Comparing $\text{LUKE}_{\text{split}}$ and $\text{BERT}_{\text{split}}$ in Table 1, the F1 score of $\text{LUKE}_{\text{split}}$ is 0.8 points higher. LUKE is pretrained with a large entity-annotated corpus from Wikipedia and is thus specialized for entity-related tasks, which likely contributed to its effectiveness in the delegation keyword extraction.

In addition, the pre-segmentation of input sentences at keyword boundaries, described in Section 4.1, improved F1 scores by 1.9 and 1.5 points for $\text{LUKE}_{\text{split}}$ and $\text{BERT}_{\text{split}}$, respectively, compared with their non-segmented counterparts. Both models showed an increase in recall of more than 5 points, accompanied by a decrease in precision of about 1.9 points, indicating that pre-segmentation tends to increase the number of detected keywords. In fact, for LUKE, the number of detected keywords increased by approximately 1,500 after pre-segmentation, suggesting that verb expressions such as "specified" were more likely to be detected as keywords.

Case studies using the predictions of the delegation keyword extraction models are provided in Appendix G.1.

F.2 Delegation Target Identification

Table 2 shows the results of delegation target identification. Focusing on the better-performing variant between those trained with and without `in-batch+hard-negative` training at the final stage, we derive the following observations.

(1) As discussed in Section 6, both E5_{step} and $\text{BERT}_{\text{step}}$ achieved a level of performance sufficient for annotation support scenarios. However, R@1 of both models at the item level remains around 55 points, indicating that further improvement is needed for fully automatic delegation target identification without human verification.

(2) As described in Section 2, the delegation keywords are often expressed generically and do not explicitly mention the title of the delegation target law. Therefore, the model must infer the appropriate target law from a large number of candidates based on the context of the delegation source provision. In this respect, both E5_{step} and $\text{BERT}_{\text{step}}$ achieve R@1 scores exceeding 90 at the law level, suggesting that the granularity-aware training enables the models to capture broader inter-law relationships and partially resolve the ambiguity of delegation keywords at the law level.

(3) Comparing E5_{step} and $\text{BERT}_{\text{step}}$, E5_{step} consistently outperforms across all granularities in both R@1 and MRR. One possible explanation is that E5 learns semantic representations over relatively long text units, such as sentences and discourse-level segments, during pretraining (Wang et al., 2024a), whereas BERT applies bidirectional self-attention at the token level, focusing on contextualized token representations (Devlin et al., 2019). Since the delegation target identification requires understanding not only the meanings of individual delegation keywords but also the overall content and abstraction level of provisions across sentences or broader textual contexts, the ability of E5 to represent longer textual units may provide a distinct advantage.

Additional case studies of the delegation target identification results are provided in Appendix G.2.

G Additional Case Studies

G.1 Delegation Keyword Extraction

G.1.1 Pre-Segmentation at Keyword Boundaries

To verify the effect of pre-segmentation of input sentences at keyword boundaries described in Section 4.1, we analyze cases that led to the higher recall and lower precision of $\text{LUKE}_{\text{split}}$ compared to $\text{LUKE}_{\text{w/o split}}$. Fisrtly, Figure 12 shows a case where $\text{LUKE}_{\text{split}}$ successfully extracted a delegation keyword that $\text{LUKE}_{\text{w/o split}}$ failed to detect. In this example, 文部科学大臣の定め (*specified by the Minister of Education, Culture, Sports, Science and Technology*) is the delegation keyword. However, the tokenizer of LUKE segments the keyword and words surrounding it into [...に関し, 、 , 文部科学, 大臣, の, 定める, 資格] ([*regarding ...*, ",", "*Education, Culture, Sports, Science and Technology*", *Minister*, *by*, *specified*, *qualifications*]), preventing $\text{LUKE}_{\text{w/o split}}$ from detecting 文部科学大臣の定め as a contiguous keyword span. On the other hand, since 定め appears as a delegation keyword in the training and development data, $\text{LUKE}_{\text{split}}$ first splits the sentence into smaller blocks such as [...に関し、 文部科学大臣の, 定め, る資格](["*by the Minster of Education, Culture, Sports, Science and Technology regarding ...*", *specified*, *qualifications*]). It then applies the tokenizer to each block individually, obtaining the token sequence [に関し, 、 , 文部科学, 大臣, の, 定め, る, 資格]. This enables $\text{LUKE}_{\text{split}}$ to correctly detect 文部科学大臣の定め as a delegation keyword. This case demonstrates that pre-segmentation of input sentences at potential keyword boundaries serves as an effective strategy to mitigate false negatives caused by tokenization errors at keyword spans.

Secondly, Figure 13 illustrates one of the cases that led to reduced precision for $\text{LUKE}_{\text{split}}$. In this example, no delegation keyword is present. $\text{LUKE}_{\text{w/o split}}$ correctly predicted the absence of a keyword, whereas $\text{LUKE}_{\text{split}}$ produced a false positive by detecting 定め (*specified*). Because the tokenizer of LUKE segments the keyword and words surrounding it into [期間を, 定めて, その, 業務, に従事] ([*period*, *specified*, *that*, *service*, "*engaging in*"]), the token 定め does not constitute a candidate span in $\text{LUKE}_{\text{w/o split}}$. In contrast, $\text{LUKE}_{\text{split}}$ pre-segments the input sentence at keyword boundaries, yielding the token sequence [期間を, 定め, て, その, 業務, に従事]. As a consequence, $\text{LUKE}_{\text{split}}$ treats 定め as a span for classification. In this particular case, it is likely that $\text{LUKE}_{\text{split}}$ labeled 定め as a keyword span due to surface-level similarity with instances observed in the training data.

Legal Text
[EN] Teachers in specialized training colleges must have qualifications **specified by the Minister of Education, Culture, Sports, Science and Technology**, regarding specialized knowledge or skills of the education they are in charge of.
[JA] 専修学校の教員は、その担当する教育に関する専門的な知識又は技能に関し、**文部科学大臣の定め**る資格を有する者でなければならない。

Delegation Keyword
[EN] "**specified by the Minister of Education, Culture, Sports, Science and Technology**"
[JA] "**文部科学大臣の定め**"

Prediction of $\text{LUKE}_{\text{w/o split}}$
No keywords

Prediction of $\text{LUKE}_{\text{split}}$
[EN] "**specified by the Minister of Education, Culture, Sports, Science and Technology**"
[JA] "**文部科学大臣の定め**"

Figure 12: Example contributing to the improved recall of $\text{LUKE}_{\text{split}}$. $\text{LUKE}_{\text{w/o split}}$ failed to detect this keyword, whereas $\text{LUKE}_{\text{split}}$ correctly predicted it.

Legal Text
[EN] The Minister of Internal Affairs and Communications may revoke a radio operator's license, or order a radio operator to cease engaging in that service for a **specified** period not exceeding three months, if the radio operator falls under one of the following items:
[JA] 総務大臣は、無線従事者が左の各号の一に該当するときは、その免許を取り消し、又は三箇月以内の期間を**定め**てその業務に従事することを停止することができる。

Delegation Keyword
No keywords

Prediction of $\text{LUKE}_{\text{w/o split}}$
No keywords

Prediction of $\text{LUKE}_{\text{split}}$
[EN] "**specified**"
[JA] "**定め**"

Figure 13: Example leading to reduced precision in $\text{LUKE}_{\text{split}}$. While $\text{LUKE}_{\text{w/o split}}$ correctly predicted that no keyword is present, $\text{LUKE}_{\text{split}}$ falsely detected 定め (*specified*).

G.1.2 Model Confidence for an Incorrect Prediction

Sequence labeling models and span classification models output a confidence score[9] for each predicted label, indicating whether each token or span in the input text is a delegation keyword. In this section, we analyze a case in which the model as-

[9] A value between 0 and 1.

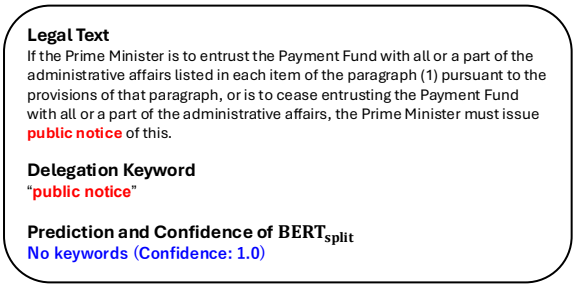

Figure 14: Example in which the model made an incorrect prediction with high confidence.

signs a high confidence score but still produces an incorrect prediction, in other words, a case that is particularly challenging for the model. We use BERT$_{split}$, a sequence labeling model, and define the confidence score of the token corresponding to the predicted keyword span as the prediction confidence of the model. When a predicted span consists of multiple tokens, we use the average of their confidence scores.

In the example shown in Figure 14, the delegation keyword is "public notice." Although the model fails to detect this keyword, the confidence score assigned to the corresponding span is 1.0. The delegation extraction dataset constructed in this study contains 20,723 keyword spans, but "public notice" appears as a keyword in only 27 of them (0.13%). In this case, the low frequency of "public notice" in the training data is likely one of the reasons the model failed to detect it.

G.2 Delegation Target Identification

To examine remaining challenges in delegation target identification, we analyze a case in Figure 15, where both E5$_{w/o\ step}$ and E5$_{step}$ failed to make a correct prediction [10]. Here, the delegation keyword is "Order of the Ministry of Land, Infrastructure, Transport and Tourism," and the delegation target provision is Article 12, paragraph (1) of the *Regulations for Enforcement of the Navigation Aids Act*. The delegation source provision of this example states that matters concerning a report to the Commandant of the Japan Coast Guard in the event of an accident shall be prescribed by the "Order of the Ministry of Land, Infrastructure, Transport and Tourism." Within Article 12 of the *Regulations*, paragraph (1) regulates the reporting duties to the Commandant of the Japan Coast Guard, while paragraph (2) defines the Commandant's subsequent actions. Therefore, the delegation target provision in this case is limited to Article 12, paragraph (1), instead of the entire Article 12.

[10]We use the prediction of the models trained with `in-batch+hard-negative` training, in the same way as in Section 7.

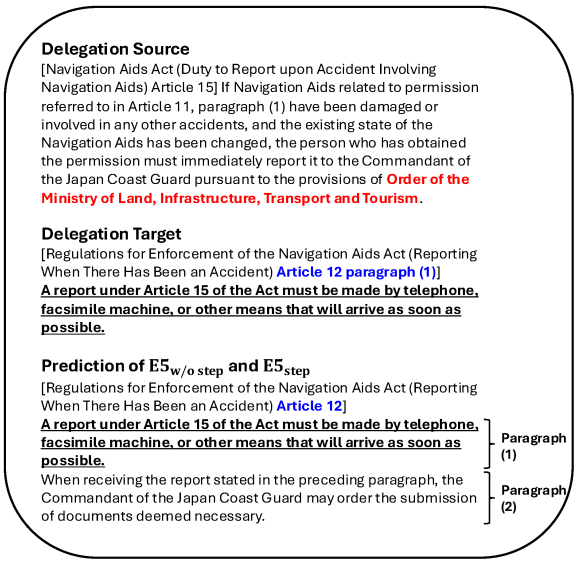

Figure 15: A case in which both E5$_{w/o\ step}$ and E5$_{step}$ predicted a provision at an incorrect granularity relative to the correct delegation target provision.

However, both E5$_{step}$ and E5$_{w/o\ step}$ incorrectly predicted the entire Article 12, including paragraph (2), as the delegation target. With respect to the challenge of selecting the appropriate granularity of the delegation target, one of the key difficulties in the delegation extraction task, this example demonstrates that fine-grained granularity decisions, such as choosing between the article level and the paragraph level, remain an unsolved challenge.

H Related Work

H.1 Legal NLP

In the legal domain, a wide range of NLP studies have been conducted on laws from various jurisdictions, including information extraction tasks such as NER and relation extraction, question answering (QA), legal judgment prediction, and document summarization (Zhong et al., 2020; Ariai et al., 2025; Siino et al., 2025). Legal information extraction aims to automatically extract elements such as legal concepts, actors, and actions, as well as the relations among them, from legal texts (Premasiri et al., 2025). Research on legal QA focuses on methods that retrieve legal documents relevant to a given legal question and generate appropriate answers (Martinez-Gil, 2023). Legal judgment

prediction seeks to infer court decisions from descriptions of the facts of a case (Cui et al., 2023). Legal document summarization focuses on developing summarization methods that address the characteristics of legal texts, including their considerable length, specialized terminology and formats, and extensive cross-references to other legal documents (Akter et al., 2025). Among these tasks, those most closely related to the delegation extraction task examined in this paper are NER and QA.

H.2 Named Entity Recognition

Research on NER in the legal domain has examined entities appearing in statutes enacted by legislatures, regulations issued by administrative agencies, and judicial decisions handed down by courts (Pre-masiri et al., 2025). In addition to standard NER categories such as persons and locations, some studies focus on legal-domain-specific categories (Hagag et al., 2024; Au et al., 2022; Kalamkar et al., 2022). Among such domain-specific categories, those similar to the delegation keywords addressed in this work include references within the input text to court decisions, statute titles, and article numbers (Gheewala et al., 2019; Ahmed et al., 2022; Sharafat et al., 2019; Chalkidis et al., 2017; Pais et al., 2021; Glaser et al., 2018; Duarte et al., 2022).

Whereas prior work mainly targets specific statute titles or case names such as "Order no. 625 from 25 April 2019" (Pais et al., 2021), the expressions extracted in our work include generic nouns indicating the type of law (e.g., "Cabinet Order," and "Order of the Ministry of Land, Infrastructure, Transport and Tourism") and abstract expressions indicating the presence of a delegated matter (e.g., "as specified by the Minister of Economy, Trade and Industry"). This difference in the target expressions is one of the distinctions between our delegation extraction task and the existing NER research.

H.3 Question Answering

Research on QA in the legal domain includes systems that retrieve statutes or case law relevant to the input text. In statute-retrieval QA, the task is to take a legal problem as a query and retrieve the statutes or provisions necessary to answer the problem. For example, the French language dataset BSARD (Louis and Spanakis, 2022), constructed from Belgian legislation, and the Chinese language dataset STARD (Su et al., 2024), constructed from Chinese legislation, introduce QA tasks in which questions such as "Is it legal to contract a lifetime lease?" are given as input, and the system retrieves the statutory provisions relevant to answering them. Additionally, using Japanese statutes, the international competition COLIEE, which focuses on legal information extraction and textual entailment, includes a task where Japanese bar exam questions are used as queries to search for relevant provisions in the Japanese Civil Code[11] (Goebel et al., 2024).

Research that retrieves case law instead uses descriptions of the factual circumstances of a case as queries and retrieves precedents involving similar fact patterns (Feng et al., 2024). For instance, COLIEE also includes a task in which decisions from the Federal Court of Canada serve as queries, and the system retrieves other decisions related to them. In addition, LexCLiPR (Upadhya and T.y.s.s, 2025), a multilingual dataset constructed from decisions of the European Court of Human Rights, proposes a task in which case-law guides, which are expository documents that describe how case law is interpreted and applied, are used as queries to retrieve the decisions they analyze.

The above studies retrieve relevant statutes, provisions, or case law at the level of the entire input, and therefore do not capture strict, localized correspondences between specific parts of the input and specific parts of the retrieved results. In contrast, our work identifies locally grounded correspondences by extracting delegation keywords from the input provision and specifying the delegation target provision associated with each keyword. Capturing such local correspondences is essential for the precise interpretation of individual provisions and for deepening the understanding of legal systems.

I Japanese Source Text and English Translation

Tables 10–12 provide the original Japanese legal texts and their corresponding English translations referenced in this study. The original Japanese texts are retrieved from e-Gov Legislation Search (e-Gov 法令検索) maintained by the Digital Agency of Japan. The English translations were prepared by the authors with reference to the Japanese Law Translation Database System maintained by the Ministry of Justice of Japan. Where portions of the text were omitted for brevity, the symbol "…" is used to indicate such omissions. Indentation and other layout formatting are simplified.

[11] Both the problem statements and the legal provisions are provided in the original Japanese and in English translation.

Japanese	English
介護保険法 （指定都道府県事務受託法人） 第二十四条の三 . . . (6)　前各項に定めるもののほか、指定都道府県事務受託法人に関し必要な事項は、政令で定める。	Long-Term Care Insurance Act (Designated and Entrusted Juridical Person for Prefectural Affairs) Article 24-3 . . . (6)　In addition to the provisions as prescribed in each of the preceding paragraphs of this Article, other necessary matters pertaining to a Designated and Entrusted Juridical Person for Prefectural Affairs are prescribed by a Cabinet Order.
介護保険法施行令 （指定都道府県事務受託法人による報告） 第十一条の九　都道府県知事は、都道府県事務の適正な実施を確保するため必要があると認めるときは、その必要な限度で、指定都道府県事務受託法人に対し、報告を求めることができる。	Order for Enforcement of the Long-Term Care Insurance Act (Report by Designated and Entrusted Juridical Person for Prefectural Affairs) Article 11-9　When the prefectural governor finds it necessary for ensuring the proper implementation of prefectural affairs, within the extent necessary, the prefectural governor may require a Designated and Entrusted Juridical Person for Prefectural Affairs to submit a report.
国土交通省設置法 （地方整備局の事務所） 第三十二条　国土交通大臣は、地方整備局の所掌事務の一部を分掌させるため、所要の地に、地方整備局の事務所を置くことができる。 ２　地方整備局の事務所の名称、位置、管轄区域、所掌事務及び内部組織は、国土交通省令で定める。	Act for Establishment of the Ministry of Land, Infrastructure, Transport and Tourism (Offices of Regional Development Bureaus) Article 32 (1)　The Minister of Land, Infrastructure, Transport and Tourism may establish offices of the Regional Development Bureaus at necessary locations, in order to allot a part of the functions under the jurisdiction of the Bureaus. (2)　The names, locations, jurisdiction, functions under the jurisdiction, and organizational structures of the offices of the Regional Development Bureaus are provided for in Order of the Ministry of Land, Infrastructure, Transport and Tourism.
地方整備局組織規則 国土交通省設置法（平成十一年法律第百号）第三十二条第二項及び国土交通省組織令（平成十二年政令第二百五十五号）第二百八条第六項の規定に基づき、並びに同法及び同令を実施するため、地方整備局組織規則を次のように定める。	Regulations for Organization of the Regional Development Bureaus Pursuant to the provisions of Article 32 paragraph (2) of the Act for Establishment of the Ministry of Land, Infrastructure, Transport and Tourism (Act No. 100 of 1999) and to the provisions of Article 208 paragraph (6) of the Order for Organization of the Ministry of Land, Infrastructure, Transport and Tourism (Cabinet Order No. 255 of 2000), in order to enforce that Act and that Order, the Regulations for Organization of the Regional Development Bureaus is established as follows.
離島航路整備法 （航路損益計算書等の提出） 第八条　補助航路事業者は、国土交通省令の定めるところにより、当該離島航路に関する航路損益計算書その他の書類を国土交通大臣に提出しなければならない。	Act on Development of Traffic Routes to Remote Islands (Submission of Traffic Route Profit and Loss Statement) Article 8　A subsidized traffic route operator must submit a traffic route profit and loss statement concerning the traffic route to the remote islands and other documents to the Minister of Land, Infrastructure, Transport and Tourism, as provided for in Order of the Ministry of Land, Infrastructure, Transport and Tourism.
離島航路整備法 （施行規定） 第十九条　この法律に定めるもののほか、この法律の施行に関し必要な事項は、国土交通省令で定める。	Act on Development of Traffic Routes to Remote Islands (Provisions on Implementation) Article 19　Beyond what is provided for in this Act, necessary matters related to the enforcement of this Act are prescribed by Order of the Ministry of Land, Infrastructure, Transport and Tourism.
離島航路整備法施行規則 （航路損益計算書等の提出） 第四条　補助航路事業者は、航路ごとに、航路補助金の交付を受けようとする会計年度の九月三十日を末日とする一年間の航路損益計算書三通を作成し、これを当該年度の十一月三十日までに、当該航路の拠点を管轄する地方運輸局長を経由して国土交通大臣に提出するものとする。. . .	Regulations for Enforcement of the Act on Development of Traffic Routes to Remote Islands (Submission of Traffic Route Profit and Loss Statement) Article 4　For each traffic route, a subsidized traffic route operator shall prepare three copies of traffic route profit and loss statements covering the one year ending on September 30 of the fiscal year for which the operator seeks to obtain a traffic route subsidy, and submit these documents to the Minister of Land, Infrastructure, Transport and Tourism via the Director of the District Transport Bureau with jurisdiction over the base of the traffic route, by November 30 of the relevant fiscal year. . .

Table 10: Original Japanese legal texts and their English translations referenced in the main text

Japanese	English
学校教育法 第百二十九条第三項　専修学校の教員は、その担当する教育に関する専門的な知識又は技能に関し、文部科学大臣の定める資格を有する者でなければならない。	School Education Act Article 129 paragraph (3)　Teachers in specialized training colleges must have qualifications specified by the Minister of Education, Culture, Sports, Science and Technology, regarding specialized knowledge or skills of the education they are in charge of.
電波法 第七十九条第一項　総務大臣は、無線従事者が左の各号の一に該当するときは、その免許を取り消し、又は三箇月以内の期間を定めてその業務に従事することを停止することができる。	Radio Act Article 79 paragraph (1)　The Minister of Internal Affairs and Communications may revoke a radio operator's license, or order a radio operator to cease engaging in that service for a specified period not exceeding three months, if the radio operator falls under one of the following items:
子ども・子育て支援法 第七十一条の十四第三項　内閣総理大臣は、第一項の規定により支払基金に同項各号に掲げる事務の全部若しくは一部を行わせることとするとき又は支払基金に行わせていた当該事務の全部若しくは一部を行わせないこととするときは、その旨を公示しなければならない。	Child and Child Care Support Act Article 71-14 paragraph (3)　If the Prime Minister is to entrust the Payment Fund with all or a part of the administrative affairs listed in each item of the paragraph (1) pursuant to the provisions of that paragraph, or is to cease entrusting the Payment Fund with all or a part of the administrative affairs, the Prime Minister must issue public notice of this.

Table 11: Original Japanese legal texts and their English translations referenced in Appendix G.1

Japanese	English
労働安全衛生法 （登録の更新） 第四十六条の二　登録は、五年以上十年以内において政令で定める期間ごとにその更新を受けなければ、その期間の経過によつて、その効力を失う。…	Industrial Safety and Health Act (Renewal of Registrations) Article 46-2　If not renewed for every five- to ten- year period specified by Cabinet Order, a registration ceases to be effective upon the expiration of that period…
労働安全衛生法施行令 （登録製造時等検査機関等の登録の有効期間） 第十五条の二　法第四十六条の二第一項（法第五十三条の三から第五十四条の二までにおいて準用する場合を含む。）の政令で定める期間は、五年とする。	Order for Enforcement of Industrial Safety and Health Act (Valid Period of Registration for Registered Manufacturing Inspection, etc. Agency) Article 15-2　The period prescribed by the Cabinet Order set forth in the paragraph (1) of the Article 46-2 of the Act (including as applied mutatis mutandis pursuant to Article 53-3, Article 54 and Article 54-2) is for 5 years.
労働安全衛生法及びこれに基づく命令に係る登録及び指定に関する省令 （登録の更新） 第一条の二の四　登録は、五年ごとにその更新を受けなければ、その期間の経過によつて、その効力を失う。…	Ministerial Order on Registration and Designation Related to Industrial Safety and Health Act and Orders based on the Act (Renewal of Registrations) Article 1-2-4 (1)　Unless the registration is renewed every five years, it expires when that period has elapsed…
都市計画法 （政令への委任） 第八十八条　この法律に定めるもののほか、この法律の実施のため必要な事項は、政令で定める。	City Planning Act (Delegation to Cabinet Order) Article 88　In addition to what is provided for in this Act, matters necessary for enforcement of this Act are prescribed by Cabinet Order.
都市計画法施行令 内閣は、都市計画法(昭和四十三年法律第百号)及び都市計画法施行法(昭和四十三年法律第百一号)の規定に基づき、この政令を制定する。	Order for Enforcement of the City Planning Act The Cabinet hereby enacts this Cabinet Order pursuant to the provisions of the City Planning Act (Act No. 100 of 1968) and the Act for Enforcement of the City Planning Act (Act No. 101 of 1968).
都市計画法施行令 （都に関する特例） 第四十六条　法第八十七条の三第一項の政令で定める都市計画は、法第十五条の規定により市町村が定めるべき都市計画のうち、次に掲げるものに関する都市計画とする…	Order for Enforcement of the City Planning Act (Special Provisions regarding Tokyo Metropolis) Article 46 (1)　The city plans to be specified by a Cabinet Order as prescribed in Article 87-3, paragraph (1) of the Act are the following city plans that a municipality is to define pursuant to the provisions of Article 15 of the Act: …
航路標識法 （航路標識に事故が発生した場合の報告義務） 第十五条　第十一条第一項の許可を受けた者は、当該許可に係る航路標識について破損その他の事故が発生し、当該航路標識の現状に変更があつたときは、国土交通省令で定めるところにより、直ちに、その旨を海上保安庁長官に報告しなければならない。	Navigation Aids Act (Duty to Report upon Accident Involving Navigation Aids) Article 15　If Navigation Aids related to permission referred to in Article 11, paragraph (1) have been damaged or involved in any other accidents, and the existing state of the Navigation Aids has been changed, the person who has obtained the permission must immediately report it to the Commandant of the Japan Coast Guard pursuant to the provisions of Order of the Ministry of Land, Infrastructure, Transport and Tourism.
航路標識法施行規則 （事故が発生した場合の報告） 第十二条　法第十五条の規定による報告は、電話、ファクシミリ装置その他なるべく早く到着するような手段によらなければならない。 2 海上保安庁長官は、前項の報告があったときは、必要と認める書類の提出を命ずることができる。	Regulations for Enforcement of the Navigation Aids Act (Reporting When There Has Been an Accident) Article 12 (1)　A report under Article 15 of the Act must be made by telephone, facsimile machine, or other means that will arrive as soon as possible. (2)　When receiving the report stated in the preceding paragraph, the Commandant of the Japan Coast Guard may order the submission of documents deemed necessary.

Table 12: Original Japanese legal texts and their English translations referenced in Appendix G.2