Summary of the Competition on Legal Information,Extraction/Entailment (COLIEE) 2023

Randy Goebel
University of Alberta
Edmonton, AB, Canada
rgoebel@ualbert.ca

Yoshinobu Kano
Shizuoka Univerity
Hamamatsu, Shizuoka, Japan
kano@kanolab.net

Mi-Young Kim
University of Alberta
Edmonton, AB, Canada
miyoung2@ualberta.ca

Juliano Rabelo
University of Alberta
Edmonton, AB, Canada
rabelo@ualberta.ca

Ken Satoh
National Institute of Informatics
Tokyo, Japan
ksatoh@nii.ac.jp

Masaharu Yoshioka
Hokkaido University
Sapporo, Hokkaido, Japan
yoshioka@ist.hokudai.ac.jp

ABSTRACT

This paper summarizes the 10th Competition on Legal Information Extraction and Entailment. In this edition, the competition included 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 based on retrieved civil code statutes (Task 4). Participation was open to any group based on any approach. Ten different teams participated in the case law competition tasks, most of them in more than one task. We received results from 8 teams for Task 1 (22 runs) and seven teams for Task 2 (18 runs). On the statute law task, there were 9 different teams participating, most in more than one task. 6 teams submitted a total of 16 runs for Task 3, and 9 teams submitted a total of 26 runs for task 4. We describe in this paper the approaches, our official evaluation, and analysis on our data and submission results.

CCS CONCEPTS

• **Information systems → Content analysis and feature selection**; **Similarity measures**; **Clustering and classification**; *Document topic models*; *Information extraction*; Specialized information retrieval.

KEYWORDS

legal textual entailment, legal information retrieval, text classification, imbalanced datasets

ACM Reference Format:
Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2023. Summary of the Competition on Legal Information,Extraction/Entailment (COLIEE) 2023. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3594536.3595176

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06…$15.00
https://doi.org/10.1145/3594536.3595176

1 INTRODUCTION

The objective of the Competition on Legal Information Extraction/Entailment (COLIEE) is 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, with the intention 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 of the International Conference on AI and Law (ICAIL), which was the case in 2017 and 2019, 2021, and again in 2023. Until 2017, 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 query case and extracting supporting cases from the provided case law corpus, hypothesized to be relevant to the query case. Task 2 is the 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. Task 3 and 4 are tasks for statute law tasks that use Japanese Bar exam to judge whether the given statement is true or not. Task 3 is an information retrieval task that retrieve relevant article for the legal entailment (task 4) and Task 4 is a legal entailment task that judge whether the given statement is true or not.

The rest of the 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 5.4 presents final some final remarks.

2 TASK 1 - CASE LAW RETRIEVAL

2.1 Task Definition

This task consists in finding what cases, amongst a set of provided candidate cases, should be "noticed" with respect to a given query

case[1]. 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 5,735 case law files. Also given is a labelled training set of 4,400, of which 959 are query cases. On average, the training data includes approximately 4.67 noticed cases per query case, which are to be identified among the 4,400 cases. To prevent competitors from merely using existing embedded conventional citations in historical cases to identify cited cases, citations are suppressed from all candidate cases and replaced by a "FRAGMENT_SUPPRESSED" tag indicating that a fragment was removed from the case contents. A test set consists of a total of 1,335 cases, with 319 query cases and a total of 859 true noticed cases (an average of 2.69 noticed cases per query case). Initially, the golden labels for that test set are not provided to competitors.

2.3 Approaches

We received 22 submissions from 8 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.

- **THUIR (3 runs)** [5] design structure-aware pre-trained language models to enhance legal case understanding. The authors also propose heuristic pre- and post- processing approaches to reduce the influence of irrelevant items. Last, learning-to-rank methods are applied to merge features with different dimensions.
- **UFAM (3 runs)** [9] explores the idea of filtering + ranking results, which was implemented by topic discovery using BERTopic followed by a ranking algorithm. The topic discovery step assigns k topics to a case (k being a parameter which is varied in the experiments). The ranking step takes whatever candidate contains the dominant query topic in its k most relevant topic list. Ranking was implemented in 3 different ways, being the best one the cosine similarity between the query and a candidate case.
- **JNLP (3 runs)** [2] implements data augmentation techniques to produce additional training data and employs large language models to capture the nuances of legal language. The data augmentation step generates synthetic cases that exhibit similar attributes to the original cases. Then, a large language model is trained on the augmented dataset and used to retrieve relevant cases (the same overall approach is also used to determine entailment in Task 2).
- **UA (3 runs)** [11] use a transformer-based model to generate paragraph embeddings, and then calculate the similarity between paragraphs of a query case and positive and negative cases. These calculated similarities are used to generate feature vectors (10-bin histograms of all pair-wise comparisons between 2 cases). They then use 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.
- **NOWJ** [14] propose a two-phase matching approach: mono matching (paragraph/decision level) and panorama matching (case level). The authors pre-processed the data by removing French content, segmenting cases into paragraphs, extracting case years, removing redundant characters, and detecting important passages. In the mono matching phase, they combined lexical and semantic matching models. Lexical matching was performed using BM25[13] to calculate relevance scores between paragraphs, while semantic matching was carried out using a fine-tuned supporting model. A lexical model was initially used to narrow down the search space and select potential candidates. In the panorama matching phase, a Longformer model was used to compare base and candidate cases based on their overall similarities.
- **IITDLI** [4] developed an approach to task 1 that can be summarized in 6 steps: 1) Pre-processing: Remove French words, extract years, and perform feature extraction using unigram/word features; 2) Term extraction: Use Kullback-Leibler Divergence for Informativeness and Term Frequency and Inverse Document Frequency for query reformulation. 3) Retrieval: Use BM25 as a ranking model to retrieve top-n results from the corpus. 4) Filtering: Apply a year filtering method to refine the results. 5) Experiments with additional filters, which ended up not being used in the final submission because showed worse results in the experiments performed. 6) Post-processing: Implement a thresholding scheme for selecting the final set of candidate relevant cases, which improves precision and overall F1 score.

The other participating teams did not send papers describing the details of their approaches.

2.4 Results and Discussion

Table 1 shows the results of all submissions received for Task 1 for COLIEE 2023. A total of 22 submissions from 8 different teams were evaluated. Similar to what happened in recent COLIEE editions, the f1-scores are generally low, which reflects the fact that the task is now more challenging than its previous formulation [2].

Most of the participating teams applied some form of traditional IR techniques such as BM25, transformer based methods such as BERT, or a combination of both. The best performing team (THUIR) employed pre-trained language models to enhance legal case understanding, pre- and post- processing heuristic approaches to reduce the influence of irrelevant items, and learning-to-rank methods at the end to merge features with different dimensions.

Specific error analysis for Task 1 would require manual analysis of the whole dataset, which is not feasible. But we can see some approaches consolidating as main trends, such as the combination of traditional IR methods with Large Language Models. We also notice the current edition presented an additional challenge, which was the shift in the noticed cases average from the training to the

[1]"Notice" is a legal technical term that denotes a legal case description that is considered to be relevant to a query case.

[2]For a description of the previous Task 1 formulation, please see the COLIEE 2020 summary [12]

Table 1: Task 1 results

Team	F1	Precision	Recall
THUIR	0.3001	0.2379	0.4063
THUIR	0.2907	0.2173	0.4389
IITDLI	0.2874	0.2447	0.3481
THUIR	0.2771	0.2186	0.3783
NOWJ	0.2757	0.2263	0.3527
NOWJ	0.2756	0.2272	0.3504
IITDLI	0.2738	0.2107	0.3912
IITDLI	0.2681	0.2063	0.3830
JNLP	0.2604	0.2044	0.3586
NOWJ	0.2573	0.2032	0.3504
UA	0.2555	0.2847	0.2317
UFAM	0.2545	0.2975	0.2224
JNLP	0.2511	0.1971	0.3458
JNLP	0.2493	0.1931	0.3516
UA	0.2390	0.3045	0.1967
UA	0.2345	0.2400	0.2293
UFAM	0.2345	0.3199	0.1851
UFAM	0.2156	0.3182	0.1630
YR	0.1377	0.1060	0.1967
YR	0.1051	0.0809	0.1502
LLNTU	0.0000	0.0000	0.0000
LLNTU	0.0000	0.0000	0.0000

test datasets. Keeping those values close is a challenge because we rely on data provided by an external partner and which we do not fully control. Still, we intend to improve the sampling methods in order to keep the distributions in the training and test datasets as similar as possible. In the current edition, we were able to remove cases that had the exact same contents but were represented as different files in the dataset. We intend to improve the method used to identify such cases to capture minor/immaterial changes in different file contents that are likely to represent the same case.

3 TASK 2 - CASE LAW ENTAILMENT

3.1 Task Definition

Given a base case and a specific text 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 $E = \{p_1, p_2, ..., p_m \mid p_i \in P\}$ 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, 625 query cases and 22,018 paragraphs were provided for training. There were 100 query cases and 3,765 paragraphs in the testing dataset. On average, there are 35.22 candidate paragraphs for each query case in the training dataset and 37.65 candidate paragraphs for each query case in the testing dataset. The average number of relevant paragraphs for Task 2 was 1.17 paragraphs for training and 1.2 paragraphs for testing. The average query length is 35.36 words in the training set and 36.57 in the test set. The average candidate length is 102.32 words in the training set and 104.71 in the test set.

3.3 Approaches

Seven teams submitted a total of 18 runs to this task. Here, we introduce six teams' approaches that described their methods in more detail in their respective papers. One has not submitted their paper to the COLIEE 2023 workshop.

- **THUIR (3 runs)** [5] implemented the following two lexical matching methods as baselines: BM25 and QLD [17]. BM25 is a classical lexical matching model with robust performance. Their calculation formula of BM25 is shown as following.

$$BM25(d, q) = \sum_{i=1}^{M} \frac{IDF(t_i)TF(t_i, d)(k_1 + 1)}{TF(t_i, d) + k_1(1 - b + b\frac{len(d)}{avgdl}} \quad (1)$$

where k_1, and b are hyperparameters.

QLD is another representative traditional retrieval model based on Dirichlet smoothing. The equation that they used as following.

$$logp(q|d) = \sum_{i:c(q_i;d)>0} log\frac{p_s(q_i|d)}{\alpha_d p(q_i|C)} + nlog\alpha_d + \sum_i logp(q_i|C) \quad (2)$$

Furthermore, contrastive learning loss is employed to fine-tune pre-trained models of different sizes. Finally, they utilize the above features to ensemble the final score.

Their run with monoT5 [8] has the third placement and the run with ensemble placed fifth.

- **UONLP (1 run)** [3] examined the potential of an agreement-based ensemble model that incorporates two differently pretrained RoBERTa [6] models by assessing their agreement on entailment decisions in order to improve overall performance. The first RoBERTa model was pretrained on a large corpus of Canadian court cases, while the other model was pre-finetuned on a corpus of annotated entailment text pairs. Since both models had a different focus in their training data, the goal of the ensemble was to leverage both the strengths of the different models by prioritizing candidate cases that both models would agree upon. Their model was ranked in 9th place in this year's Task 2 competition.

- **JNLP (3 runs)** [2] utilized N transformer models, denoted as $M_1, M_2, ..., M_N$, respectively, where each model is associated with a specific loss function. For each query-candidate paragraph pair (q, p), they fed the pair into each of the N models to obtain the corresponding similarity scores $s_1(q, p), s_2(q, p), ..., s_N(q, p)$, where each $s_i(q, p)$ represents the similarity score computed by the i-th model. Then they added all $s_i(q, p)$ values from i=1 to N as the final similary score.

- **CAPTAIN (3 runs)** [7] proposes an approach based on the pre-trained MonoT5 sequence-to-sequence model, which is fine-tuned with hard negative mining and ensembling techniques. The approach utilizes a straightforward input template to represent the point-wise classification aspect of the model and captures the relevancy score of candidate paragraphs using the probability of the "true" token (versus the "false" token). The ensembling stage involves hyperparameter searching to find the optimal weight for

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

Team	F1-score	Precision	Recall
CAPTAIN	**0.7456**	0.7870	0.7083
CAPTAIN	0.7265	0.7864	0.6750
THUIR	0.7182	0.7900	0.6583
CAPTAIN	0.7054	0.7596	0.6583
THUIR	0.6930	0.7315	0.6583
JNLP	0.6818	0.7500	0.6250
IITDLI	0.6727	0.7400	0.6167
JNLP	0.6545	0.7200	0.6000
UONLP	0.6387	0.6441	0.6333
THUIR	0.6091	0.6700	0.5583
NOWJ	0.6079	0.6449	0.5750
NOWJ	0.6036	0.6569	0.5583
NOWJ	0.5982	0.6442	0.5583
IITDLI	0.5304	0.5545	0.5083
JNLP	0.5182	0.5700	0.4750
IITDLI	0.5091	0.5600	0.4667
LLNTU	0.1818	0.2000	0.1667
LLNTU	0.1000	0.1100	0.0917

each checkpoint. The approach achieves state-of-the-art performance in Task 2 this year, demonstrating the effectiveness of their proposed techniques.

- **NOWJ (3 runs)** [14] relies on BERT and LONGFORMER pretraining models without using any external data. Additionally, they employ an internal data generation method based on Vuong et al [13] method to overcome the lack of data and enhance the legal case retrieval process.

- **IITDLI (3 runs)** [4] has explored sparse retrieval models like BM25, as well as dense retrieval models like zero-shot T5 and GPT3.5 based reranker.

3.4 Results and Discussion

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 CAPTAIN team attained the best results. Among the three submissions from CAPTAIN, two submissions were ranked first and second. Some teams have pointed out the problem of sparse training data, so their ranking method did not achieve satisfactory performance. It may indicate that the answer paragraphs cannot be simply confirmed by information retrieval techniques. Therefore, Task 1 (information retreival task) and Task 2 (information entailment task) should be solved in a different way. Some experimental results have shown that more parameters and more legal knowledge contribute to better legal text understanding.

4 TASK 3 - STATUTE LAW INFORMATION RETRIEVAL

4.1 Task Definition

Task 3 is a task to retrieve an appropriate subset (S_1, S_2,..., S_n) of Japanese Civil Code Articles from the Civil Code texts for answering a legal bar exam question statement Q.

An appropriate subset means that the entailment system can judge whether the statement Q is true $Entails(S_1, S_2, ..., S_n, Q)$ or not $Entails(S_1, S_2, ..., S_n, notQ)$.

4.2 Dataset

For Task 3, questions related to the Japanese Civil Code were selected from the Japanese bar exam. We use a part of the Japanese Civil Code that has an official English translation (the number of articles used in the dataset is 768). The training data (the questions and corresponding article pairs) were constructed using previous COLIEE data (996 questions). For the test data, new questions selected from the 2022 bar exam are used (101 questions). 72 questions have a single relevant article and 29 questions have 2 relevant articles.

4.3 Approaches

The following 6 teams submitted their results (16 runs in total). There are two main approaches for the basic IR system. One is to use a Large Language Model (LLM) based ranking model. CAPTAIN and JNLP use monoT5 for English. HUKB and CAPTAIN use tohoku BERT [3] for Japanese. NOWJ uses bert-base-multilingual-uncased [4] for multilingual settings. The other is the keyword-based approach. HUKB, NOWJ, JNLP, UA use BM25. LLNTU and UA use TF-IDF. Four teams (CAPTAIN, HUKB, JNLP, and NOWJ) use the ensemble approach to generate final results using output from IR systems with different settings. Three teams (CAPTAIN, JNLP, and NOWJ) use ensemble to obtain results using Japanese and English.

- **CAPTAIN (3 runs)** [7] uses LLM-based ranking models; Tohoku BERT for Japanese and monoT5 for English. The best performance system uses the ensemble of these two results.
- **HUKB (3 runs)** [16] uses ensembles of keyword-based IR with different settings and LLM-based ranking models using Tohoku BERT.
- **JNLP (3 runs)** [2] uses ensembles of BM25 for Japanese and LLM-based ranking model for English; monoT5.
- **LLNTU (3 runs)** uses ordinal keyword based system (TF-IDF) and emphasizes keywords identified by named entity recognition system.
- **NOWJ (1 run)**[5] [14] uses a two-stage retrieval system that selects candidates using BM25 and re-ranks the results using an LLM-based ranking model; bert-base-multilingual-uncased for English and Japanese. They use both English and Japanese text to calculate the final score.
- **UA (3 runs)** [11] uses BM25 (UA.BM25), TF-IDF (UA.tfidf) for IR module.

[3] https://github.com/cl-tohoku/bert-japanese-whole-word-ma
[4] https://huggingface.co/bert-base-multilingual-uncased
[5] Due to the system error, two runs are withdrawn from the official evaluation.

Table 3: Evaluation results of Task 3 (Best run by teams)

Team	return	retr.	F2	prec.	rec.	MAP
CAPTAIN	143	92	0.757	0.726	0.792	0.692
JNLP	196	98	0.745	0.645	0.822	0.710
NOWJ	156	90	0.727	0.682	0.767	0.790
HUKB	174	85	0.673	0.628	0.708	0.740
LLNTU	101	74	0.653	0.733	0.644	0.764
UA	110	67	0.564	0.620	0.564	0.655

retr.: retrieved, prec.: precision, rec.: recall

4.4 Results

The table 3 shows the results of the evaluation of the submitted runs. The official metrics used in this task were macro average (average of the scores for each question over all questions) of the questions) of the F2 measure, precision and recall.

$$precision = \frac{number\ of\ retrieved\ relevant\ articles}{number\ of\ returned\ articles} \quad (3)$$

$$recall = \frac{number\ of\ retrieved\ relevant\ articles}{number\ of\ relevant\ articles} \quad (4)$$

$$f2 = \frac{5 \times precision \times recall}{4 \times precision + recall} \quad (5)$$

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) using the long ranking list (100 articles).

Table 3 shows the results of the evaluation of the submitted results. Due to the limitation of the paper length, the best performance run in terms of F2 is selected from each team run.

This year, CAPTAIN is the best run among all runs. The top four systems use ensemble settings for the various IR modules, including the LLM-based ranking model. The others use only keyword based IR. These results confirm the effectiveness of using the LLM-based ranking model.

There are a good number of questions with a single relevant article where all systems can find the relevant article. For 17 questions, all systems can find the relevant article without adding non-relevant articles (precision and recall = 1). For 12 questions, all systems can find the relevant article, but some of the systems add non relevant articles for the candidates (precision < 1 and recall = 1). On the contrary, for the answer with multiple relevant articles, there is no question that all systems can find the relevant articles. In addition, there are 5 questions where none of the systems can find the relevant article. 3 of them are questions with 2 relevant articles and 2 of them are questions with 1 relevant article.

The figures 1 and 2 show the average of the evaluation measure of all submission runs for the questions with a single relevant article[6] and those with multiple relevant articles. As we can see from comparing these two graphs, questions with multiple relevant articles are more difficult than those with a single relevant article. As we can see from the figure 2, precision is good compared to recall for the question with multiple relevant articles. This means that most systems succeed in finding the relevant article without adding irrelevant articles, but fail to find the secondary relevant article.

[6]Due to space limitations, we exclude 29 questions, all systems can find all relevant articles (recall = 1).

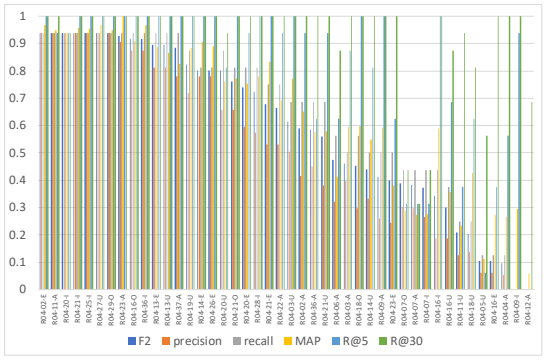

Figure 1: Averages of precision, recall, F2, MAP, R_5, and R_30 for questions with a single relevant article

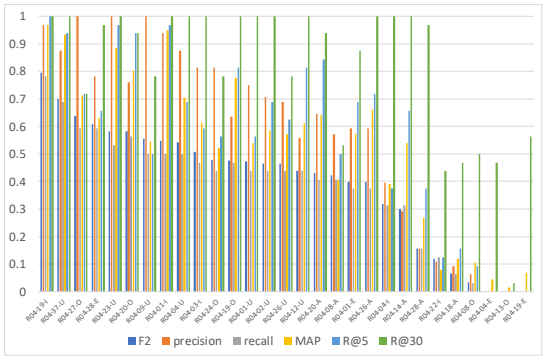

Figure 2: Averages of precision, recall, F2, MAP, R_5, R_10, and R_30 for multiple relevant articles

4.5 Discussion

The top three systems use both English and Japanese questions with LLM. The effect of using different languages (English and Japanese) for the COLIEE task was previously discussed in [15]. For example, the Japanese IR system works well for the question of when the legal terminology represented by Chinese characters is effective for finding the relevant article. Therefore, the ensemble of results from the multilingual IR system improves the performance of the monolingual IR system. However, it is not common that the user can provide such information in two languages. It may be better to prohibit the use of questions in both languages. Even for such a case, we can use machine translation system to have questions in two languages.

One of the characteristics of the difficult questions are those that use anonymized symbols such as “A” and “B” to refer to persons or other entities. There are 41 questions that use such symbols. The table 4 shows the number of questions for the F2 measure (average) that are classified as having an anonymized symbol or not. From this table we can see that the questions with anonymized symbols are more difficult than those without.

Table 4: Number of questions classified by F2 score and question type

F2	Anonymize	Other
0-0.2	6	6
0.2-0.4	7	5
0.4-0.6	13	9
0.6-0.8	2	8
0.8-1.0	13	32

Another key factor is the number of relevant articles. HUKB tries to deal with the effectiveness of combining two or more articles, but other teams only check the relevance between the question and the articles to select the relevant ones. More research is needed to find secondary relevant articles.

Another type of difficult question is the existence of other articles that share common terms. For example, question R04-16-E is about "tender of services" which is discussed in article 492. However, there are many keywords related to "acceptance" which is discussed in article 413. The LLM based system can handle the context, but it is still difficult to select the most important part to find the relevant article.

5 TASK 4 - STATUTE LAW TEXTUAL ENTAILMENT AND QUESTION ANSWERING

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 after 2019, 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. We also require the participants to make their system reproducible in the academic standard, i.e. they should describe which methods and what datasets were used to a reproducible extent. 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 organizers provided a data set used for previous campaigns as training data (??? questions) and new questions selected from the 2023 bar exam as test data (101 questions).

5.3 Approaches

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

- **AMHR (3 runs)** [1] **AMHR01** employed 2-shot prompting using the FlanT5-XXL model from Google Research on HuggingFace [7], where the shots, balanced by label, were chosen from the train set using TF-IDF similarity metric to each example at inference time. **AMHR02** used A couple of publicly available models to assemble an ensemble of few-shot prompted models. These models and their hyperparameters were chosen using grid search. **AMHR03** employed 6-shot prompting using the GPT-4 model [8]. The shots were chosen from the train set using TF-IDF similarity metric to each example at inference time. Five runs were performed, and majority vote was used to select the final prediction. This run is excluded in the formal run list due the OpenAI API provides insufficient reproducibility.
- **CAPTAIN (3 runs)** [7] **CAPTAIN.run1** split each article and query to pairs of (condition, statement) and consider the consensus of the conditions and the statements between the query and an article by Electra[9]. **CAPTAIN.run2** chunked articles to phrase (using n-gram model), encoded all phrases and query by BERT and train a SVM model for classifying. The result finally is ensemble with **CAPTAIN.run1** to get the final result. **CAPTAIN.gen** matched pair question with summaries of relevant article for classifying the label by BERT[10].
- **HUKB (3 runs)** [16] **HUKB1** used Japanese pretrained BERT[11]. **HUKB2** used Task 3 retrieval system for subarticles to select appropriate part of the article and applied the same BERT system for generating final result. **HUKB3** used their BERT-based Task 3 retrieval system for the subarticles to select appropriate part of the article and apply same BERT system for generating final result. Their systems are almost equivalent to their system in COLIEE 2022.
- **JNLP (3 runs)** [2] **JNLP** used zero-shot models of LLMs, by gathering all the prompts from the GLUE tasks available in the PromptSource library, selected 56 prompts. **JNLP1** used google/flan-t5-xxl model[12] , **JNLP2** used google/flan-ul2 model[13], **JNLP3** used declare-lab/flan-alpaca-xxl model[14], respectively, to run the prompts which were the given problem-article pairs inserted.
- **KIS (3 runs)** [10] **KIS** extended their previous system which performs data augmentation and ensemble of BERT-based models and rule-based models, to integrate LUKE[15], the named entity enhanced Transformer. **KIS1** uses the pretrained LUKE model, **KIS2** used a fine-tuned LUKE model for alpabetical person included dataset, **KIS3** used another

[7] https://huggingface.co/google/flan-t5-xxl
[8] https://openai.com/blog/openai-api
[9] google/electra-base-discriminator
[10] cl-tohoku/bert-base-japanese-whole-word-masking
[11] cl-tohoku/bert-base-japanese-whole-word-masking
[12] https://huggingface.co/google/flan-t5-xxl
[13] https://huggingface.co/google/flan-ul2
[14] https://huggingface.co/declare-lab/flan-alpaca-xxl
[15] luke-japanese-base-lite

fine-tuned LUKE model without the alpabetical person included dataset.

- **LLNTU (3 runs) LLNTU** used Disjunctive Union of Longest Common Subsequence, and adjusting them from similarity and length.
- **NOWJ (3 runs)** [14] **NOWJ** used multi-task model with pre-trained Multilingual BERT[16] as backbone; **NOWJ.multiv1-en** employed the English data for the training phase, **NOWJ.multi-v1-en** employed Japanese data for the training phase, and **NOWJ.multijp** also utilized Japanese data with different inference strategy.
- **TRLABS (3 runs)** Their three runs directly use GPT-4 with zero-shot prompting, prompt with IRAC legal reasoning approach (**TRLABS_I**), prompt with TREACC legal reasoning approach (**TRLABS_T**), no-legal reasoning approach prompted just asked to analyze Hypothesis given the Premise (**TRLABS_D**). Due to the reproducibility issue of GPT-4, these runs are not regarded as formal results.
- **UA (2 runs)** [11] Their system incorporates the semantic information into the BERT to help the pragmatic reasoning, for natural language inference. **UA_V1** fine-tuned on DeBERTa-small and **UA_V2** fine-tuned on DeBERTa-large model.

5.4 Results and Discussion

Table 5 shows the COLIEE 2023 Task 4 formal run results. The Formal Run (R04) column shows the result of the COLIEE 2023 formal run using the latest Japanese legal bar exam (Year R04). The columns of R02 and R01 are the results using the past formal run datasets, which we required participants to submit, in order to compare different datasets for reference due to the smallness of our datasets, while these datasets were already made public as part of our training dataset.

The lower part of the table shows runs with prefixes of "*", which used external services where its detailed architecture, training datasets, model weights are not available, resulting in inreproducible outputs that are prohibited in our participation call.

The best runs by team **JNLP** used LLMs in a straightforward way. The second best runs by team **KIS** used BERT and rule-base systems, which is an extension of their previous system, the best one in COLIEE 2022. Comparing results of the past formal run settings (R02 and R01), we found that the rankings switch between these runs from this year's formal run. In the R01 dataset, the best run was **AMHR01**, which also uses an LLM. These suggest that the accuracies still depend on the characteristics of each year's dataset, while LLMs are, at least, comparable or better to the existing models.

A concern of the LLMs is that we do not completely grasp what texts are used to train the LLMs; they could include very similar texts with the COLIEE's problem/answer texts. This is fine if we simply expect the systems to answer Yes/No in any way, but would not work, especially when logics are required in the statute law, in practical use cases.

Another issue to discuss is the reproducibility of the external resources, e.g. OpenAI's ChatGPT and GPT-4. Some of the teams employed those services, but they could change monthly, weekly, or even daily; we do not know what dataset was used in their training. Usage of such inreproducible services would not fit with academic discussions, as things become just a guess.

Even though, there are certain interests to what extent those services could solve COLIEE problems. We asked ChatGPT (GPT-3.5) to answer the COLIEE problems, by straightforward prompts of "please answer yes or no given the following question:" (in Japanese) with the given problem texts as they are. ChatGPT sometimes shows evidences which are inappropriate or wrong even if the Yes/No answer itself is correct, such as information which are not related to answer, almost repeats of the problem text, not handling "except listed below".

As these wrong examples suggest, LLMs would create their answers not logically, but their huge stack of similar contents led the answers and evidences; LLMs might not perform logical calculations. Because Task 4 is intended to a pure textual entailment task, superficial similarities without logical calculations would not make much sense. However, as a practical legal application, it can be useful when there are, to some extent, similar contents available as previous existing cases. As our future work, we need to detect which answers are answered with strong certainty, excluding not just hallucinations but such uncertain wrong answers, if we use LLMs in this way.

6 CONCLUSION

We have summarized the systems and their performance as submitted to the COLIEE 2023 competition. For Task 1,[please add any conclusion message]. In Task 2, the winning team was CAPTAIN, and they used an approach based on the pre-trained MonoT5 sequence-to-sequence model, which is fine-tuned with hard negative mining and ensembling techniques to achieve an F1 score of 0.7456. For Task 3, [please add any conclusion message]. Lastly, for Task 4, [please add any conclusion message]. We intend to further continue to improve dataset quality in future editions of COLIEE so the tasks more accurately represent real-world problems.

7 ACKNOWLEDGMENTS

This competition would not be possible without the significant support of Colin Lachance from vLex, Compass Law and Jurisage, 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), [including DGECR-2022-00369, RGPIN-2022-0346], and the support of Alberta Innovates. This work was supported by JSPS KAKENHI Grant Numbers, 22H00543 and JST, AIP Trilateral AI Research, Grant Number JPMJCR20G4.

REFERENCES

[1] Onur Bilgin, Logan Fields, Antonio Laverghetta Jr., Zaid Marji, Animesh Nighojkar, Stephen Steinle, and John Licato. 2023. AMHR Lab 2023 COLIEE Competition Approach. In *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL).*

[16] bert-base-multilingualuncased

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

Team	Submission ID	L	Formal Run (R04)		R02		R01	
			Correct Answers	Accuracy	Correct Answers	Accuracy	Correct Answers	Accuracy
No to All	BaseLine	-	No 52/All 101	0.5149	No 43/All 81	0.5309	Yes 59/All 111	0.5315
JNLP	JNLP3	E	79	0.7822	65	0.8025	72	0.6486
JNLP	JNLP1	E	76	0.7525	66	0.8148	75	0.6757
JNLP	JNLP2	E	76	0.7525	63	0.7778	75	0.6757
KIS	KIS2	J	70	0.6931	58	0.7160	77	0.6937
KIS	KIS1	J	68	0.6733	56	0.6914	74	0.6667
UA	UA_V2	?	67	0.6634	N/A	N/A	N/A	N/A
AMHR	AMHR01	E	66	0.6535	65	0.8025	79	0.7117
KIS	KIS3	J	66	0.6535	54	0.6667	73	0.6577
AMHR	AMHR03	E	65	0.6436	63	0.7778	49	0.4414
LLNTU	LLNTUdulcsL	J	63	0.6238	42	0.5185	55	0.4955
UA	UA	?	63	0.6238	61	0.7531	67	0.6036
HUKB	HUKB2	J	60	0.5941	50	0.6173	60	0.5405
CAPTAIN	CAPTAIN.gen	J	59	0.5842	55	0.6790	65	0.5856
CAPTAIN	CAPTAIN.run1	E	58	0.5743	41	0.5062	67	0.6036
LLNTU	LLNTUdulcsS	J	57	0.5644	44	0.5432	50	0.4505
HUKB	HUKB1	J	56	0.5545	41	0.5062	67	0.6036
HUKB	HUKB3	J	56	0.5545	48	0.5926	61	0.5495
LLNTU	LLNTUdulcsO	J	56	0.5545	44	0.5432	49	0.4414
NOWJ	NOWJ.multi-v1-jp	J	55	0.5446	N/A	N/A	N/A	N/A
CAPTAIN	CAPTAIN.run2	E	53	0.5248	42	0.5185	67	0.6036
NOWJ	NOWJ.multijp	J	53	0.5248	N/A	N/A	N/A	N/A
NOWJ	NOWJ.multi-v1-en	E	49	0.4851	N/A	N/A	N/A	N/A
AMHR	*AMHR02	E	82	0.8119	66	0.8148	89	0.8018
TRLABS	*TRLABS_D	E	79	0.7822	68	0.8395	90	0.8108
TRLABS	*TRLABS_I	E	79	0.7822	71	0.8765	87	0.7838
TRLABS	*TRLABS_T	E	76	0.7525	71	0.8765	87	0.7838

[2] Quan Minh Bui, Dinh-Truong Do, Nguyen-Khang Le, Dieu-Hien Nguyen, Khac-Vu-Hiep Nguyen, Trang Pham Ngoc Anh, and Minh Nguyen Le. 2023. JNLP COLIEE-2023: Data Argumentation and Large Language Model for Legal Case Retrieval and Entailment. In *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL).*

[3] Michel Custeau and Diana Inkpen. 2023. Individual Models Can Perform Better than Agreement-Based Ensembles. In *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL).*

[4] Rohan Debbarma, Pratik Prawar, Abhijnan Chakraborty, and Srikanta Bedathur. 2023. IITDLI: Legal Case Retrieval Based on Lexical Models. In *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL).*

[5] Haitao Li, Weihang Su, Changyue Wang, Yueyue Wu, Qingyao Ai, and Yiqun Liu. 2023. THUIR@COLIEE 2023: Incorporating Structural Knowledge into Pre-trained Language Models for Legal Case Retrieval. In *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL).*

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692.*

[7] Chau Nguyen and Minh-Le Nguyen. 2023. CAPTAIN at COLIEE 2023: Efficient Methods for Legal Information Retrieval and Entailment Tasks. In *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL).*

[8] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020.* Association for Computational Linguistics, Online, 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[9] Luisa Pereira Novaes, Daniela Vianna, and Altigran da Silva. 2023. A Topic-Based Approach for the Legal Case Retrieval Task. In *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL).*

[10] Takaaki Onaga, Masaki Fujita, and Yoshinobu Kano. 2023. Japanese Legal Bar Problem Solver Focusing on Person Names. In *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL).*

[11] Juliano Rabelo, Mi-Young Kim, and Randy Goebel. 2023. Transformer-based Legal Information Extraction. In *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL).*

[12] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2021. *COLIEE 2020: Methods for Legal Document Retrieval and Entailment.* 196–210. https://doi.org/10.1007/978-3-030-79942-7_13

[13] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (apr 2009), 333–389. https://doi.org/10.1561/1500000019

[14] Thi-Hai-Yen Vuong, Hai-Long Nguyen, Tan-Minh Nguyen, Hoang-Trung Nguyen, Thai-Binh Nguyen, and Ha-Thanh Nguyen. 2023. NOWJ at COLIEE 2023 - Multi-Task and Ensemble Approaches in Legal Information Processing. In *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL).*

[15] Masaharu Yoshioka. 2018. Analysis of COLIEE Information Retrieval Task Data. In *New Frontiers in Artificial Intelligence*, Sachiyo Arai, Kazuhiro Kojima, Koji Mineshima, Daisuke Bekki, Ken Satoh, and Yuiko Ohta (Eds.). Springer International Publishing, Cham, 5–19.

[16] Masaharu Yoshioka and Yasuhiro Aoki. 2023. HUKB at COLIEE 2023 Statute Law Task. In *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL).*

[17] ChengXiang Zhai. 2008. Statistical language models for information retrieval. *Synthesis lectures on human language technologies* 1, 1 (2008), 1–141.