The Review of Socionetwork Strategies (2024) 18:123–143
https://doi.org/10.1007/s12626-024-00155-5

ARTICLE

Contribution Analysis of Large Language Models and Data Augmentations for Person Names in Solving Legal Bar Examination at COLIEE 2023

Takaaki Onaga[1] · Masaki Fujita[1] · Yoshinobu Kano[1]

Received: 5 September 2023 / Accepted: 11 January 2024 / Published online: 8 March 2024

Abstract

This paper describes our system for COLIEE 2023 Task 4, which automatically answers Japanese legal bar exam problems. We propose an extension to our previous system in COLIEE 2022, which achieved the highest accuracy among all submissions using data augmentation. We focus on problems that include mentions of person names. In this paper, we present two main contributions. First, we incorporate LUKE as our deep learning component, which is a named entity recognition model trained on RoBERTa. Second, we fine-tune the pretrained LUKE model in multiple ways, comparing fine-tuning on training datasets that include alphabetical person names and ensembling different fine-tuning models. We confirmed that LUKE and its fine-tuned model on person type problems improve their accuracies. Our formal run results show that LUKE and our fine-tuning approach using alphabetical person names were effective, achieving an accuracy of 0.69 in the COLIEE 2023 Task 4 formal run.

Keywords COLIEE 2023 · Question answering · Legal information entailment · Predicate argument structure analysis

✉ Yoshinobu Kano
kano@kanolab.net

Takaaki Onaga
tonaga@kanolab.net

Masaki Fujita
mfujita@kanolab.net

[1] Faculty of Informatics, Shizuoka University, Hamamatsu, Shizuoka, Japan

1 Introduction

Competition for Legal Information Extraction (COLIEE) is an annual international competition held in conjunction with the International Conference on Artificial Intelligence and Law (ICAIL) and Juris-informatics (JURISIN) conferences [1, 5–9, 12–14]. COLIEE 2023 consists of four tasks: Tasks 1 and 2 are case law tasks that use datasets from the Canadian Federal Court, while Tasks 3 and 4 are statute law tasks that use the Japanese Legal Bar exam. In Task 3, a participant system is given a problem text and asked to retrieve relevant articles from Japanese Civil Law to solve the problem. In Task 4, a participant system is given a problem text and its relevant articles, and asked to determine whether the articles entail the problem text or not by answering *Yes* or *No*. We participated in Task 4. The analysis of problem types in previous COLIEE tasks [13] showed that the COLIEE dataset includes diverse types of problems. Some are relatively easy to solve, because the texts in the pairs are very similar, while others are complex and difficult, requiring parsing, semantics, anaphora, logic, etc. Previous Task 4 participant systems have included rule-based and deep learning-based systems, such as BERT [19], ELECTRA [11], and GNN [17]. However, previous systems have not performed well on problems that require inferences about person roles. In this paper, we focus on person name resolution, where person names/roles are represented using alphabetical letters. We propose a system that extends our previous system in COLIEE 2022, which achieved the highest accuracy among all submissions using data augmentations. Our proposed system provides two main contributions. First, while we use an ensemble of a rule-based component and a deep learning-based component, we adopt LUKE as our deep learning-based component, which is a named entity trained model based on RoBERTa, instead of BERT. Second, we fine-tune the pretrained LUKE model in multiple ways, comparing fine-tuned training datasets that include alphabetical person names and an ensemble of different fine-tuned models. Our formal run results show that LUKE and our fine-tuning approach for alphabetical person names are effective.

2 Related Works

LUKE [18] is a language model based on RoBERTa [10], which is a derivative of BERT [2]. BERT is a deep learning model that is commonly used in various NLP tasks, and it utilizes the encoder part of the Transformer [16] architecture. LUKE, on the other hand, uses a unique mechanism called Entity-aware Self-attention. LUKE treats not only words, but also entities as independent tokens, and computes intermediate and output representations for all tokens using the Transformer (Fig. 1). Since entities are treated as tokens, LUKE can directly model the relationships between entities. In this paper, we focus on the person type problems which include named entities of persons, thus LUKE is expected to work well with these issues. Furthermore, at the time of its development, LUKE achieved

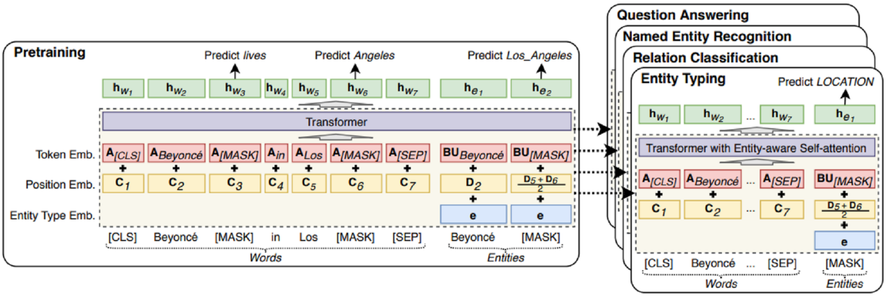

Fig. 1 Architecture of LUKE using the input sentence "Beyonce lives in Los Angeles." LUKE outputs contextualized representation for each word and entity in the text. The model is trained to predict randomly masked words (e.g., lives and Angeles in the figure) and entities (e.g., Los Angeles in the figure). Downstream tasks are solved using its output representations with linear classifiers. Cited from [18]

the highest accuracy in several NLP tasks. We adopt LUKE as the base model and fine-tune the pretrained LUKE model.

Hoshino et al. [4] is our previous work presented in COLIEE 2019. They proposed a rule-based system that parses sentences into clauses based on their original definition. The parsing results were then used to extract the set of clauses, including subject, predicate, and object for each clause, and compared these sets. They developed several modules, such as the Precise Match module, which compared the relevant civil law clauses with the clause set of the problem text and answered *Yes* if all the elements in the clause sets matched. Fujita et al. [3] is another recent work of ours in COLIEE 2022, which proposed an ensemble of the rule-based system developed by Hoshino et al.'s rule-based system and a BERT-based system. This system achieved the highest accuracy in the formal run of COLIEE 2022 Task 4. To address the issue of limited training data, we performed data augmentation such as logical inversion, replacement of person terms, and replacement of article numbers. In this paper, we extended our previous system by replacing BERT with LUKE and modifying the ensemble method to build different fine-tuned models depending on the type of problem.

3 System

3.1 System Overview

Our system comprises a rule-based component and an LUKE-based component. The LUKE-based component utilizes an LUKE model, which is fine-tuned on three different datasets: all training datasets provided by COLIEE, and two types of training datasets extracted from different problem types. The rule-based and LUKE-based components are integrated through ensemble, which performs binary classification, predicting either *Yes* or *No* based on the higher probability value. In the COLIEE Task 4 dataset, alphabetical characters are used to represent persons in the problem text, as illustrated in Fig. 2, which shows an example of a problem involving

<id="R02-4-I", label="Y">
<article>
A person who has contracted as an agent of another person shall be liable to the other party for performance or damages at the other party's option, unless he has proved his own agency or has obtained his own additional authorization.
<problem>
A, purporting to be B's agent, entered into a purchase agreement with C to sell land owned by B to C, but did not actually have the agency to enter into the agreement; if B ratified the purchase agreement, A is not liable to C as an unauthorized agent.

<関連条文>
他人の代理人として契約をした者は、自己の代理権を証明したとき、又は本人の追認を得たときを除き、相手方の選択に従い、相手方に対して履行又は損害賠償の責任を負う。
<問題文>
Aは， B の代理人と称して， C との間で B の所有する土地を C に売却する旨の売買契約を締結したが，実際にはその契約を締結する代理権を有していなかった。 B が売買契約を追認した場合， A は C に対する無権代理人の責任を負わない。

Fig. 2 An example of a problem where alphabetical person characters appears

alphabetical person characters. It is necessary to determine the relationship between each person indicated by an alphabetical character and the person role described in the civil law text. In the example, A in the problem text represents a person who contracted as an agent of another person, B represents a different person, and C corresponds to a counterparty, as defined in the civil code text. Such problems are considered to be among the most challenging to solve automatically.

We focus on problems that involve alphabetical person names, and create separate LUKE models trained on such problems and trained on other problems. For the LUKE-based part, we prepare three LUKE models for comparison: an LUKE model trained on all data (**LUKE-all**), an LUKE model trained on problems with alphabetical person names (**LUKE-person**), and an LUKE model trained on problems without alphabetical person names (**LUKE-nonperson**).

While our previous system [4] had different modules with different matching methods for the clause sets, our previous study [3] showed that the Precise Match module was the most effective, answering *Yes* only when all pairs of subjects, objects, and predicates match. Therefore, we adopt the Precise Match module as our rule-based part. We fine-tuned a publicly available LUKE model (studio-ousia/luke-japanese-base-lite[1]) which was pretrained on Wikipedia articles, to output binary probabilities of *Yes* or *No*, given a problem text and a relevant civil law article as input.

In this section, we describe the design of our system as follows. First, we create additional training data using civil law articles (3.2). Second, after preprocessing the data, we select the most relevant civil law article for solving a given problem statement, based on the similarity of their texts (3.3). Third, we expand the training data by performing logical inversion and replacing person terms (3.4). Fourth, we fine-tune the LUKE model using these datasets. We split the datasets by year and create multiple models for all possible combinations of the training and validation datasets

[1] https://huggingface.co/studio-ousia/luke-japanese-base-lite.

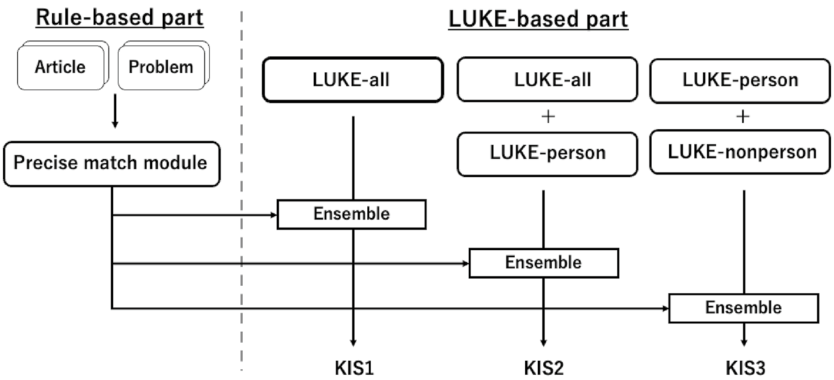

Fig. 3 System overview

(3.5). Based on the methods above, we created three different submission models for our formal run results: **KIS1**, **KIS2**, and **KIS3**, which were designed for different types of problems (3.6). Among the three formal run submissions, **KIS2** was our proposed system. **KIS1** was an ensemble of an LUKE-based model using all of the training data and the rule-based system. **KIS2** was an ensemble of **KIS1** and a model trained specifically for problems in which alphabetical person names appear. **KIS3** was an ensemble of a model trained specifically for problems in which alphabetical person names appear and a model trained specifically for problems in which they do not appear. Figure 3 illustrates these relationships. We applied our article selection preprocess (3.3) to the formal run test dataset.

3.2 Create Training Data from Article(s)

To increase the size of the official training dataset, we created an additional training dataset using the civil code articles without problem texts. In this subsection, we will refer to the relevant articles in COLIEE as premise (t1) and the problem text in COLIEE as hypothesis (t2) to avoid confusion, since both are taken from the articles. First, we divided the distributed civil law articles into sections and created pairs of identical civil code sections, setting their correct answer labels to *Yes*. For example, "A minor must obtain the consent of his/her legal representative to perform a legal act. However, this shall not apply to acts merely to obtain rights or to be relieved of obligations. (Civil Code Article 5)" and the same paragraph is paired with the label *Yes*. If the text of the article contains an exception sentence or proviso, such as "Provided, however, [...], this shall not apply.", we divided the original article texts into a text before the sentence (a principle part) and after the sentence (a proviso part). If "However, [...], this shall not apply" describes an act, person, or right, we manually replace that act, person, or right in the principle part with an act, person, or right in the proviso part. Then, we invert the logic of the predicate as described in 3.4. In the example in Fig. 4, Article 5 of the Civil Code "However, this shall not apply to acts by which a minor merely acquires a right or is relieved of a duty." was rewritten as "A minor need not obtain the consent of his or her legal

<Article 5 of the Civil Code (Original)>
A minor must obtain the consent of the minor's legal representative to perform a juridical act. provided, however, that this does not apply to a juridical act for merely acquiring a right or being released from an obligation.
・Split by "however."
<principle part>
A minor must obtain in the consent of his/her legal representative to perform a legal act.
<exception part>
A minor need not obtain the consent of his or her legal representative for a juridical act for merely acquiring a right or being released from an obligation.

<民法第 5 条 (原文)>
未成年者が法律行為をするには、その法定代理人の同意を得なければならない。 ただし、単に権利を得、又は義務を免れる法律行為については、この限りでない。
・「ただし、」で分割する。
<原則部分>
未成年者が法律行為をするには、その法定代理人の同意を得なければならない。
<例外部分>
未成年者が、単に権利を得、又は義務を免れる法律行為をするには、その法定代理人の同意を得なくてもよい。

Fig. 4 Divide into principle and exception

Pair 1: <t1>and <t2>are identical, the original text of Article 5 of the Civil Code.
<t1>
A minor must obtain the consent of the minor's legal representative to perform a juridical act. provided, however, that this does not apply to a juridical act for merely acquiring a right or being released from an obligation.
未成年者が法律行為をするには、その法定代理人の同意を得なければならない。ただし、単に権利を得、又は義務を免れる法律行為については、この限りでない。
<t2>
A minor must obtain the consent of the minor's legal representative to perform a juridical act. provided, however, that this does not apply to a juridical act for merely acquiring a right or being released from an obligation.
未成年者が法律行為をするには、その法定代理人の同意を得なければならない。ただし、単に権利を得、又は義務を免れる法律行為については、この限りでない。

Pair 2: Proviso part is not needed to solve the problem.
<t1>
Same as pair 1' s <t1>
<t2>
A minor must obtain the consent of his/her legal representative to perform a legal act.
未成年者が法律行為をするには、その法定代理人の同意を得なければならない。

Pair 3: Proviso part is needed to solve the problem.
<t1>
Same as pair 1' s <t1>
<t2>
A minor need not obtain the consent of his or her legal representative for a juridical act for merely acquiring a right or being released from an obligation.
未成年者が、 単に権利を得、又は義務を免れる法律行為をするには、その法定代理人の同意を得なくてもよい。

Fig. 5 < t1 >< t2 > Pairs created using exceptions representative to commit an act merely to obtain a right or to be relieved of a duty." The subject normally appears in the principle part, but sometimes it appears in the proviso part. When the subject appears in the proviso part, we revert the affirmative/negation of the principle part using the method described later (3.4) and add it to the training dataset, sharing the same original premise (t1). Figure 5 shows an example.

3.3 Preprocess and Article Selection

First, we apply the following preprocessing steps to the articles and then select the relevant ones. A problem statement may have multiple related articles. If we concatenate the texts of all these articles as input, the input to the model may become too long, exceeding the upper limit (in our case, 512 tokens), and important parts may be lost when we truncate the input. To address this issue, we split the relevant articles into sections (each article consists of one or more sections). Then, we create all possible combinations of the divided sections (Fig. 6). We discard any combination in which the total number of tokens of the combined sections and the given problem text exceeds the upper limit.

If the generated text contains reference notations such as "preceding paragraph" or "Article XX", we search the given relevant articles for the referred article and replace the reference notations with the text from the referred article (as shown in Fig. 7). The replaced version is then added to the training dataset. Notations such as

<id="R01-3-E", label="Y">
<article>
1, Article 106 A sub agent shall represent the principal with respect to acts within his/her authority.
2,(2) A sub agent shall have the same rights and assume the same obligations as an agent with respect to the principal and third parties within the scope o f his/her authority.
・Generate combinations for each divided term
1, A sub agent shall represent the principal with respect to acts within his/her authority.
2, A sub agent shall have the same rights and assume the same obligations as an agent with respect to the principal and third parties within the scope of his/her authority.
1 + 2, A sub agent shall represent the principal with respect to acts within his/her authority. A sub agent shall have the same rights and assume the same obligations as an agent with respect to the principal and third parties within the scope of his/her authority.

<関連条文>
1, 第百六条 復代理人は、その権限内の行為について、本人を代表する。
2, 2 復代理人は、本人及び第三者に対して、その権限の範囲内において、代理人と同一の権利を有し、義務を負う。
・項ごとの組合せを生成する
1, 復代理人は、その権限内の行為について、本人を代表する。
2, 復代理人は、本人及び第三者に対して、その権限の範囲内において、代理人と同一の権利を有し、義務を負う。
1+2, 復代理人は、その権限内の行為について、本人を代表する。復代理人は、本人及び第三者に対して、その権限の範囲内において、代理人と同一の権利を有し、義務を負う。

Fig. 6 An example of combinations reconstruction

<id=” R03-5-A” , label=” N” >
<article>
Article 150
When a demand is made, the prescription shall not be completed until six months have elapsed from that time.
(2) Another demand made while the completion of prescription is postponed by a demand shall not have the effect of postponing the completion of prescription under the preceding paragraph.
・Substitute the first paragraph for **“preceding paragraph”** in the second paragraph.
Another demand made while the completion of prescription is postponed by a demand shall not have the effect of postponing the completion of prescription under the paragraph that state: “When a demand is made, the prescription shall not be completed until six months have elapsed from that time.”

<関連条文>
第百五十条
催告があったときは、その時から六箇月を経過するまでの間は、時効は、完成しない。
２ 催告によって時効の完成が猶予されている間にされた再度の催告は、 前項 の規定による時効の完成猶予の効力を有しない。
・**“前項”** を置き換える。
催告によって時効の完成が猶予されている間にされた再度の催告は、「催告があったときは、その時から六箇月を経過するまでの間は、時効は、完成しない。」の規定による時効の完成猶予の効力を有しない。

Fig. 7 An example of article reference

<pair id=“H30-4-I”, label=“N”>
<article>
Article 103 An agent without prescribed authority shall have the authority to perform only the following acts
(i) acts of preservation
(ii) acts for the purpose of utilizing or improving the object or right for which the agent is acting, to the extent that the nature of such object or right is not changed
・Substitute each item for “the following acts”.
(i) An agent with no defined authority is authorized only to perform acts of preservation.
(ii) An agent without prescribed authority is authorized only to perform acts for the purpose of using or improving the thing or right that is the object of the representation, to the extent that the nature of the thing or right is not changed.

<関連条文>
第百三条
権限の定めのない代理人は、 次に掲げる行為のみをする権限を有する。
一 保存行為
二 代理の目的である物又は権利の性質を変えない範囲内において、その利用又は改良を目的とする行為
・次に掲げる に各号を代入する。
(一) 権限の定めのない代理人は、 保存行為のみをする権限を有する。
(二) 権限の定めのない代理人は、 代理の目的である物又は権利の性質を変えない範囲内において、その利用又は改良を目的とする行為 のみをする権限を有する。

Fig. 8 An example of substituting each item for “lited below”

<id=” R02-5-U” , label=” Y” >
<article>
Article 724
The right to claim damages for a tort shall be extinguished by prescription in the following cases
(i) When the victim or his/her legal representative has not exercised it for three years from the time when he/she became aware of the damage and the perpetrator
(ii) When the right is not exercised for 20 years from the time of the tortious act.
<problem>
The right to claim damages based on a tort shall be extinguished by prescription if not exercised for 20 years from the time of the tortious act.

・Find the similarity between the problem statement and the article paragraph by paragraph.
1, **The right to claim damages in tort shall be extinguished by prescription if not exercised for twenty years from the time of the tortious act. (The similarity of this statement was the highest.)**
2, The right to claim damages for a tort shall be extinguished by prescription if the victim or his/her legal representative does not exercise the right for three years from the time when he/she learned of the damage and the perpetrator.

<関連条文>
第七百二十四条
不法行為による損害賠償の請求権は、次に掲げる場合には、時効によって消滅する。
一 被害者又はその法定代理人が損害及び加害者を知った時から三年間行使しないとき。
二 不法行為の時から二十年間行使しないとき。
<問題>
不法行為に基づく損害賠償請求権は，不法行為の時から２０年間行使しない場合，時効によって消滅する。
・問題文と条文の組合せごとの類似度を求める。
1，不法行為による損害賠償の請求権は、不法行為の時から二十年間行使しないときには、時効によって消滅する。(この文の類似度が最も高くなった。)
2, 不法行為による損害賠償の請求権は、被害者又はその法定代理人が損害及び加害者を知った時から三年間行使しないときには、時効によって消滅する。

Fig. 9 An example of article selection

“listed below” are substituted with the specified items in the article. Figure 8 provides an example of this process.

As shown in Fig. 4, the proviso part of an article describes an exceptional situation where the principle part does not apply. To understand the meaning of the proviso part, we need to include the principle part as well. Therefore, we concatenate the proviso part with its principle part, inverting the affirmation/negation of the latter. If the proviso part includes an act, person, or right, we replace the corresponding item in the principle part with the one in the proviso part. Among these preprocessed articles, we select most relevant article to solve the given problem by the similarity scores of the vectors obtained by Sentence Luke (sonoisa/sentence-luke-japanese-base-lite[2]). Sentence LUKE is a tool for creating advanced sentence vectors using the LUKE model (LUKE version of the Sentence BERT [15] in other words), which was pretrained by the Japanese Wikipedia and the Siamese network. We remove the

2 https://huggingface.co/sonoisa/sentence-luke-japanese-base-lite.

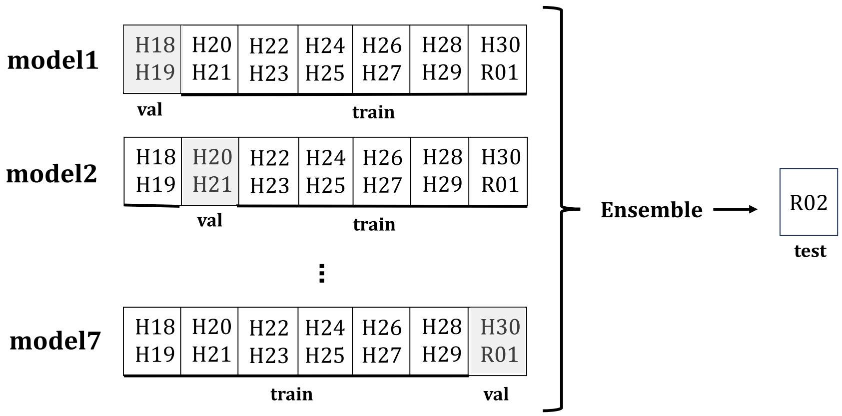

Fig. 10 A conceptual figure of training data split

suffixes of predicates, which could contain negation expressions. This is because we search for the most similar content regardless of affirmative/negative. Figure 9 shows an example.

3.4 Data Augmentation

Our previous COLIEE 2022 system [3] consisted of two expansions: negation expansion and person term replacement, which we describe below. In this year's formal run, we have added more negative words and person terms to our manual dictionary. For negation expansion, we create a new sample by reversing the logic at the end of a sentence, along with its *Yes* or *No* answers, using a predefined list of affirmative and negation expression pairs. We apply this expansion to both pairs created from the Civil Code articles as described in the previous sections and the given problem text. However, we do not apply this expansion to problems with a gold standard answer of *No*, since the negative form at the end of a sentence does not always result in a *Yes* when the original answer is *No*. The COLIEE problems sometimes use alphabetical characters, such as A or B, to represent person names. Our person term replacement expansion addresses this issue by creating a dataset from the training data that replaces person names with alphabetical characters. We assign the alphabetical letters in the order of appearance, holding identical person names to be identical characters.

3.5 Combinatorial Split of Training and Validation Dataset

To fully utilize the COLIEE official training dataset, we created multiple models trained with different parts of the official dataset. We split the official dataset using various patterns, such as the cross-validation method, where we selected each 2-year period as a validation dataset and used the rest of the official dataset as its training dataset. After fine-tuning for each pattern, we applied an ensemble of these multiple models. We chose 2 years as our splitting unit, because it would be too many combinations if we split by year. Figure 10 illustrates this split method.

3.6 Fine-Tune for Alphabetical Person Names

When alphabetical letters are used as person names in the given problem text, a different approach is required to solve the problem, as it becomes necessary to determine which person the alphabetical character corresponds to in the relevant civil law article. Therefore, we fine-tune a model specifically for such problems. Additionally, we fine-tune a model for problems in which alphabetical person names do not appear. Each model internally performs an ensemble of the combinatorial split fine-tunes described in Sect. 3.5, and thus, the preprocessing steps described in Sects. 3.1 to 3.4 are applied before the fine-tuning. We regard a problem as an alphabetical person name type problem if it contains any single alphabetical character (as the original text is in Japanese except for these characters). As mentioned earlier, **KIS2** and **KIS3** use the model fine-tuned with problems containing alphabetical characters, while **KIS1** uses the model fine-tuned without them. During binary classification, a fully connected linear transformation is performed on the output of the last layer's node corresponding to the "<s>" token (or the "[CLS]" token in the case of BERT) for both *Yes* and *No* answers. Then, the classification scores are compared to determine whether the answer is *Yes* or *No*. For fine-tuning, the classification scores are converted into probabilities for each label using the Softmax function, and the loss is calculated using cross-entropy.

3.7 Ensemble Prediction

Finally, we perform an ensemble of our rule-based part and our LUKE-based part. The rule-based (precise match module) is the same as in our previous work, which has high precision but a low number of answerable problems. Therefore, we first apply the rule-based part when applicable, and then apply the LUKE-based part when the rule-based part is not applicable. For the LUKE-based part, we have prepared three models: **LUKE-all** (fine-tuned on all of our datasets), **LUKE-person** (fine-tuned on problems with alphabetical person names), and **LUKE-nonperson** (fine-tuned on problems without alphabetical person names). **KIS3** applies **LUKE-person** when the problem includes alphabetical person names and applies **LUKE-nonperson** when the problem does not include any alphabetical person names. Similarly, **KIS2** applies **LUKE-person** in the same way but uses **LUKE-all** when the problem does not include any alphabetical person names. If the rule-based part is not applicable, KIS always applies **LUKE-all**.

Table 1 COLIEE 2023 Task 4's formal run results for each participant's submission. # represents the number of correct answers; Acc represents Accuracy. The submission IDs in bold are our submissions

Submission ID	#	Acc
Total	101	N/A
BaseLine	52	0.5149
AMHR02	82	0.8119
JNLP3	79	0.7822
TRLABS_D	79	0.7822
TRLABS_I	79	0.7822
JNLP1	76	0.7525
JNLP2	76	0.7525
TRLABS_T	76	0.7525
KIS2	**70**	**0.6931**
KIS1	**68**	**0.6733**
UA_V2	67	0.6634
AMHR01	66	0.6535
KIS3	**66**	**0.6535**
AMHR03	65	0.6436
LLNTUdulcsL	63	0.6238
UA	63	0.6238
HUKB2	60	0.5941
CAPTAIN.gen	59	0.5842
CAPTAIN.run1	58	0.5743
LLNTUdulcsS	57	0.5644
HUKB1	56	0.5545
HUKB3	56	0.5545
LLNTUdulcsO	56	0.5545
NOWJ.multi-v1-jp	55	0.5446
CAPTAIN.run2	53	0.5248
NOWJ.multijp	53	0.5248
NOWJ.multi-v1-en	49	0.4851

Table 2 Numbers of correct answers and accuracies in previous formal run datasets

	#	KIS1	KIS2	KIS3
H30 dataset	70	**44 (0.62)**	**44 (0.62)**	43 (0.61)
R01 dataset	111	75 (0.67)	**77 (0.69)**	73 (0.65)
R02 dataset	81	56 (0.69)	**58 (0.71)**	52(0.62)

4 Experiments and Results

4.1 Fine-Tune Parameters

We performed our fine-tuning with the following parameters: maximum tokens length of 512, batch size of 32, learning rate of 1e-5, and a maximum number of epochs of 10 but terminates early due to Early Stopping.

Table 3 The number of correct answers by problem types (**person**: problems of person type, **nonperson**: others, **all**: including both of **person** and **nonperson**), for BERT and LUKE

Dataset	H30			R01			R02			R04			Total		
	B	L	#	B	L	#	B	L	#	B	L	#	B	L	#
All	42	**44**	70	**73**	72	111	56	56	81	61	**66**	101	232	**238**	363
Person	7	7	13	20	20	34	**23**	22	35	27	**29**	41	77	**78**	123
Nonperson	35	35	57	**53**	52	77	33	**34**	46	34	**37**	60	155	**158**	240

The column names of "B", "L", and # stand for BERT, LUKE, and the number of each problem type, respectively

Table 4 The number of correct answers by problem types (other than **person**) for BERT and LUKE

Dataset	H30			R01			R02			R04			Total		
	B	L	#	B	L	#	B	L	#	B	L	#	B	L	#
Condition	32	**36**	55	**44**	43	74	18	18	31	49	**51**	69	143	**148**	229
Persons role	21	**26**	44	28	**29**	48	21	**22**	33	43	**46**	53	113	**123**	178
Person relation	21	**26**	44	28	**29**	48	21	**22**	33	49	**50**	63	119	**127**	188
Morpheme	8	8	11	25	**26**	32	1	1	1	8	**11**	13	42	**46**	57
Anaphora	8	**11**	15	15	**16**	23	8	**10**	15	30	**37**	42	61	**74**	95
Case role	8	**10**	16	**9**	8	14	**1**	0	2	**6**	5	8	**24**	23	40
Verb paraphrases	11	11	20	26	**27**	41	**9**	8	10	9	**10**	11	55	**56**	82
Normal terms	5	5	7	9	**10**	16	11	11	21	1	1	4	**26**	25	48
Predicate argument	5	**7**	10	46	**49**	72	12	12	17	**8**	7	12	71	**75**	111
Negation	21	21	32	41	**44**	68	9	9	16	**25**	24	27	96	**98**	143
Legal fact	21	21	40	37	**39**	67	10	10	14	15	15	26	83	**85**	147
Entailment	1	**2**	3	**6**	5	9	1	1	1	7	**9**	12	15	**17**	25
Dependency	6	**9**	16	18	18	24	5	5	10	2	**3**	3	31	**35**	53
Article search	3	**4**	5	6	6	14	2	2	3	16	**17**	19	27	**29**	41
Paraphrases	1	1	8	1	**2**	3	1	1	1	8	**10**	11	11	**14**	23
Itemized	1	1	3	4	**6**	11	**3**	2	5	3	**4**	5	11	**13**	26
Calculation	1	1	2	1	0	1	3	**4**	4	4	4	4	9	9	11

The column names of "B", "L", and # stand for BERT, LUKE, and the number of each problem type, respectively

4.2 COLIEE 2023 Formal Run Results

Table 1 shows the results of all teams in the COLIEE 2023 Task 4's formal run, where KIS is our team name.

4.3 Previous COLIEEs' Formal Run results

Table 2 shows the results of our experiments using previous formal runs of COLIEE 2019, 2020, and 2021 (test datasets are H30, R01, and R02, respectively) as required by the organizers.

4.4 Comparison of BERT and LUKE

Table 3 shows the results of the experiments on the formal run and the past formal runs using BERT and LUKE. Each cell shows numbers of correct answers with total numbers of problems from H30 to R04; the **all** column shows the total numbers, the **person** column shows the numbers for problems containing characters of the alphabetical person names, and the **nonperson** column shows the numbers for problems without the alphabetical person names (Table 4). The results of this table show that the correct numbers of the LUKE model is larger than the BERT model in H30 and R04. Especially in R04, LUKE improved the performance of the alphabetical person names problems. On the other hand, BERT had higher performance in R01 and similar performance in R02.

4.5 Evaluation of Fine-Tune Models Without Ensemble Using Previous Formal Runs

Table 5 shows the evaluation results of the individual fine-tuned LUKE models on the formal run of COLIEE 2023 and the formal runs of the past three years. Each fine-tuned model was evaluated independently without any ensemble. We evaluated the models separately for the problems with alphabetical person names (**person**) and others (**nonperson**). The results show that the **person** model, which is fine-tuned by person type problems, worked better than other models in all of the datasets.

Table 5 Number of correct answers of three patterns of fine-tuned LUKE models (**all**, **person**, and **nonperson**), for each training/test datasets (H30, R01, R02, and R04), dividing into person type problems (**P**) and others (**N**)

Fine-tune model		Total	All	Person	Nonperson
H30	N	57	37	35	35
	P	13	7	**8**	5
R01	N	77	52	51	51
	P	34	20	**23**	18
R02	N	46	34	32	31
	P	35	22	**25**	22
R04	N	60	37	36	34
	P	41	29	**31**	27

5 Discussion

The individual results of the fine-tuned models (Table 5) demonstrate that the fine-tuning was effective for the corresponding type of problems but not for the other types. Our team's formal run results (Table 1) and the results of our experiments using past formal runs (Table 2) also showed that **KIS2**, which is an ensemble using the fine-tuned model for alphabetical person names, achieved the highest score. Table 3 shows that LUKE and BERT have different percentages of correct answers. We analyzed the patterns in which either LUKE or BERT answered problems correctly. Figure 11 shows an example problem that can be answered without analyzing the alphabetical person names, even though they appear in the problem text. Such problems could be correctly answered by BERT. As shown in Fig. 12, R04–08-A is an example of a person name problem where LUKE was correct and BERT was incorrect. In this problem, the gold label is "*No*", because "B consented to this" in the problem text is different from "a third party consented to this" in the article, since B is an agent and C is a third party. LUKE was able to predict that the label for this problem would be "*No*". This example suggests that LUKE might be more proficient in understanding personal relationships compared to BERT. While

<pair id="R04-04-I" label="Y">
<problem>
A, a person under curatorship, may not rescind a judicial act by A after the legal representative ratifies the judicial act.

<問題文>
被保佐人Aがした法律行為を法定代理人が追認したときは、Aは、以後、その法律行為を取り消すことができない。

Fig. 11 An example problem which can be solved without analyzing the alphabetical person names

<pair id="R04-08-A" label="N">
<article>
Article 178, An assignment of a real right over movables may not be asserted against a third party without delivery of the movables.
Article 184, In cases of possession by an agent, if the principal has ordered the agent to take possession of the thing for a third party thereafter and the third party has consented thereto, the third party shall acquire the right of possession.
第百七十八条 動産に関する物権の譲渡は、その動産の引渡しがなければ、第三者に対抗することができない。
第百八十四条 代理人によって占有をする場合において、本人がその代理人に対して以後第三者のためにその物を占有することを命じ、その第三者がこれを承諾したときは、その第三者は、占有権を取得する。
<problem>
If A sells to C a painting A owned by A while leaving it with B, and A orders B to take possession of A for C thereafter, and B agrees to this, C may assert against the third party the acquisition of the ownership of A.
Aがその所有する絵画甲をBに預けたままCに売却した場合において、AがBに対して以後Cのために甲を占有すべきことを命じ、Bがこれを承諾したときは、Cは、甲の所有権の取得を第三者に対抗することができる。

Fig. 12 An example of a problem where LUKE provided the correct answer

LUKE itself slightly improved its performance compared with BERT (Table 3), LUKE works significantly better when fine-tuned with the person type problems, which corresponds to the highlighted cells in Table 5; the person type problems (**P**) were better solved by the **person** fine-tuned model than other models in any case. By manually checking the problems, we found that among 13 problems, that were correctly answered by LUKE and its **person** fine-tuned model than BERT, 11 problems were the type of the above explanation, which require to analyze the alphabetical person names.

We analyzed the results of our article selection by Sentence LUKE and found an unsuccessful example shown in Fig. 13. In this example, our system selected Article 5, "A minor shall obtain the consent of his/her legal representative in order to perform a legal act. Any legal act contrary to the provisions of the preceding paragraph may be revoked", while Article 124-2, item 2 was required to solve the problem. The non-relevant article our system selected shares similar tokens with the problem text, such as "minor" and "consent", but the relevant article also shares these tokens. This may be because abstract paraphrases like "Any legal act contrary to the provisions of the preceding paragraph may be revoke" make the cosine similarities larger. Pretraining and fine-tuning on legal documents and paraphrase preprocessing into everyday language may help improve this issue.

Next, we compare the extent to which our three data extension methods have contributed to improve the accuracy of the model. Our three data extension methods applied to the training data augmentations include: (i) the data created from civil law articles described in Sect. 3.2, (ii) the negation expansion, and (iii) the person term replacement described in Sect. 3.4. As a comparison analysis, We applied one of the three data extension methods before fine-tune the BERT model. We also compare the fine-tuned BERT models with all three extensions (our proposed model), and without any of the three extensions, thus five patterns in total. We use the dataset of each of the years H30, R01, R02, and R04 for evaluation, use the dataset of years prior to the year used in evaluation for training; these training-evaluation pairs correspond to the past formal run settings. Within the training dataset, we performed 11-fold cross-validations, resulting in 11 fine-tuned models. Our final prediction results are decided by majority votes between these 11 models. Using our human-created problem type classifications, we counted the number of correctly answered problems for each fine-tuned model for each problem type (Table 6 correspond to H30, R01, R02, and R04, respectively, and Table 7 shows the total number of correct answers for each model question type.). When expanding data using articles, accuracy improvements were observed in many problem types. The negation expansion showed significant contributions in problems involving negation problem types as expected. Data augmentation by person replacement was expected to contribute to the **Person** problem type (where person names are represented as alphabetical symbols such as A and B); H30 and R01 showed a positive contribution, while we could not observe positive contributions in other years. These result would suggest that the training dataset is still insufficient after augmenting the **Person** problem type by person replacement, as the alphabetical symbols could appear as a variety of different roles.

<pair id="R04-01-E" label="Y">
<article>
Article 5, A minor shall obtain the consent of his/her statutory representative before performing a legal act. However, this shall not apply to legal acts merely to obtain rights or to be relieved of obligations.
(2) Any legal act in violation of the provisions of the preceding paragraph may be rescinded.
Article 124, A supplementary acknowledgment of a revocable act shall not be effective unless it is made after the circumstances causing the revocation have ceased to exist and the rescuer becomes aware of his/her right to rescind.
(2) In any of the following cases, the ratification set forth in the preceding paragraph shall not be required to be made after the circumstances that were the cause of the rescission have ceased to exist.
(ii) Where a person with limited capacity to act (excluding an adult ward) (iii) When a person with limited capacity to act (excluding an adult ward) makes a supplementary acknowledgment with the consent of his/her statutory representative, conservator or assistant.
<problem>
A minor who has entered into a contract without the consent of a person with parental authority may not, until he or she reaches the age of majority, follow up on the contract on his or her own without the consent of the person with parental authority.

・The bold text is the important part to solve the problem, our article selection system by Sentence-luke selected below.
A minor shall obtain the consent of his/her statutory representative before performing a legal act. Any legal act in violation of the provisions of the preceding paragraph may be rescinded.

<関連条文>
第五条 未成年者が法律行為をするには、その法定代理人の同意を得なければならない。ただし、単に権利を得、又は義務を免れる法律行為については、この限りでない。
２ 前項の規定に反する法律行為は、取り消すことができる。
第百二十四条 取り消すことができる行為の追認は、取消しの原因となっていた状況が消滅し、かつ、取消権を有することを知った後にしなければ、その効力を生じない。
２ 次に掲げる場合には、前項の追認は、取消しの原因となっていた状況が消滅した後にすることを要しない。
二 制限行為能力者（成年被後見人を除く。）が法定代理人、保佐人又は補助人の同意を得て追認をするとき。
<問題文>
親権者の同意を得ずに契約を締結した未成年者は、成年に達するまでは、親権者の同意を得なければ、自らその契約の追認をすることができない。

・問題を解くためには、関連条文の太字部分が重要であるが、私達の条文選択システムは以下の条文を選択した。
未成年者が法律行為をするには、その法定代理人の同意を得なければならない。前項の規定に反する法律行為は、取り消すことができる。

Fig. 13 Examples of article selection failures

6 Conclusion and Future Works

We extended our previous system from COLIEE 2022 by performing an ensemble of the rule-based part and the LUKE-based part for COLIEE 2023 Task 4. We discriminated problems into two types based on whether they included alphabetical person names or not, and fine-tuned three different datasets on these two types of problems and all problems. We confirmed that our fine-tuned model for alphabetical person names improved the overall accuracy for those types of problems, achieving

Table 6 Problem type counts by year and model.
Model **Non** represents the baseline without any data augmentation applied. Model **Art** reflects extension based on articles, Model **Neg** corresponds to the negation augmentation, Model **Per** involves replacing person term, and Model **All** encompasses results obtained from applying all data augmentation

	H30					R01				
	Non	Art	Neg	Per	All	Non	Art	Neg	Per	All
Condition	26	29	34	28	32	38	47	42	45	43
Persons role	21	22	25	22	21	25	25	26	30	28
Person relationship	21	22	25	22	21	25	24	25	30	28
Morpheme	8	9	7	9	8	21	22	16	22	25
Anaphora	3	10	7	8	8	8	12	13	13	15
Case role	7	7	9	10	8	9	9	7	10	9
Verb paraphrases	10	9	12	9	11	25	23	21	23	26
Normal terms	2	5	6	3	5	9	7	6	11	9
Predicate argument	5	7	6	7	5	37	39	39	41	46
Negation	14	13	**21**	15	**21**	40	43	40	39	41
Legal fact	18	20	22	22	21	37	34	32	39	37
Entailment	1	3	2	1	1	5	7	2	6	6
Dependency	6	7	8	7	6	14	15	13	12	18
Article search	1	4	4	2	3	5	7	3	10	6
Paraphrases	4	2	1	5	1	2	2	1	2	1
Itemized	1	1	1	1	1	4	7	4	8	4
Calculation	1	1	1	2	1	0	1	1	0	1
Person	5	7	8	8	6	18	17	18	20	20
	R02					R04				
	Non	Art	Neg	Per	All	Non	Art	Neg	Per	All
Condition	18	19	19	17	18	37	44	42	40	49
Persons role	19	22	23	18	21	35	38	37	32	43
Person relationship	19	22	23	18	21	37	45	41	37	49
Morpheme	1	1	1	1	0	7	6	8	5	8
Anaphora	5	9	10	6	8	25	28	29	23	30
Case role	0	0	0	1	0	5	5	5	4	6
Verb paraphrases	5	8	8	6	9	8	6	8	6	9
Normal terms	11	11	12	11	11	2	1	1	1	1
Predicate argument	11	13	12	11	12	4	4	6	3	8
Negation	7	8	**9**	8	**9**	19	16	**22**	18	**25**
Legal fact	9	11	11	9	10	14	15	13	17	15
Entailment	1	1	1	1	1	7	7	7	6	7
Dependency	4	4	5	4	5	2	1	2	2	2
Article search	0	2	2	1	2	14	15	15	10	16
Paraphrases	0	1	1	0	1	8	7	7	7	8
Itemized	2	1	2	1	3	5	3	4	4	5
Calculation	3	4	4	3	3	5	4	4	5	4

Table 6 (continued)

	R02					R04				
	Non	Art	Neg	Per	All	Non	Art	Neg	Per	All
Person	20	22	23	19	21	22	25	22	17	27

Table 7 Problem type counts by model

	None	Articles	Negation	Person	All
Condition	119	139	137	130	142
Persons role	100	107	111	**102**	**113**
Person relationship	102	113	114	**107**	**119**
Morpheme	37	38	32	37	41
Anaphora	41	59	59	50	61
Case role	21	21	21	25	23
Verb paraphrases	48	46	49	44	55
Normal terms	24	24	25	26	26
Predicate argument	57	63	63	62	71
Negation	80	80	**92**	80	**96**
Legal fact	78	80	78	87	83
Entailment	14	18	12	14	15
Dependency	26	27	28	25	31
Article search	20	28	24	23	27
Paraphrases	14	12	10	14	11
Itemized	12	12	11	14	13
Calculation	9	10	10	10	9
Person	65	71	71	64	74

0.69 accuracy in the formal run for COLIEE 2023 Task 4. Our future work includes improving the data split method and processing other types of problems, as well as working on improving the accuracy of article selection.

Acknowledgements This research was partially supported by MEXT KAKENHI 00271635, JP22H00804, Japan, and SECOM Science and Technology Foundation.

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Competition on legal information extraction/entailment (coliee-14) workshop on ju-ris-informatics (jurisin) 2014 (2014). http://webdocs.cs.ualberta.ca/miyoung2/jurisin_task/index.html
2. Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423
3. Fujita, M., Onaga, T., Ueyama, A., & Kano, Y. (2022). Legal textual entailment using ensemble of rule based and bert based method with data augmentaion by rekated articke generation. In: Proceedings of the Sixteenth International Workshop on Jurisinformatics (JURISIN 2022), pp. 84–97.
4. Hoshino, R., Kiyota, N., & Kano, Y. (2019). Question answering system for legal bar examination using predicate argument structures focusing on exceptions. In: Proceedings of the Sixth International Competition on Legal Information Extraction/Entailment (COLIEE), pp. 38–42.
5. Kano, Y., Kim, M.Y., Goebel, R., & Satoh, K. (2017). Overview of coliee 2017. In K. Satoh, M.Y. Kim, Y. Kano, R. Goebel, T. Oliveira (Eds.) COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment, *EPiC Series in Computing*, vol. 47, pp. 1–8. EasyChair. https://doi.org/10.29007/fm8f. https://easychair.org/publications/paper/Fglr
6. Kano, Y., Kim, M. Y., Yoshioka, M., Lu, Y., Rabelo, J., Kiyota, N., Goebel, R., & Satoh, K. (2019). Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In K. Kojima, M. Sakamoto, K. Mineshima, & K. Satoh (Eds.), *New Frontiers in Artificial Intelligence* (pp. 177–192). Cham: Springer International Publishing.
7. Kim, M.Y., Goebel, R., Kano, Y., & Satoh, K. (2016). Coliee-2016: Evaluation of the competition on legal information extraction and entailment.
8. Kim, M.Y., Goebel, R., & Satoh, K. (2015). Coliee-2015: Evaluation of legal question answering.
9. Kim, M.Y., Rabelo, J., Goebel, R., Yoshioka, M., Kano, Y., & Satoh, K. (2023). Coliee 2022 summary: Methods for nbsp; legal document retrieval and nbsp; entailment. In: New Frontiers in Artificial Intelligence: JSAI-IsAI 2022 Workshop, JURISIN 2022, and JSAI 2022 International Session, Kyoto, Japan, June 12–17, 2022, Revised Selected Papers, pp. 51–67. Springer-Verlag, Berlin, Heidelberg. https://doi.org/10.1007/978-3-031-29168-5_4
10. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. In: arXiv preprint arXiv:1907.11692
11. Minh-Quan, B., Chau, N.M., Do, D.T., Le, N.K., Nguyen, D.H., & Nguyen, T.T.T. (2022). Jnlp team: Using deep learning approaches for tackling legal's challenges in coliee 2022. In: Proceedings of the Sixteenth International Workshop on Jurisinformatics (JURISIN 2022), pp. 70–83.
12. Rabelo, J., Goebel, R., Kim, M. Y., Kano, Y., Yoshioka, M., & Satoh, K. (2022). Overview and discussion of the competition on legal information extraction/entailment (Coliee) 2021. *Review of Socionetwork Strategies, 16*(1), 111–133.
13. Rabelo, J., Kim, M.Y., Goebel, R., Yoshioka, M., Kano, Y. & Satoh, K. (2020) Coliee 2020: Methods for legal document retrieval and entailment. In: New Frontiers in Artificial Intelligence: JSAI-IsAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15–17, 2020, Revised Selected Papers, p. 196–210. Springer-Verlag, Berlin, Heidelberg. https://doi.org/10.1007/978-3-030-79942-7_13
14. Rabelo, J., Kim, M. Y., Goebel, R., Yoshioka, M., Kano, Y., & Satoh, K. (2020). A summary of the Coliee 2019 competition. In M. Sakamoto, N. Okazaki, K. Mineshima, & K. Satoh (Eds.), *New Frontiers in Artificial Intelligence* (pp. 34–49). Cham: Springer International Publishing.
15. Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing

(EMNLP-IJCNLP), pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1410. https://aclanthology.org/D19-1410
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, pp. 6000–6010. Curran Associates Inc., Red Hook, NY, USA.
17. Wehnert, S., Kutty, L. & Luca, E.W.D. (2022). Using textbook knowledge for statute retrieval and entailment classification. In: Proceedings of the Sixteenth International Workshop on Jurisinformatics (JURISIN 2022), pp. 137–146.
18. Yamada, I., Asai, A., Shindo, H., Takeda, H. & Matsumoto, Y. (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), pp. 6442–6454. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.emnlp-main.523. https://aclanthology.org/2020.emnlp-main.523
19. Yoshioka, M., Suzuki, Y., & Aoki, Y. (2022). Hukb at the coliee 2022 statute law task. In: Proceedings of the Sixteenth International Workshop on Jurisinformatics (JURISIN 2022), pp. 33–46.