COLIEE 2022 Summary: Methods for Legal Document Retrieval and Entailment

Mi-Young Kim[1,3], Juliano Rabelo[1,2], Randy Goebel[1,2], Masaharu Yoshioka[4], Yoshinobu Kano[5], and Ken Satoh[6]

[1] Alberta Machine Intelligence Institute, Edmonton AB, Canada
[2] Department of Computing Science, University of Alberta, Edmonton AB, Canada
[3] Department of Science, Augustana Faculty, Camrose AB, Canada
{miyoung2,rabelo,rgoebel}@ualberta.ca
[4] Faculty of Information Science and Technology, Hokkaido University, Kita-ku, Sapporo-shi, Hokkaido, Japan
yoshioka@ist.hokudai.ac.jp
[5] Faculty of Informatics, Shizuoka University, Naka-ku, Hamamatsu-shi, Shizuoka, Japan
kano@inf.shizuoka.ac.jp
[6] National Institute of Informatics, Hitotsubashi, Chiyoda-ku, Tokyo, Japan
ksatoh@nii.ac.jp

Abstract. We present a summary of the 9th Competition on Legal Information Extraction and Entailment. The competition consists of four tasks on case law and statute law. The case law component includes an information retrieval task (Task 1), and the confirmation of an entailment relation between an existing case and an unseen case (Task 2). The statute law component includes an information retrieval task (Task 3) and an entailment/question answering task (Task 4). Participation was open to any group who could use any approach. Ten different teams participated in the case law competition tasks, most of them in more than one task. We received results from 9 teams for Task 1 (26 runs) and 5 teams for Task 2 (14 runs). On the statute law task, there were 11 different teams participating, most in more than one task. Five teams submitted a total of 15 runs for Task 3, and 6 teams submitted a total of 18 runs for Task 4. We summarize their approaches, our official evaluation, and provide analysis on our data and submission results.

Keywords: Legal Documents Processing · Textual Entailment · Information Retrieval · Classification · Question Answering.

1 Introduction

The Competition on Legal Information Extraction/Entailment (COLIEE) intends to develop the state of the art for information retrieval and entailment using legal texts. It is usually co-located with JURISIN, the Juris-Informatics workshop series, which was created to promote community discussion on both fundamental and practical issues on legal information processing. The intention is to embrace various disciplines, including law, social sciences, information processing, logic and philosophy, including the existing conventional "AI and law" area. In alternate years, COLIEE is organized as a workshop the International Conference on AI and Law (ICAIL), which was the case in 2017, 2019, and 2021. Until 2018, COLIEE consisted of two tasks: information retrieval (IR) and entailment using Japanese Statute Law (civil law). Since COLIEE 2018, IR and entailment tasks using Canadian case law were introduced.

Task 1 is a legal case retrieval task, and it involves reading a new case Q, and extracting supporting cases S_1, S_2, ..., S_n from the provided case law corpus, which are hypothesized to support the decision for Q. Task 2 is a legal case entailment task, which involves the identification of a paragraph or paragraphs from existing cases, which entail a given fragment of a new case. For the information retrieval task (Task 3), based on the discussion about the analysis of previous COLIEE IR tasks, we modify the evaluation measure of the final results and ask participants to submit ranked relevant articles results to further understand details of the difficulty of the questions. For the entailment task (Task 4), we performed categorized analyses to expose a variety of issues with the problems and characteristics of the submissions, in addition to the evaluation accuracy as in previous COLIEE tasks.

The rest of our paper is organized as follows: Sections 2, 3, 4, 5 describe each task, presenting their definitions, datasets, list of approaches submitted by the participants, and results attained. Section 6 presents final some final remarks.

2 Task 1 - Case Law Information Retrieval

2.1 Task Definition

This task consists in finding which cases, in the set of provided candidate cases, should be "noticed" with respect to a given query case. "Notice" is a legal technical term that denotes a legal case description that is considered to be relevant to a query case. More formally, given a query case q and a set of candidate cases $C = \{c_1, c_2, ..., c_n\}$, the task is to find the supporting cases $S = \{s_1, s_2, ..., s_n \mid s_i \in C \wedge noticed(s_i, q)\}$ where $noticed(s_i, q)$ denotes a relationship which is true when $s_i \in S$ is a noticed case with respect to q.

2.2 Dataset

The dataset is comprised of a total of 5,978 case law files. Provided is a labelled training set of 4,415, of which 900 query cases. On average, there are approximately 4.9 noticed cases per query case in the provided training dataset, which

should be identified among the 4,415 cases. To prevent competitors to merely use citations of past cases in order to identify cited cases, citations are suppressed from the case contents and replaced by a "FRAGMENT_SUPPRESSED" tag indicating that fragment was removed from the case contents.

A test set is given, consisting of a total of 1,563 cases, with 300 query cases and a total of 1,263 true noticed cases, which means there are on average 4.21 noticed cases per query case in the test dataset. Initially, the golden labels for that test set is not provided to competitors.

2.3 Approaches

We received 26 submissions from 9 different teams for Task 1. In this section, we present an overview of the approaches taken by the 7 teams which submitted papers describing their methods. Please refer to the corresponding papers for further details.

- **TUWBR (2 runs)** [5] start from two assumptions: first, that there is a topical overlap between query and notice cases, but that not all parts of a query case are equally important. Secondly, they assume that traditional IR methods such as BM25 provide competitive results in Task 1. They perform document level and passage level retrieval, and also augment the system by adding external domain knowledge by extracting statute fragments mentioned in the cases and explicitly adding those fragments to the documents.
- **JNLP (3 runs)** [3] applies an approach that consists first splits the documents into paragraphs, then calculates the similarities between cases by combining term-level matching and semantic relationships on the paragraph level. They also propose an attention model to encode the whole query in the context of candidate paragraphs, then infer the relationship between cases.
- **DoSSIER (3 runs)** [1] combined traditional and neural based techniques in Task 1. The authors investigate lexical and dense first stage retrieval methods aiming for a high recall in the initial retrieval and then compare shallow neural re-ranking with the MTFT-BERT model to the BERT-PLI model. They then investigate which part of the text of a legal case should be taken into account for re-ranking. The achieved results show that BM25 shows a consistently high effectiveness across different test collections in comparison to the neural network re-ranking models.
- **LeiBi (3 runs)** [2] applied an approach which consists of the following steps: first, given a legal case as a query, they reformulate it by heuristically extracting various meaningful sentences or n-grams. The authors then use the pre-processed query case to retrieve an initial set of possible relevant legal cases, which are further re-ranked. Finally, the team aggregates the relevance scores obtained by the first stage and the re-ranking models to improve retrieval effectiveness. The query cases are reformulated to be shorter, using three different statistical methods (KLI, PLM, IDF-r), in addition to models that leverage embeddings (e.g., KeyBERT). Moreover, the authors investigate if automatic summarization using Longformer-Encoder-Decoder

(LED) can produce an effective query representation for this retrieval task. Furthermore, the team proposes a re-ranking cluster-driven approach, which leverages Sentence-BERT models to generate embeddings for sentences from the query and candidate documents. Finally, the authors employ a linear aggregation method to combine the relevance scores obtained by traditional IR models and neural-based models.

- **UA (3 runs)** [9] use a transformer-based model to generate paragraph embeddings, and then calculate the similarity between paragraphs of query and positive and negative cases. The calculated similarities are used to generate feature vectors (10-bin histograms of all pair-wise between 2 cases), and then using a Gradient Boosting classifier to determine if those cases should be noticed or not. The UA team also applies pre- and post-processing heuristics to generate the final results, which were ranked first in Task 1 of the current COLIEE edition;
- **nigam (3 runs)** [8] developed an approach which was a combination of transformer-based and traditional IR techniques; more specifically, they used Sentence-BERT and Sent2Vec for semantic understanding combined with BM25. First, the nigam team selects top-K candidates according to the BM25 rankings, and then they use pre-trained Sentence-BERT and Sent2Vec to generate representation features of each sentence. The authors also used cosine similarity with the max-pooling strategy to get the final document score between the query and noticed cases.
- **siat (3 runs)** [15] show how longformer-based contrastive learning is able to process sequences of thousands of tokens, thus overcoming a well-known limitation of common transformer-based methods which are usually restricted to between 512 and 1024 tokens. In addition to that longformer-based approach, the siat team also explores traditional retrieval models. They achieved second place overall in Task 1 of COLIEE 2022.

2.4 Results

Table 1 shows the results of all submissions received for Task 1 for COLIEE 2022. A total of 26 submissions from 9 different teams have been received. Similar to what happened in COLIEE 2021 [11], the f1-scores are generally low, which reflects the fact that the task is now more challenging than its previous formulation (for a description of the previous Task 1 formulation, please see the COLIEE 2020 summary [10]). However, in the current edition, we already could see a relevant improvement in those scores, with the top teams achieving scores above 0.35 (up from 0.19 as the best score in the COLIEE 2021 edition).

Most of the participating teams applied traditional IR techniques such as BM25, transformer based methods such as BERT, or a combination of both. The best performing team was UA, with an f1-score of 0.3715, with an approach that relied on creating an embedding representation for the cases, and then calculating the similarity between each query case and positives and negatives samples from the training dataset. The resulting distances are then bucketed into 10-bin histograms, which are used to train a Gradient Boosting classifier. Also

worth mentioning is the siat team, whose approach made use of a longformer-based model and achieved second place overall.

Table 1. Task 1 results

Team	File	F1 Score	Precision	Recall
UA	pp_0.65_10_3.csv	0.3715	0.4111	0.3389
UA	pp_0.7_9_2.csv	0.3710	0.4967	0.2961
siat	siatrun1.txt	0.3691	0.3005	0.4782
siat	siatrun3.txt	0.3680	0.3026	0.4695
UA	pp_0.65_6.csv	0.3559	0.3630	0.3492
siat	siatrun2.txt	0.2964	0.2522	0.3595
LeiBi	run_bm25.txt	0.2923	0.3000	0.2850
LeiBi	run_weighting.txt	0.2917	0.2687	0.3191
JNLP	run3.txt	0.2813	0.3211	0.2502
nigam	bm25P3M3.txt	0.2809	0.2587	0.3072
JNLP	run2.txt	0.2781	0.3144	0.2494
DSSR	DSSR_01.txt	0.2657	0.2447	0.2906
JNLP	run1.txt	0.2639	0.2446	0.2866
DSSR	DSSR_03.txt	0.2461	0.2267	0.2692
TUWBR	TUWBR_LM_law	0.2367	0.1895	0.3151
LeiBi	run_clustering.txt	0.2306	0.2367	0.2249
TUWBR	TUWBR_LM	0.2206	0.1683	0.3199
nigam	sbertP3M3RS.txt	0.1542	0.1420	0.1686
nigam	s2vecP3M3RS.txt	0.1484	0.1367	0.1623
DSSR	DSSR_02.txt	0.1317	0.1213	0.1441
LLNTU	2022.task1.LLNTUtfidCos	0.0000	0.0000	0.0000
LLNTU	2022.task1.LLNTUtanadaT	0.0000	0.0000	0.0000
LLNTU	2022.task1.LLNTU3q4clii	0.0000	0.0000	0.0000
Uottawa	Task1Run3_UottawaLegalBert.txt	0.0000	0.0000	0.0000
Uottawa	Task1Run1_UottawaMB25.txt	0.0000	0.0000	0.0000
Uottawa	Task1Run2_UottawaSentTrans.txt	0.0000	0.0000	0.0000

For future editions of COLIEE, we intend to make the distributions of the training and test datasets more similar with respect to average and standard deviation of number of noticed cases. There are still some issues we need to fix in the dataset, such as two different files with the exact same contents (i.e., the same case represented as two separate files). This is a problem with the original dataset from where the competition's data is drawn, and knowing that dataset presents those issues we will improve our collection methods to correct them. Fortunately, those issues were rare and did not have a relevant impact on the final results.

3 Task 2 - Case Law Entailment

3.1 Task Definition

Given a base case and a specific fragment from it, together with a second case relevant to the base case, this task consists in determining which paragraphs of the second case entail that fragment of the base case. More formally, given a base case b and its entailed fragment f, and another case r represented by its paragraphs $P = \{p_1, p_2, ..., p_n\}$ such that $noticed(b, r)$ as defined in section 2 is true, the task consists in finding the set $E = \{p_1, p2, ..., p_m \mid p_i \in P\}$ where $entails(p_i, f)$ denotes a relationship which is true when $p_i \in P$ entails the fragment f.

3.2 Dataset

In Task 2, 525 query cases were provided for training against 18,740 paragraphs. There were 100 query cases against 3278 candidate paragraphs as part of the testing dataset. On average, there are 35.627 candidate paragraphs for each query case in the training dataset and 32.455 candidate paragraphs for each query case in the testing dataset. The average number of relevant paragraphs for Task 2 was 1.14 paragraphs for training and 1.18 paragraphs for testing.

3.3 Approaches

Five teams submitted a total of 14 runs to this task. They used LegalBERT, BM25, zero shot models, and some other heuristic approaches. More details on the approaches are shown below. Here, we introduce three teams' approaches that described their methods in their papers.

- **JNLP (3 runs)** [3] applied LegalBERT and BM25. In their first run, they combined the two scores from LegalBERT and BM25, and then ranked the outputs. In the second run, they used a knowledge representation technique called "Abstract Meaning Representation" (AMR) to capture the most important words in the query and corresponding candidate paragraph. In the third run, instead of combining the two scores from LegalBERT and BM25, they identified relevant paragraphs through the interaction between the top N candidates in LegalBERT and top M candidates in AMR + BM25.

- **nigam (2 runs)** [8] submitted two official runs both of which are based on the BM25 model. Run-1 uses only entailed fragment text as a query, and run-2 uses the filtered base case such that they combine search entailed fragment text into the base case, then provided as input to the model as searched results (along with previous and following sentences to capture the other relevant information). Every query case predicts one case law paragraph because the average number of relevant paragraphs is approximately 1.14 in the training dataset.

– **NM (3 runs)** [13] used monoT5, which is an adaptation of the T5 model. During inference, monoT5 generates a score that measures the relevance of a document to a query by applying a softmax function to the logits of the tokens "true" and "false." NM also extends a zero-shot approach, and they fine-tune the T5-base and T5-3B models for 10k steps, which corresponds to almost one epoch or approximately 530,000 query-passage pairs from the MS MARCO training set. They refer to the resulting models as monoT5-base-zero-shot and monoT5-3b-zero-shot. In the third run, they combine the answers from the two models. They apply their own answer selection method to select the final set of answers from the two models. Their ensemble model was ranked first in the COLIEE 2022 Task 2 competition.

3.4 Results

The F1-measure is used to assess performance in this task. The actual results of the submitted runs by all participants are shown on table 2, from which it can be seen that the NM team attained the best results. Among the three submissions from NM, two submissions were ranked first and second.

Table 2. Results attained by all teams on the test dataset of task 2.

Team	Submission File	F1-score	Precision	Recall
NM	**monot5-ensemble.txt**	**0.6783**	0.6964	0.6610
NM	monot5-3b.txt	0.6757	0.7212	0.6356
JNLP	run2_bert_amr_remove_reduntdant_filter.txt	0.6694	0.6532	0.6864
JNLP	run3_bert_BM25.txt	0.6612	0.6452	0.6780
jljy	run2_task2.txt	0.6514	0.7100	0.6017
jljy	run3_task2.txt	0.6514	0.7100	0.6017
JNLP	run1_bert_amr_remove_reduntdant.txt	0.6452	0.6154	0.6780
jljy	run1_task2.txt	0.6330	0.6900	0.5847
NM	monot5-base.txt	0.6325	0.6379	0.6271
UA	res_score-0.95_max-1.txt	0.5446	0.6105	0.4915
UA	res_score-0.5_max-1.txt	0.5363	0.7869	0.4068
UA	res_score-0.95_max-5.txt	0.4121	0.3049	0.6356
nigam	bm25EF.txt	0.3204	0.1980	0.8390
UA	bm25BC.txt	0.2104	0.1300	0.5508

4 Task 3 - Statute Law Retrieval

4.1 Task Definition

Task 3 is a pre-processing step for legal textual entailment (Task 4), whose goal is to extract a subset of Japanese Civil Code Articles S_1, S_2,...,S_n from the entire Civil Code articles considered appropriate for answering the legal bar exam question Q such that
$Entails(S_1, S_2, ..., S_n, Q)$ or $Entails(S_1, S_2, ..., S_n, notQ)$.

Given a question Q and the all Civil Code Articles, the participants are required to retrieve the set of "$S_1, S_2, ..., S_n$" as the answer of this task.

4.2 Dataset

For Task 3, questions related to Japanese civil code articles were selected from the Japanese bar exam. However, since (updated in 2020), we use civil law articles that have official English translation (768 articles in total) as the target civil code.

The number of questions classified by the number of relevant articles is listed in Table 3.

Table 3. Number of questions classified by number of relevant articles

number of relevant article(s)	1	2	3	4	5	total
number of questions	94	11	2	1	1	109

4.3 Approaches

The following 5 teams submitted their results (15 runs in total). All teams had experience in submitting results in the previous competition and extend their previous approaches. Ordinal IR models such as BM25 [12], TF-IDF are still good models with better performance. Deep learning (DL) based approaches (using BERT [4] and other variants) are also effective for the task. There are several runs that combines outputs of these different approaches.

- **HUKB (3 runs)** [16] uses BM25 IR model with different document article databases (original article, rewritten articles using reference, and the judicial decision part of the articles). Final results are generated by using these results.
- **JNLP (3 runs)** [3] uses a deep learning (DL) based approach with identification of the use-case questions. Due to the different characteristics of the data for use-case question and others, they propose to make two models: one is for ordinal questions and the other is for use-cases.
- **LLNTU (3 runs)** previously used a BERT-based method, but there paper has no clear explanation about any adjustments for this year's task.
- **OvGU (3 runs)** [14] uses the scores of a TF-IDF model combined with a sentence-embedding based similarity score to produce answer rankings. They also use external knowledge (texts related to the articles) to calculate sentence-embedding.
- **UA (3 runs)** uses the TF-IDF model and BM25 model as an IR module.

4.4 Results

Table 4 shows the evaluation results of submitted runs. The official evaluation measures used in this task were macro average (average of evaluation measure values for each query over all queries) of F2 measure[7], precision, and recall.

$$precision = \frac{number\ of\ retrieved\ relevant\ articles}{number\ of\ returned\ articles} \tag{1}$$

$$recall = \frac{number\ of\ retrieved\ relevant\ articles}{number\ of\ relevant\ articles} \tag{2}$$

$$f2 = \frac{5 \times precision \times recall}{4 \times precision + recall} \tag{3}$$

We also calculate the mean average precision (MAP), recall at k (R_k: recall calculated by using the top k ranked documents as returned documents) by using the long ranking list (100 articles).

Table 4 shows the results of the evaluation of submitted results. Due to the limitation of the size of paper, the best performance run in terms of F2 are selected from each team runs.

Table 4. Evaluation results of submitted runs (Task 3) and the corresponding organizers' run

sid	return	retrieved	F2	Precision	Recall	MAP
HUKB2	136	101	0.820	0.818	0.841	0.843
OVGU_run3	161	96	0.779	0.778	0.805	0.836
JNLP.longformer	178	101	0.770	0.687	0.838	0.793
UA_TFIDF2	115	90	0.764	0.807	0.764	0.829
LLNTU0066cc	114	74	0.642	0.674	0.639	0.700

Figure 1,2,3 shows average of evaluation measure for all submission runs. The number of questions whose relevant article is 1 and F2 (average of all submission) is higher than 0.8 is 65.9% (62/94). This is better than that for the COLIEE 2021 29.2% (19/65). This may reflect the different characteristics of the dataset and participation of the well-experienced team, but this result shows that we have almost succeeded to develop a method for identifying easy questions. From Fig. 1, we confirmed that there are many easy questions for which almost all system can retrieve the relevant article. The easiest question is R03-2-A "The obligee may not exercise the right of the obligor, if the right is immune from attachment." whose relevant article contains sentence with same vocabularies. However, there are also many queries for which none of the system can retrieve the relevant articles. R03-07-E is an example of this question: "After A sold land

[7] Since task 3 is a preprocess for legal textual entailment (task 4), it is important to have all relevant articles in the retrieved results. So we put emphasis on recall in this evaluation

X owned by A to B, B resold land X to C, and each of them was registered as such. Later, the sales contract between A and B was voided on the grounds that A was an adult ward. If C did not know without negligence that A was an adult ward, A may not claim that the ownership of land X belongs to A," and relevant article is Article 121 "An act that has been rescinded is deemed void ab initio." This article uses specific legal terminology such as "rescinded" and "void ab initio." This is almost impossible to handle with an ordinal keyword-based IR system. It is also difficult for the DL-based approach, because it is not a simple semantic association matching.

For the question with multiple relevant articles, it is difficult to determine the significant difference from the viewpoint of overall evaluation. We still have problems to determine the whole set of relevant articles compared with single relevant article cases.

A new approach proposed in this year's competition is as follows:

- Using different document database (e.g., article text database and texts for judicial decision part database) for calculating the score to merge [16].
- Using external knowledge (text related to the articles) to enrich the sentence-embedding information of the article. [14]
- Classification of the query types; use-case queries or others. [3]

Those approaches are confirmed to be effective in this year's test data, and we expect that the combination of those approaches may improve the retrieval performance compared with this year's system.

5 Task 4 - Statute Law Entailment

5.1 Task Definition

Task 4 is a task to determine entailment relationships between a given problem sentence and article sentences. Competitor systems should answer "yes" or "no" regarding the given problem sentences and given article sentences. Until COLIEE 2016, the competition had pure entailment tasks, where t1 (relevant article sentences) and t2 (problem sentence) were given. Due to the limited number of available problems, COLIEE 2017, 2018 did not retain this style of task. In the Task 4 of COLIEE 2019, 2020 and 2022, we returned to the pure textual entailment task to attract more participants, allowing more focused analyses. Participants can use any external data, however assuming that they do not use the test dataset and/or something which could directly contains the correct answers of the test dataset, because this task is intended to be a pure textual entailment task. Towards deeper analysis, we asked the participants to submit their outputs when using any fragment of the training dataset (H30-R02), in addition to the formal runs.

5.2 Dataset

Our training dataset and test dataset are the same as for Task 3. Questions related to Japanese civil law were selected from the Japanese bar exam. The

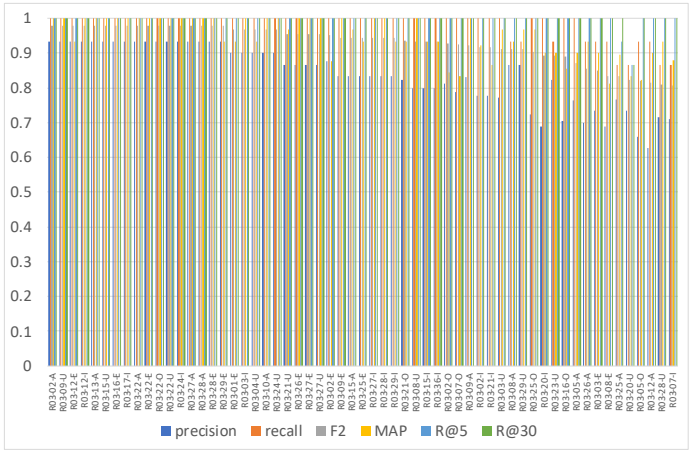

Fig. 1. Averages of precision, recall, F2, MAP, R_5, and R_30 for easy questions with a single relevant article

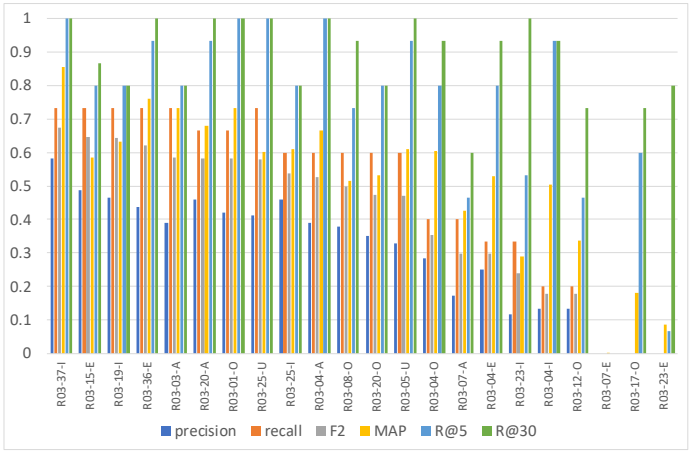

Fig. 2. Averages of precision, recall, F2, MAP, R_5, and R_30 for noneasy questions with a single relevant article

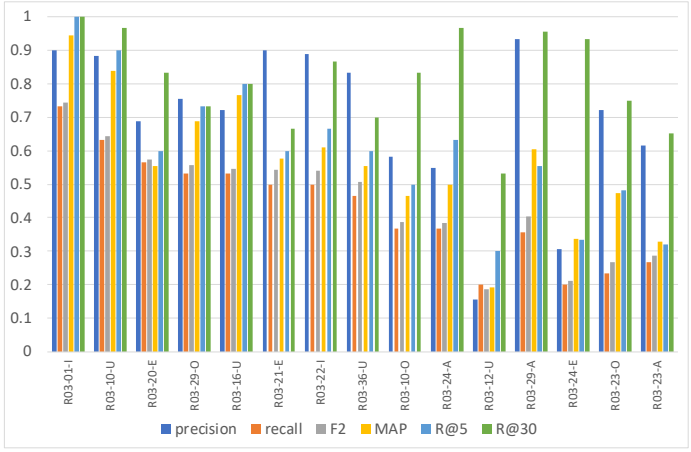

Fig. 3. Averages of precision, recall, F2, MAP, R_5, R_10, and R_30 for noneasy questions with a single relevant article

organizers provided a data set used for previous campaigns as training data (887 questions) and new questions selected from the 2022 bar exam as test data (109 questions).

5.3 Approaches

We describe approaches for each team as follows, shown as a header format of **Team Name (number of submitted runs)**.

- **HUKB (3 runs)** [16] proposed a method to select relevant part from the articles (**HUKB-2**) and a new data augmentation method (**HUKB-1**) in addition to their system in COLIEE 2021 (**HUKB-3**) which uses an ensemble of BERT with data augmentation, extracting judicial decision sentences, creating positive/negative data from articles.
- **JNLP (3 runs)** [3] compared ELECTRA, RoBERTa, and LegalBERT, which is pretrained using large legal English texts. They also compared impacts of negation data augmentation, and paragraph-level entailments.
- **KIS (3 runs)** [6] employed an ensemble of their rule-based method using predicate-argument structures which extends their previous work, and BERT-based methods. Their BERT-based methods commonly use data augmentation (**KIS2**), with data selection (**KIS1**), and with person name inference (**KIS3**). They also employed an ensemble of different trials of fine-tunings.
- **LLNTU (2 runs)** [7] restructured given data to a dataset of the disjunctive union strings from training queries and articles, and established a longest uncommon subsequence similarity comparison model, without stopwords (**LLNTUdiffSim**), and with stopwords (**LLNTUdeNgram**). One of their runs was retracted because they used their dataset via web crawling that could potentially include the correct answers of the test dataset.
- **OvGU (3 runs)** [14] employed an ensemble of graph neural networks (GNNs) as their previous work (**OvGU1**), concatenated with referring textbook nodes (**OvGU2** and averaging sentence embeddings (**OvGU3**). No submission for the past training datasets.
- **UA (3 runs)** [9] provides no description for Task 4.

5.4 Results

Table 5 shows evaluation results of Task 4, including the formal run results and training dataset. The evaluation results of the training dataset regard either of R02, R01, or H30 as test dataset, using corresponding former years' dataset (-R01, -H30, -H29) as training datasets; these configurations correspond to the formal runs of COLIEE 2021, 2020, and 2019, respectively.

6 Conclusion

We have summarized the systems and their performance as submitted to the COLIEE 2022 competition. For Task 1, UA submitted by the University of

Table 5. Evaluation results of submitted runs (Task 4). L: Dataset Language (J: Japanese, E: English), #: number of correct answers

Team	Submission ID	L	Formal Run		R02		R01		H30	
			#	Accuracy	#	Accuracy	#	Accuracy	#	Accuracy
N/A	Total	-	109	1.0000	81	1.0000	111	1.0000	70	1.0000
N/A	BaseLine	-	58	0.5320	43	0.5309	59	0.5315	36	0.5143
KIS	KIS2	J	74	0.6789	44	0.5432	71	0.6396	49	0.7000
HUKB	HUKB-1	J	73	0.6697	50	0.6173	73	0.6577	42	0.6000
KIS	KIS1	J	72	0.6606	48	0.5926	68	0.6126	46	0.6571
KIS	KIS3	J	72	0.6606	49	0.6049	68	0.6126	47	0.6714
HUKB	HUKB-2	J	69	0.6330	48	0.5926	74	0.6667	41	0.5857
HUKB	HUKB-3	J	69	0.6330	58	0.7160	72	0.6486	44	0.6286
LLNTU	LLNTUdeNgram	J	66	0.6055	47	0.5802	60	0.5405	33	0.4714
LLNTU	LLNTUdiffSim	J	63	0.5780	49	0.6049	56	0.5045	36	0.5143
OVGU	OVGU3	?	63	0.5780	-	-	-	-	-	-
UA	UA_e	?	59	0.5413	44	0.5432	69	0.6216	32	0.4571
UA	UA_r	?	59	0.5413	44	0.5432	69	0.6216	32	0.4571
UA	UA_structure	?	59	0.5413	44	0.5432	69	0.6216	32	0.4571
JNLP	JNLP1	J	58	0.5321	50	0.6173	59	0.5315	40	0.5714
JNLP	JNLP2	J	58	0.5321	49	0.6049	58	0.5225	40	0.5714
OVGU	OVGU2	?	58	0.5321	-	-	-	-	-	-
JNLP	JNLP3	J	56	0.5138	48	0.5926	65	0.5856	42	0.6000
OVGU	OVGU1	?	52	0.4771	-	-	-	-	-	-

Alberta team was the best performing team with an F1 score of 0.3715. In Task 2, the winning team combined T5-base and T5-3B models, and achieved an F1 score of 0.6783. For Task 3, the top ranked team is HUKB and achieved an F2 score of 0.8204. KIS was the Task 4 winner, with an Accuracy of 0.6789.

We intend to further improve the datasets quality in future editions of COLIEE so the tasks more accurately represent real-world problems.

Acknowledgements

This competition would not be possible without the significant support of Colin Lachance from vLex and Compass Law, and the guidance of Jimoh Ovbiagele of Ross Intelligence and Young-Yik Rhim of Intellicon. Our work to create and run the COLIEE competition is also supported by our institutions: the National Institute of Informatics (NII), Shizuoka University and Hokkaido University in Japan, and the University of Alberta and the Alberta Machine Intelligence Institute in Canada. We also acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [DGECR-2022-00369, RGPIN-2022-0346]. This work was also supported by JSPS KAKENHI Grant Numbers, JP17H06103 and JP19H05470 and JST, AIP Trilateral AI Research, Grant Number JPMJCR20G4.

References

1. Abolghasemi, A., Althammer, S., Hanbury, A., Verberne, S.: Dossier@coliee2022: Dense retrieval and neural re-ranking for legal case retrieval. In: Sixteenth International Workshop on Juris-informatics (JURISIN) (2022)
2. Askari, A., Peikos, G., Pasi, G., Verberne, S.: Leibi@coliee 2022: Aggregating tuned lexical models with a cluster-driven bert-based model for case law retrieval. In: Sixteenth International Workshop on Juris-informatics (JURISIN) (2022)
3. Bui, M.Q., Nguyen, C., Do, D.T., Le, N.K., Nguyen, D.H., Nguyen, T.T.T.: Using deep learning approaches for tackling legal's challenges (coliee 2022). In: Sixteenth International Workshop on Juris-informatics (JURISIN) (2022)
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018)
5. Fink, T., Recski, G., Kusa, W., Hanbury, A.: Statute-enhanced lexical retrieval of court cases for coliee 2022. In: Sixteenth International Workshop on Juris-informatics (JURISIN) (2022)
6. Fujita, M., Onaga, T., Ueyama, A., Kano, Y.: Legal textual entailment using ensemble of rule-based and bert-based method with data augmentations including generation without excess or deficiency. In: Sixteenth International Workshop on Juris-informatics (JURISIN) (2022)
7. Lin, M., Huang, S.C., Shao, H.L.: Rethinking attention: An attempting on revaluing attention weight with disjunctive union of longest uncommon subsequence for legal queries answering. In: Sixteenth International Workshop on Juris-informatics (JURISIN) (2022)
8. Nigam, S.K., Goel, N.: nigam@coliee-22: Legal case retrieval and entailment using cascading of lexical and semantic-based models. In: Sixteenth International Workshop on Juris-informatics (JURISIN) (2022)
9. Rabelo, J., Kim, M.Y., Goebel, R.: Semantic-based classification of relevant case law. In: Sixteenth International Workshop on Juris-informatics (JURISIN) (2022)
10. Rabelo, J., Kim, M.Y., Goebel, R., Yoshioka, M., Kano, Y., Satoh, K.: COLIEE 2020: Methods for Legal Document Retrieval and Entailment, pp. 196–210 (06 2021). https://doi.org/10.1007/978-3-030-79942-7_13
11. Rabelo, J., Kim, M.Y., Goebel, R., Yoshioka, M., Kano, Y., Satoh, K.: Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. In: The Review of Socionetwork Strategies. vol. 16, pp. 111–133 (04 2022). https://doi.org/10.1007/s12626-022-00105-z
12. Robertson, S.E., Walker, S.: Okapi/Keenbow at TREC-8. In: Proceedings of TREC-8. pp. 151–162 (2000)
13. Rosa, G.M., Bonifacio, L.H., Jeronymo, V., de Alencar Lotufo, R., Nogueira, R.: 3b parameters are worth more than in-domain training data: A case study in the legal case entailment task. In: Sixteenth International Workshop on Juris-informatics (JURISIN) (2022)
14. Wehnert, S., Kutty, L., Luca, E.W.D.: Using textbook knowledge for statute retrieval and entailment classification. In: Sixteenth International Workshop on Juris-informatics (JURISIN) (2022)
15. Wen, J., Zhong, Z., Bai, Y., Zhao, X., , Yang, M.: Siat@coliee-2022: Legal case retrieval with longformer-based contrastive learning. In: Sixteenth International Workshop on Juris-informatics (JURISIN) (2022)
16. Yoshioka, M., Suzuki, Y., Aoki, Y.: Hukb at the coliee 2022 statute law task. In: Sixteenth International Workshop on Juris-informatics (JURISIN) (2022)