Noname manuscript No.
(will be inserted by the editor)

Overview and Discussion of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021

Juliano Rabelo · Randy Goebel · Mi-Young Kim · Yoshinobu Kano · Masaharu Yoshioka · Ken Satoh

Received: date / Accepted: date

Abstract We summarize the 8th Competition on Legal Information Extraction and Entailment. In this edition, the competition included five 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), an entailment/question answering task based on retrieved civil code statutes (Task 4) and an entailment/question answering task without retrieved civil code statutes (Task 5). Participation was open to any group based on any approach. Eight different teams participated in the case law competition tasks, most of them in more than one task. We

J. Rabelo
Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Alberta, Canada
E-mail: rabelo@ualberta.ca

R. Goebel
Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Alberta, Canada
E-mail: rgoebel@ualberta.ca

M. Kim
Dept. of Science, Augustana Faculty, and Alberta Machine Intelligence Institute, University of Alberta, Alberta, Canada
E-mail: miyoung2@ualberta.ca

Y. Kano
Faculty of Informatics, Shizuoka University, Hamamatsu, Shizuoka, Japan
E-mail: kano@inf.shizuoka.ac.jp

M. Yoshioka
Faculty of Information Science and Technology, Hokkaido University, Sapporo-shi, Hokkaido, Japan
E-mail: yoshioka@ist.hokudai.ac.jp

K. Satoh
National Institute of Informatics, Chiyoda-ku, Tokyo, Japan
E-mail: ksatoh@nii.ac.jp

received results from 6 teams for Task 1 (16 runs) and 6 teams for Task 2 (17 runs). On the statute law task, there were 8 different teams participating, most in more than one task. Six teams submitted a total of 18 runs for Task 3, 6 teams submitted a total of 18 runs for Task 4, and 4 teams submitted a total of 12 runs for Task 5. Here we summarize the approaches, our official evaluation, and analysis on our data and submission results.

Keywords COLIEE2021 · legal information retrieval · legal information entailment

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.

1 Introduction

The objective of the Competition on Legal Information Extraction/Entailment (COLIEE) is to build a research community and establish 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 with the International Conference on AI and Law (ICAIL), which was the case in 2017, 2019, and again in 2021. 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, and the 2021 edition included a fifth task (entailment in statute law text without relying on previously retrieved data).

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 are hypothesized to 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 relevant to the difficulty of the questions. For the entailment task (Task 4), we performed categorized analyses to expose different issues of the problems and characteristics of the submissions, in addition to the evaluation accuracy as in previous COLIEE tasks. Task 5 is similar to Task 4, but competitors can not rely on previously retrieved statute data.

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 6 presents some final remarks.

2 Task 1 - Case Law Retrieval

2.1 Task Definition

The Case Law Retrieval Task consists in finding which cases should be "noticed"[1] with respect to a given query case. More formally, given a set of cases C, a set of query cases Q, a set of the true noticed cases N, and a set of false noticed cases F, such that $C = \{Q \cup N \cup F\}$, the Task is to find the set of answers $A = \{A_1 \cup A_2 ... \cup A_n\}$, such that $n = |Q|$ and each $A_i \subset N$ contains

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

all the true noticed cases and only the true noticed cases with respect to the query case $q_i \in Q$.

2.2 Dataset

The dataset is comprised of 4,415 case law files. A labelled training set of 650 cases is provided, together with a total of 3,311 true noticed cases. At first glance, the task may seem simple, as one could think competitors need to identify the 3,311 cases among the 4,415 total cases. However, the task actually requires competitors to identify the noticed cases for each given query case. On average, there are approximately 5 noticed cases per query case in the provided training dataset, which should be identified among the 4,415 cases. To prevent merely using citations of past cases, citations are suppressed from the case contents and replaced by a "FRAGMENT_SUPPRESSED" tag indicating that fragment was removed.

A test set is given with 250 query cases and a total of 900 true noticed cases, which means there are on average 3.6 noticed cases per query case in the test dataset. In future editions, we intend to ensure that the training and test datasets have similar distributions. Initially, the golden labels for that test set is not provided to competitors.

2.3 Approaches

We received 15 submissions from 7 different teams for Task 1, but only 5 teams submitted papers describing their approaches. Their methods are briefly described below. Please refer to the corresponding papers for further details.

- Li et al. [11] (**team name: siat**) propose a pipeline method based on statistical features and semantic understanding models, which enhances the retrieval method with both recall and semantic ranking. siat's best submission had an f1-score of 0.030.
- Schilder et al. [21] (**team name: TR**) applies a two-phase approach for Task 1: first, they generate a candidate set which tentatively contains all true noticed cases but eliminates some of the false candidates (i.e., this step is optimized for recall). The second step is a binary classifier which receives as input the pair $(query\ case, candidate\ case)$ and predicts whether they represent a true noticed relationship.
- Rosa et at. [20] (**team name: NM**) presents a vanilla application of BM25 to the case law retrieval problem. They do that by first indexing all base and candidate cases contained in the dataset. Before indexing, each document is split into segments of texts using a context window of 10 sentences with overlapping strides of 5 sentences (which are called 'candidate case segments'). BM25 is then used to retrieve candidate case segments for each base case segment. The relevance score for a $(base\ case, candidate\ case)$ pair is the maximum score among all their base case segment and candidate

case segment pairs. The candidates are then ranked according to threshold-based heuristics. The NM team submitted only one run, which was ranked second place among all submissions with an f1-score of 0.0937.
- Ma et al. [13] (**team name: TLIR**) was the top ranked team for Task 1. They apply two methods: the first is a traditional language model for IR (LMIR) [2], which consists of an application of LMIR on a pre-processed version of the dataset. The TLIR team did not use the full case contents, but cleverly made use of the tags inserted in the text to indicate a fragment has been suppressed in order to heuristically identify the potentially most relevant text fragments. The fact this approach ranked first place among all Task 1 competitors indicates traditional IR methods can achieve good results in the case law retrieval task. The second approach is a transformer based method, which factors a document into paragraphs and then computes measures on interactions between paragraphs using BERT. Compared with other neural models, BERT-PLI can take long text representations as an input without truncating them at some threshold. Yet, the results attained with this approach in COLIEE 2021 were not as good as the simpler IR-based approach, ranking at third and fifth places among all submission with an f1-score of 0.0456 and 0.0330.
- Althammer et al. [1] (**team name: DSSIR**) combine retrieval methods with neural re-ranking methods using contextualized language models like BERT. Since the cases are typically long documents exceeding BERT's maximum input length, the authors adopt a two phase approach. The first phase combines lexical and dense retrieval methods on the paragraph-level of the cases. They then re-rank the candidates by summarizing the cases and then apply a fine-tuned BERT re-ranker on said summaries. Their best ranking submission attained fourth place overall, with an f1-score of 0.0411.

2.4 Results and Discussion

Table 1 shows the results of all submissions received for Task 1 in COLIEE 2021. A total of 15 submissions from 7 different teams have been received. It can be seen the f1-scores were, in general, much lower than in previous editions, reflecting the fact the task is now more challenging than its previous formulation. The best performing team in Task 1 in the 2020 edition, for example, achieved an f1-score of 0.6774. For more information on the previous task formulation and approaches, please see the COLIEE 2020 summary [16].

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 TLIR, with an f1-score of 0.1917, with an approach that combined traditional IR methods with simple heuristics to identify the most relevant fragments in a case law. Also worth mentioning is the NM team, whose approach was a vanilla application of BM25 and achieved the second place overall.

Table 1: Task 1 results

Team	File	F1
TLIR	run1.txt	0.1917
NM	NM_Run_Task 1_BM25.txt	0.0937
TLIR	run3.txt	0.0456
DSSIR	run_test_bm25.txt	0.0411
TLIR	run2.txt	0.0330
siat	siatEMB_result-Task 1.txt	0.0300
siat	siatEMB2_result-Task 1.txt	0.0291
DSSIR	run_test_vanillabert.txt	0.0279
DSSIR	run_test_bm25_dpr.txt	0.0272
MAN01	[MAN01] Task 1 run0.txt	0.0073
TR	TR_run1.csv	0.0046
JNLP	JNLP.taks1.BM25SD_3_7.txt	0.0019
JNLP	JNLP.taks1.BM25SD_7_3.txt	0.0019
JNLP	JNLP.taks1.SD.txt	0.0009
TR	TR_run2.csv	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. Besides that, we will fix a few minor issues which were found in the dataset, such as two different files with the exact same contents (i.e., the same case represented as two separate cases). 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 an impact on the final results.

A known issue with the dataset is that tags inserted to indicate suppression of fragments provide an artificial clue as to where there is potentially highly relevant contents. That aspect was exploited by the winning team in COLIEE 2021. Whereas that is not a problem with that team's approach, we would like our datasets to represent as accurately as possible real-world problems, so options to improve such datasets will be explored in future editions.

3 Task 2 - Case Law Entailment

3.1 Task Definition

Task 2 is a legal case entailment task and it involves the identification of a paragraph from existing cases that can be claimed to entail the decision of a new case. Given a decision Q of a new case and a relevant case R, the challenge is to identify a specific paragraph in R that entails the decision Q. The organizers have confirmed that the answer paragraph cannot be identified merely by information retrieval techniques using some examples. Because the case R is a relevant case to Q, many paragraphs in R could be relevant to Q, regardless of confirming entailment. This task requires one to identify a paragraph which entails the decision of Q, so required is a specific entailment method

Table 2: Dataset information in Task 2

Task 2	**Train**	**Test**
# Query case	426	100
# Candidate paragraphs/Query	35.72	35.24
# Entailing paragraphs/Query	1.17	1.17

that compares the meaning of each paragraph in R and the decision in Q. The data are drawn from an existing collection of predominantly Federal Court of Canada case law documents. The evaluation measure will be precision, recall and F-measure.

For COLIEE 2021, the Task 2 training and testing sets contain 426 and 100 base cases respectively. Table 2 shows the dataset information for Task 2.

Training data is provided in the form of triples, each consisting of a query, a noticed case, and a paragraph number of the noticed case by which the decision of the query is entailed. Here, "noticed case" means the relevant case of the query. An example is shown in Table 3.

3.2 Approaches

Seven teams participated in Task 2, and a total of 17 results were submitted (average 2.43 results per team). Each team was allowed to submit a maximum of three results. Table 4 shows the approaches that teams used in Task 2. Althammer et al. [1] (team name:DSSIR) used either BM25 or DPR [8] model to produce the first two results, which were trained on the entailing paragraph pairs in order to rank each paragraph in the noticed case, given the query paragraph. They also combined the ranking of BM25 and DPR as their third result.

Schilder et al. [21] (team name: TR) used hand-crafted similarity features and applied a classical random forest classifier. Using n-gram vectors, universal sentence encoder vectors, and averaged word embedding vectors, they computed the similarity between each paragraph in the noticed case and the decision fragment in the query. After selecting the most similar k paragraphs, they trained a random forest classifier.

Kim et al. [9] (team name: UA) used BERT pre-trained on a large (general purpose) dataset by fine-tuning on the provided training dataset. If the tokenization step produced more than the 512 token limit, they apply another transformer-based model to generate a summary of the input text, and then process the pair again. Since the input text often includes text in French, they apply a simple language detection model based on naive Bayesian filter to remove those fragments. There are usually very few actual entailing paragraphs in a case (by far, most of the cases only have one entailing paragraph). So in the post-processing step they establish limits for the maximum number of outputs allowed per case. At the same time, they observe a minimum score in an attempt to reduce the number of the false positives.

Table 3: Training data Example in Task 2

base case	B232 arrived in Canada with 491 other persons aboard the MV Sun Sea...
decision	Given that the Respondent remains a security risk whom the Minister has...
p#1 in noticed case	Previous decisions to detain the individual must be...
p#2 in noticed case	The Ministers are requesting an order...
...	...
p#32 in noticed case	THIS COURT ORDERS that the stay motion be granted until the final ...
entailing paragraph	#27

Table 4: Approaches in Task 2

Team	**Approaches**
DSSIR	BM25 or DPR model
TR	hand-crafted similarity features and random forest classifier
UA	BERT and naive bayesian filtering
siat	BERT, n-gram masking, data augmentation and Fast Gradient method
JNLP	supporting model, lexical model and NSFP model
NM	monoT5-zero-shot and DeBERTa

Li et al. [11] (team name: siat) proposed a pre-training Task on BERT (BERT-base-uncased) with dynamic N-gram masking, to get a special BERT model with legal knowledge (BERTLegal). They utilized n-gram masking to generate masked inputs for what they call "masked language model" targets. The length of each n-gram mask is randomly selected amongst 1, 2, and 3. They also did data augmentation and used a Fast Gradient method.

Nguyen et al. [14] (team name: JNLP) used the supporting model and lexical model for two submissions, and in the last submission, they used a neighbouring structures fingerprint (NSFP) model.

[19] (team name: NM) used monoT5-zero-shot, monoT5 and DeBERTa [7]. They also evaluated an ensemble of their monoT5 and DeBERTa models. The model monoT5-zero-shot is a sequence-to-sequence adaptation of the T5 [17] model.

We were not able to identify the approach of the team MAN01 as there was no corresponding paper submission.

3.3 Evaluation Measure

Task 2 uses micro-average precision, recall and F1-measure as evaluation metrics, which are formulated as follows:

$$Precision = \frac{N_{TP}}{N_{TP} + N_{FP}}, \tag{1}$$

$$Recall = \frac{N_{TP}}{N_{TP} + N_{FN}}, \tag{2}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{3}$$

where N_{TP} denotes the number of true positive prediction for all queries, $N_{TP} + N_{FP}$ is the total positive prediction number for all queries, and $N_{TP} + N_{FN}$ is the ground truth positive case number.

3.4 Results and Discussion

Table 5 shows the Task 2 results. NM team's three submissions are all ranked no.1 to no.3. In particular, their Ensemble of DeBERTa and monoT5 showed the best performance with the F1 score of 0.6912. As shown in Table 6, the systems of the the winning team (NM) show balanced performance between precision and recall. This task is to find the paragraph(s) that entails the decision of the query, and in most cases, only one paragraph is the correct answer. So, systems are likely to show better precision than recall. An interesting observation in Table 6 is that the system *monoT5* showed better recall than precision.

Most of the systems combined the traditional BM25 information retrieval algorithm and BERT Transformer language model. They showed that the traditional BM25 system is still useful in legal information retrieval and entailment. To solve the issue of the dataset imbalance, some teams tried data augmentation. In addition, some approaches tried to extract semantic relationships between paragraphs using BERT. Finally, there was an approach to use LEGAL-BERT, a BERT system optimized for the legal domain, but the performance was not promising.

Participants have stated that the extreme class-imbalance nature of the problem and the limited data size make it challenging to train an efficient and generalizable classification model. Because of the limited data size, the winning team (NM) adopted zero-shot models, and they showed that zero-shot models can have at least equivalent performance to models that have been fine-tuned on a legal case entailment task. They also confirmed a counter-intuitive result: that models with little or no adaption to the target task can be more robust to changes in the data distribution than models that have been carefully fine-tuned to the task at hand.

4 Task 3 - Statute Law Information Retrieval

4.1 Task Definition

Task 3 requires the retrieval of an appropriate subset (S_1, S_2,..., S_n) of Japanese Civil Code Articles from the Civil Code texts dataset, used for answering a Japanese legal bar exam question Q.

Table 5: Task 2 official results

Team	File	F1
NM	Run_Task2_DebertaT5.txt	0.6912
NM	Run_Task2_monoT5.txt	0.6610
NM	Run_Task2_Deberta.txt	0.6339
UA	UA_reg_pp.txt	0.6274
JNLP	JNLP.Task2.BM25Sup._Den..txt	0.6116
JNLP	JNLP.Task2.BM25Sup._Den._F..txt	0.6091
UA	UA_def_pp.txt	0.5875
JNLP	JNLP.Task2.NFSP_BM25.txt	0.5868
siat	siatCLS_result-Task2.txt	0.5860
DSSIR	run_test_bm25.txt	0.5806
siat	siatFGM_result-Task2.txt	0.5670
UA	UA_loose_pp.txt	0.5603
TR	Task 2_TR.txt	0.5438
DSSIR	run_test_bm25_dpr.txt	0.5161
DSSIR	run_test_dpr.txt	0.5161
MAN01	[MAN01] Task 2 run1.txt	0.5069
MAN01	[MAN01] Task 2 run0.txt	0.2500

Table 6: Task 2 winning team's detailed performance

Submission name	F1	Prec	Recall
Deberta	0.6339	0.6635	0.6068
monoT5	0.6610	0.6554	0.6666
DebertaT5	0.6912	0.7500	0.6410

An appropriate subset means the identification of a subset of statutes for which an entailment system can judge whether the statement Q is true $Entails(S_1, S_2, ..., S_n, Q)$ or not $Entails(S_1, S_2, ..., S_n, \neg Q)$.

4.2 Dataset

For Task 3, questions related to Japanese civil law were selected from the Japanese bar exam. Since there were some updates of Japanese Civil Code on April 2020, we revised the text database to reflect this revision for Civil Code, and its translation into English. However, since the English translated version is not provided for a portion of this code, we exclude those untranslated parts from the civil code text and their related questions. As a result, the number of civil code articles used in the dataset is 768, or about half of previous COLIEE competitions. Training data (the questions and relevant article pairs) were constructed by using previous COLIEE data (806 questions). In this data, questions related to revised articles are reexamined and those for excluded articles are removed from the training data. For the test data, new questions were selected from the 2020 bar exam (81 questions).

The number of questions classified by the number of relevant articles is listed in Table 7.

Table 7: Number of questions classified by number of relevant articles

number of relevant article(s)	1	2	4	*total*
number of questions	65	14	2	*81*

4.3 Approaches

The following 6 teams submitted their results (18 runs in total). We describe approaches for each team as follows, using a header format of the form **Team Name (number of submitted runs)**. All teams had experience in submitting results in previous competition. Because the best performance system [22] of COLIEE 2020 uses BERT [5], most of the teams (HUKB, JNLP OvGU, and TR) use BERT and ensemble results with an ordinary IR system (HUKB and OvGU). One characteristic feature proposed in this year's task is extension of training data for BERT-based IR system training. OvGU proposed a method to extend the contents of original article using text data related to the article (metadata, text from the website). JNLP proposed a method to select a corresponding part of the article for the query using a sliding window mechanism. HUKB proposed a method to add detailed information from the referred articles. Other common techniques used in the system were well known IR engine mechanisms such as BM25, TF-IDF, Indri [23], and Word Movers' Distance (WMD) [10].

- **HUKB (three runs)** [27] uses a BERT-based IR system and Indri for the IR module, and compares the result of each system output to create final results. They construct a new article database with the following two types: one expands the detailed information using the referred article, and the other uses text splitting for describing one judicial decision. They submitted three runs with almost similar settings and the best run is HUKB-3.
- **JNLP (three runs)** [14] uses a BERT-based IR models that combines multiple BERT models for generating results. They also construct training data of relevant articles by selecting the most relevant part of the article using a sliding window. The best run is JNLP.CrossLMultiLThreshlod that uses an ensemble of three different systems outputs by selecting the highest result among them.
- **LLNTU (three runs)** has not submitted a paper describing their methods.
- **OvGU (three runs)** [25] uses a variety of BERT models with different data enrichment techniques. The best run is OvGU_run1 that uses sentence-BERT embedding [18] with TF-IDF by enriching the articles in the training data by using metadata, text from the web data related to the article and relevant queries from training data.
- **TR (three runs)** [21] submits three runs and the best run is TR_HB uses Word Mover's Distance (WMD) approach to calculate the similarity between query and articles.

Table 8: Evaluation results of submitted runs (Task 3)

sid	ret.	retr.	F2	Prec.	Rec.	MAP	R_5	R_{10}	R_{30}
OvGU_run1	134	71	0.743	0.687	0.790	0.762	0.762	0.822	0.861
JNLP. CrossLMultiL Threshold	156	76	0.735	0.612	0.815	0.805	0.792	0.891	0.950
BM25.UA	81	62	0.722	0.765	0.716	0.768	0.723	0.743	0.822
R3.LLNTU	114	67	0.692	0.653	0.731	0.779	0.792	0.832	0.911
TR_HB	162	55	0.533	0.340	0.630	0.675	0.723	0.752	0.851
HUKB-3	241	63	0.531	0.294	0.710	0.621	0.693	0.752	0.871

ret. (return), retr. (retrieved), Prec. (Precision), Rec. (Recall)

– **UA (three runs)** [9] uses ordinary IR modules for generating results. The best run is BM25.UA that uses BM25 as an IR module.

4.4 Results and Discussion

Table 8 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 the F2 measure, precision, and recall (See Appendix 7 for the definition of those measures).

We also calculate the mean average precision (MAP) and recall at k (R_k: recall is calculated by using the top k ranked documents as returned documents) by using the long ranking list (100 articles). Table 8 shows the results of the evaluation of submitted results[2].

This year, OvGU is the best run among all runs. JNLP achieves almost similar score and have higher MAP. This year, ordinary IR model BM25 achieves good performance for finding 1 relevant article for the question. From this results, we confirm the effectiveness of using deep learning technology such as BERT for this task.

Figures 1, 2, and 3 show the average of evaluation measure for all submission runs. As we can see from Figure 1, there are many easy questions for which almost all system can retrieve the relevant article. The easiest question is R02-10-E "An underground space or airspace may be established as the subject matter of superficies for ownership of structures, through the specification of upper and lower extents." whose relevant article (Article 269-2) has the same sentence in the text.

However, there are five queries for which none of the system can retrieve the relevant articles. All questions (R02-9-E, R02-15-I, 02-15-U, 02-15-E, and R02-23-E) are based on the use case of the article that requires semantic matching and handling anonymized symbols such as "A" and "B" for referring person or other entities. For example, question of R02-9-E is "B obtained A's bicycle by fraud. In this case, A may demand the return of the bicycle against B by

[2] Due to errors in the evaluation data, this result is different from the one used in the workshop paper. However, the order of teams is same as that in the workshop.

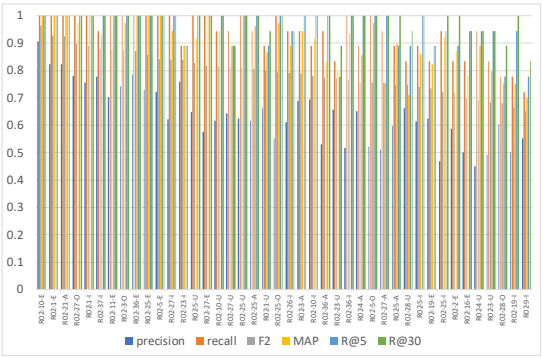

Fig. 1: Averages of precision, recall, F2, MAP, R_5, and R_30 for easy questions with a single relevant article

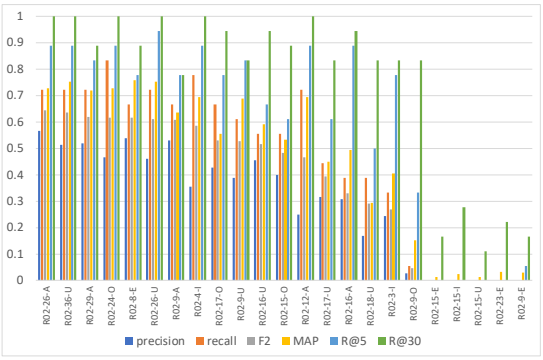

Fig. 2: Averages of precision, recall, F2, MAP, R_5, and R_30 for non-easy questions with a single relevant article

filing an action for recovery of possession." A related article is "Article 192 A person that commences the possession of movables peacefully and openly by a transactional act acquires the rights that are exercised with respect to the movables immediately if the person possesses it in good faith and without negligence."[3] It is necessary to recognize following semantic relationship ("bicycle" as "movables" and "A" and "B" as persons, and conflict between "by fraud" and "peacefully"). This semantic interpretation of the statue statements is an instance of the greater challenge of identifying relationships between abstract statutes and specific texts.

3

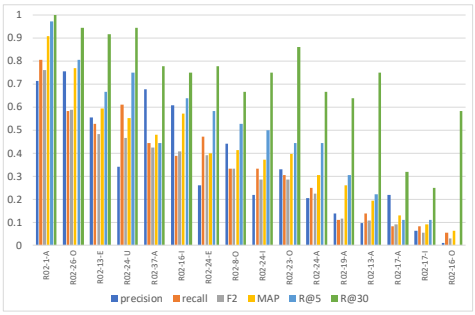

Fig. 3: Averages of precision, recall, F2, MAP, R_5, R_10, and R_30 for non-easy questions with a single relevant article

F2	Anonymize	Other
0-0.1	6	0
0.1-0.2	0	0
0.2-0.3	2	0
0.3-0.4	2	0
0.4-0.5	2	0
0.5-0.6	4	0
0.6-0.7	6	7
0.7-0.8	2	15
0.8-0.9	3	13
0.9-1.0	0	3

Table 9: Number of questions classified by F2 score and query type (Single relevant articles)

4.5 Discussion

Since the statute law retrieval task is one of the oldest tasks of COLIEE, it is appropriate to discuss which kind of issues have been addressed over the development process. As we can see, there are three different types of questions for which we can describe the challenges.

One of the characteristics of difficult questions of this year are those that uses anonymized symbols as pronouns or placeholders, such as “A” and “B” for referring person or other entities. In the test case of COLIEE 2021, 35 questions contain such anonymized symbol and 27 (out of 35) questions have one related article.

Table 9 represents the number of query with one relevant article for the F2 measure (average) classified by one with anonymized symbol or other. Table 10 represents the number of query with multiple relevant article for the F2 measure classified by one with anonymized symbol.

From Table 9, we confirm that most of the retrieval questions without anonymized symbol can be identified by most of the submitted systems (there

F2	Anonymize	Other
0-0.1	3	0
0.1-0.2	1	1
0.2-0.3	1	2
0.3-0.4	1	1
0.4-0.5	1	2
0.5-0.6	1	0
0.7-0.8	0	2

Table 10: Number of questions classified by F2 score and query type (multiple relevant articles)

is no question whose F2 measure (average) is lower than 0.6). However, it is still difficult for the system to retrieve relevant articles for the question with anonymized symbols (16 out of 27 questions has F2 measure (average) lower than 0.6).

This result reflects the different characteristics of the question with anonymized symbol or not. In most of the cases, questions with anonymized symbols represent question about use cases of the articles, therefore they require the handling of semantic relationships that we discussed in Section 4.4. On the contrary, most of the questions without anonymized symbols do not require the handling of such semantic relationships. In addition, since deep learning based NLP such as BERT can handle the context information, it is helpful to select appropriate relevant articles from the ones that use a similar vocabulary. However, the similarity of terms in the legal domain may not be same as ones in the usual texts. For example, "jewelry," "car" and "paintings" are similar terms in the context of valuable movables in the legal domain, but those terms are not similar context in the ordinary texts. Usage of legal-BERT [25] is one of the possible solution for this problem, but their performance is not good as the best run. It is necessary to investigate appropriate model of the transformer (including BERT and other variations) for this task.

For the questions with multiple relevant articles, we still have difficulties to retrieve all relevant articles (Table 10). This is because most of the systems tried to deal this problem as simple rank-based retrieval problems. For example, the best performance system OvGU [21] and the 2nd best team JNLP [14] also use a thresholding approach to select relevant articles. These selection processes can be interpreted as one for deciding of number of relevant documents using rank-based retrieval results.

However, it is better to consider the relationships among statute law articles using article reference information from the legal perspective. HUKB [27] tried to identify the relationships among articles based on the reference information with rank-based retrieval approach. However, their performance is not currently as good as expected.

Based on the discussion, we can confirm that success can use conventional IR methods for retrieving simple questions whose topic are not use cases and have one relevant article. However, we still have difficulty to handle questions about use cases and ones with multiple relevant articles.

For possible future directions, it is necessary to propose a framework to encourage participants to tackle these problems.

5 Tasks 4 and 5 - Statute Law Entailment and Question Answering

5.1 Task Definition

Task 4 is a task to determine textual entailment relationships between a given problem sentence and relevant article sentences. Competitor systems should answer "yes" or "no" regarding the given problem sentences and given article sentences. Until COLIEE 2016, the competition had only 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 and 2020, we returned to the pure textual entailment task to attract more participants, which produced more focused analyses. In COLIEE 2021, we revived the question answering task as Task 5, and retained the textual entailment task as Task 4; Task 5 requires a system to answer "yes" or "no" given a problem sentence(s) only. Participants can use any external data, however this assumes that they do not use the test dataset.

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 (806 questions) and new questions selected from the 2020 bar exam as test data (81 questions). The Task 5 dataset is the same as Task 4. We performed Task 5 before Task 4 in order not to reveal the gold standard article labels which are included in the Task 4 dataset.

5.3 Approaches

All teams submitted three runs for each of Tasks 4 and 5, except that the OvGU and HUKB teams participated Task 4 only.

- **HUKB (three runs)** [26] used an ensemble architecture of BERT methods with data augmentation. They prepared an ensemble of 10 models. Their data augmentation extracts judical decision sentences, then makes positive/negative data from articles.
- **JNLP (three runs)** [15] uses *bert-base-japanese-whole-word-masking* with tf-idf based data augmentation. Their models are trained with different numbers of pretrained/fine-tuned epochs (**JNLP.Enss5a** and **JNLP.Enss5b**), and an ensemble of these two models (**JNLP.EnssBest**). For Task 4, their

proposed methods use their proposed Next Foreign Sentence Prediction (**JNLP. NFSP**) which trains to determine if semantic of two sentences in different languages belong to two consecutive sentences in a document, and Neighbor Multilingual Sentence Prediction (**JNLP. NMSP**) which adds pairs of same-language sentences in two languages to the bilingual pairs of NFSP, together with the original multilingual BERT (**JNLP. BERT_Multilingual**) for Task 5.

- **KIS (three runs)** [6] extended their previous work using a classic NLP approach, to be explainable, based on predicate-argument structure analysis, original legal dictionary, negation detection, and ensemble of modules with different thresholds and combinations of these features.
- **OvGU (three runs)** [25] employed an ensemble of graph neural networks where each node represents either a query or an article, sentences embedded by a pre-trained *paraphrase-distilroberta-base-v1* (**OvGU_run1**), and LEGAL-BERT based on *legalbert- base-uncased* with different training phases (**OvGU_run2** and **OvGU_run3**).
- **TR (three runs)** [21] uses existing models: **TR-Ensemble** using T5 [17]-based ensemble, **TR-MTE** using Multee [24], and **TR_ Electra** using Electra [4] for Task 4; (**TRDistill-Roberta**) using distilled version of RoBERTa [12], **TRGPT3Davinci** using the largest model of GPT-3 [3] and **TRGPT3Ada** using the smaller one for Task 5.
- **UA (three runs)** [9] uses BERT (**UA_dl**), with semantic information (using the Kadokawa thesaurus concept number) (**UA_parser**).

5.4 Results and Discussion

Tables 11 and 13 show evaluation results of Tasks 4 and 5, respectively. Tables 12 and 14 show our categorization results of Tasks 4 and 5, respectively. Because an entailment task is essentially a complex composition of different subtasks, we manually categorized our test data into linguistic categories, depending on what sort of technical issues require resolution. As this is a composite task, overlap is allowed between categories. Our categorization is based on the original Japanese version of the legal bar exam. The **BL** column in Table 12 shows correct answer ratios for each category when answering the majority answer "No" to all problems. Interestingly, all runs are under the baseline in the Negation category, which is expected to answer easier than other categories. This comparison supports the discussion that the task is complex and composite one, the result is not simply regarded as it is better when the overall score is better.

The test dataset characteristics seems not to be coherent throughout these years of the COLIEE series. For example, we observe more problems which require handling of anonymized symbol such as "A" and "B" for referring persons (discussed in the Task 3 part as well) than previous years. Such problems should be still very difficult for any NLP method to solve, except similar possible patterns could be sufficiently covered by some external training dataset.

The **Anaphora** rows of Tables 12 The best team in Task 4 would have solved "easier" problems well, while remaining "difficult" linguistic issues remain for future work.

Team		L	Correct	Accuracy
N/A	BaseLine	N/A	Yes 43/All 81	0.5309
HUKB	HUKB-2	J	57	0.7037
HUKB	HUKB-1	J	55	0.6790
HUKB	HUKB-3	J	55	0.6790
UA	UA_parser	E	54	0.6667
JNLP	JNLP.Enss5Ca	J	51	0.6296
JNLP	JNLP.Enss5Cb	J	51	0.6296
JNLP	JNLP.EnssBest	J	51	0.6296
OVGU	OVGU_run3	E	48	0.5926
TR	TR-Ensemble	J	48	0.5926
TR	TR-MTE	J	48	0.5926
OVGU	OVGU_run2	E	45	0.5556
KIS	KIS1	J	44	0.5432
KIS	KIS3	J	44	0.5432
UA	UA_1st	E	44	0.5432
KIS	KIS2	E	43	0.5309
UA	UA_dl	E	43	0.5309
TR	TR_Electra	J	41	0.5062
OVGU	OVGU_run1	E	36	0.4444

Table 11: Evaluation results of submitted runs (Task 4). sid: submission id, L: Dataset Language (J: Japanese, E: English), Correct: number of correct answers (81 problems in total). JNLP.Enss5Ca and JNLP.Enss5Cb stand for JNLP.Enss5C15050 and JNLP.Enss5C15050SilverE2E10, respectively

6 Conclusion

We have summarized the systems and their performance as submitted to the COLIEE 2021 competition. For Task 1, TLIR was the best performing team with an F1 score of 0.1917, whose approach applied a combination of LMIR and a BERT-based method. In Task 2, the winning team ensembled DeBERTa and monoT5 and achieved an F1 score of 0.6912. For Task 3, the top ranked team (OvGU) employed sentence-BERT embeddings and augmented the training data with metadata, web data related to the articles and relevant queries from the training data, to achieve an F2 score of 0.73. HUKB was the Task 4 winner, with an Accuracy of 0.7037. They applied an ensemble of BERT models and data augmentation. In Task 5, JNLP was the best performing team and applied a variety of BERT-based models, achieving an Accuracy of 0.6049.

In this edition, we introduced a new task on statute law question answering (Task 5) and a new formulation for the case law retrieval task (Task 1). We intend to further improve the datasets quality in future editions of COLIEE so the tasks more accurately represent real-world problems.

Type	#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	BL
Condition	65	60	**63**	60	58	58	58	45	45	45	45	52	54	52	52	49	51	45	58	54
Pers. reltnshp.	44	61	**66**	64	55	57	55	45	43	45	41	52	57	59	55	52	43	50	61	55
Anaphora	37	**68**	**68**	65	65	57	59	62	57	59	41	57	54	51	51	49	54	46	65	57
Pers. role	34	68	**71**	65	53	59	56	44	41	44	44	47	53	56	56	53	44	50	65	59
Pred. argument	32	75	**81**	78	69	69	66	44	41	44	50	50	59	47	59	34	54	47	69	69
Negation	28	64	64	64	61	64	57	50	46	50	54	43	54	**68**	57	57	43	50	**68**	79
Verb paraphrs.	23	61	70	65	61	61	51	48	48	48	35	57	70	70	74	48	65	57	**83**	74
Legal fact	22	68	**73**	64	55	55	55	55	55	55	41	55	64	55	50	45	36	41	59	52
Dependency	22	**73**	**73**	**73**	50	64	59	55	50	50	41	55	59	45	59	32	50	36	59	55
Morpheme	21	**81**	**81**	**81**	76	71	76	**81**	**81**	**81**	48	67	62	71	71	57	71	76	**81**	24
Itemized	11	55	55	55	55	45	55	55	55	55	**73**	64	64	55	45	55	73	55	55	64
Article search	9	**67**	**67**	**67**	44	56	44	**67**	56	56	44	56	**67**	44	56	44	56	33	44	67
Case role	2	0	0	0	**50**	0	0	**50**	0	0	**50**	**50**	**50**	**50**	0	**50**	0	**50**	0	50
Paraphrase	2	50	50	50	**100**	50	**100**	50	50	50	50	50	0	**100**	**100**	0	50	**100**	50	50

Table 12: Task 4's Linguistic category statistics of problems, and correct answers of submitted runs for each category in numbers of counts and percentages.
Type column shows the category names, # column shows the number of problems for each category, alphabetical header names in other columns correspond to formal run names as follows, showing correct answer ratio percentage for each run. A: HUKB-1, B: HUKB-2, C: HUKB-3, D: JNLP.Enss5C15050, E: Enss5C15050SilverE2E10, F: JNLP.EnssBest, G: KIS1, H: KIS2, I: KIS3, J: OVGU_run1, K: OVGU_run2, L: OVGU_run3, M: TR-Ensemble, N: TR-MTE, O: TR_Electra, P: UA_1st, Q: UA_dl, R: UA_parser, BL: Baseline (Answering No to All)

Team	**sid**	**L**	**Correct**	**Accuracy**
N/A	BaseLine	N/A	No 43/All 81	0.5309
JNLP	JNLP.NFSP	J	49	0.6049
UA	UA_parser	E	46	0.5679
JNLP	JNLP.NMSP	J	45	0.5556
UA	UA_dl	E	45	0.5556
TR	TRDistillRoberta	J	44	0.5432
KIS	KIS_2	J	41	0.5062
KIS	KIS_3	J	41	0.5062
UA	UA_elmo	E	40	0.4938
JNLP	JNLP.Task5.BERT	J	38	0.4691
KIS	KIS_1	J	35	0.4321
TR	TRGPT3Ada	J	35	0.4321
TR	TRGPT3Davinci	J	35	0.4321

Table 13: Evaluation results of submitted runs (Task 5). sid: submission id, L: Dataset Language (J: Japanese, E: English), Correct: number of correct answers (81 problems in total). JNLP.Task5.BERT_Multilingual is abbreviated as JNLP.Task5.BERT

7 Conflict of Interest

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

type	#	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
Condition	65	49	52	**57**	54	52	43	37	45	45	51	45	45	48	45	49
Pers. reltnshp.	44	52	48	52	52	43	43	39	48	50	**55**	41	41	50	39	36
Anaphora	37	59	**62**	59	51	49	46	46	59	**62**	51	46	46	51	43	51
Pers. role	34	**62**	56	59	50	44	41	38	47	47	56	38	38	47	44	35
Pred. argument	32	56	**72**	**72**	66	53	50	34	47	44	53	44	44	44	50	56
Negation	28	50	57	**68**	61	46	46	36	39	39	50	29	29	43	50	57
Verb praphrs.	23	57	52	**74**	65	43	48	26	30	35	57	39	39	52	52	65
Legal fact	22	50	59	**64**	**64**	55	55	41	45	45	50	32	32	45	36	41
Dependency	22	55	55	50	**77**	55	55	36	41	41	59	41	41	55	36	55
Morpheme	21	62	57	52	71	76	43	71	71	71	67	48	48	76	71	**81**
Itemized	11	55	45	55	64	45	45	45	73	64	**82**	45	45	73	45	55
Article search	9	56	56	67	56	**67**	56	33	33	33	44	33	33	44	33	33
Case role	2	50	50	50	50	0	0	50	50	50	50	**100**	**100**	**100**	0	0
Paraphrase	2	50	50	50	0	0	0	0	50	50	**100**	**100**	**100**	**100**	50	50

Table 14: Task 5's Linguistic category statistics of problems, and correct answers of submitted runs for each category in numbers of counts and percentages. **Type** column shows the category names, # column shows the number of problems for each category, alphabetical header names in other columns correspond to formal run names as follows, showing correct answer ratio percentage for each run. a: HUKB-1, b: HUKB-2, c: HUKB-3, d: JNLP.NFSP, e: JNLP.NMSP, f: JNLP.Task5.BERT_Multilingual, g: KIS_1, h: KIS_2, i: KIS_3, j: TRDistillRoberta, k: TRGPTAda, l: TRGPT3Davinci, m: UA_dl, n: UA_elmo, o: UA_parser

References

1. Althammer, S., Askari, A., Verberne, S., Hanbury, A.: Dossier@coliee 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval. In: Proceedings of the COLIEE Workshop in ICAIL (2021)
2. Banerjee, P., Han, H.: Language modeling approaches to information retrieval. JCSE **3**, 143–164 (2009). DOI 10.5626/JCSE.2009.3.3.143
3. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020)
4. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. CoRR **abs/2003.10555** (2020). URL http://arxiv.org/abs/1910.10683
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018)
6. Fujita, M., Kiyota, N., Kano, Y.: Predicate's argument resolver and entity abstraction for legal question answering: Kis teams at coliee 2021 shared task. In: Proceedings of the COLIEE Workshop in ICAIL (2021)
7. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced BERT with disentangled attention. CoRR **abs/2006.03654** (2020). URL https://arxiv.org/abs/2006.03654
8. Karpukhin, V., Oguz, B., Min, S., Wu, L., Edunov, S., Chen, D., Yih, W.: Dense passage retrieval for open-domain question answering. CoRR **abs/2004.04906** (2020). URL https://arxiv.org/abs/2004.04906
9. Kim, M.Y., Rabelo, J., Goebel, R.: Bm25 and transformer-based legal information extraction and entailment. In: Proceedings of the COLIEE Workshop in ICAIL (2021)
10. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, pp. 957–966. JMLR.org (2015)
11. Li, J., Zhao, X., Liu, J., Wen, J., Yang, M.: Siat@coliee-2021: Combining statistics recall and semantic ranking for legal case retrieval and entailment. In: Proceedings of the COLIEE Workshop in ICAIL (2021)

12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)
13. Ma, Y., Shao, Y., Liu, B., Liu, Y., Zhang, M., Ma, S.: Retrieving legal cases from a large-scale candidate corpus. In: Proceedings of the 18th International conference on Artificial Intelligence and Law (ICAIL) (2021)
14. Nguyen, H.T., Nguyen, P.M., Vuong, T.H.Y., Bui, Q.M., Nguyen, C.M., Dang, B.T., Tran, V., Nguyen, M.L., Satoh, K.: Jnlp team: Deep learning approaches for legal processing tasks in coliee 2021. In: Proceedings of the COLIEE Workshop in ICAIL (2021)
15. Nguyen, H.T., Tran, V., Minh, N.L., Nguyen, M.P., Vuong, T.H.Y., Bui, M.Q., Nguyen, M.C., Dang, B., Satoh, K.: Paralaw nets - cross-lingual sentence-level pretraining for legal text processing. In: Proceedings of the COLIEE Workshop in ICAIL (2021)
16. 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 (2021). DOI 10.1007/978-3-030-79942-7$_1$3
17. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. CoRR **abs/1910.10683** (2019). URL http://arxiv.org/abs/1910.10683
18. Reimers, N., Gurevych, I.: 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. 3973–3983 (2019)
19. Rosa, G.M., Rodrigues, R.C., de Alencar Lotufo, R., Nogueira, R.: To tune or not to tune? zero-shot models for legal case entailment. In: Proceedings of the 18th International conference on Artificial Intelligence and Law (ICAIL) (2021)
20. Rosa, G.M., Rodrigues, R.C., Lotufo, R., Nogueira, R.: Yes, bm25 is a strong baseline for legal case retrieval. In: Proceedings of the 18th International conference on Artificial Intelligence and Law (ICAIL) (2021)
21. Schilder, F., Chinnappa, D., Madan, K., Harmouche, J., Vold, A., Bretz, H., Hudzina, J.: A pentapus grapples with legal reasoning. In: Proceedings of the COLIEE Workshop in ICAIL (2021)
22. Shao, H.L., Chen, Y.C., Huang, S.C.: BERT-based ensemble model for the statute law retrieval and legal information entailment. In: The Proceedings of the 14th International Workshop on Juris-Informatics (JURISIN2020), pp. 223–234. The Japanese Society of Artificial Intelligence, (2020)
23. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language model-based search engine for complex queries. In: Proceedings of the International Conference on Intelligent Analysis, pp. 2–6 (2005)
24. Trivedi, H., Kwon, H., Khot, T., Sabharwal, A., Balasubramanian, N.: Repurposing entailment for multi-hop question answering tasks. In: Proc. of NAACL (2019)
25. Wehnert, S., Sudhi, V., Dureja, S., Kutty, L., Shahania, S., Luca, E.W.D.: Legal norm retrieval with variations of the bert model combined with tf-idf vectorization. In: Proceedings of the 18th International conference on Artificial Intelligence and Law (ICAIL) (2021)
26. Yoshioka, M., Aoki, Y., Suzuki, Y.: Bert-based ensemble methods with data augmentation for legal textual entailment in coliee statute law task. In: Proceedings of the COLIEE Workshop in ICAIL (2021)
27. Yoshioka, M., Suzuki, Y., Aoki, Y.: Bert-based ensemble methods for information retrieval and legal textual entailment in coliee statute law task. In: Proceedings of the COLIEE Workshop in ICAIL (2021)

Evaluation Measure

In the COLIEE tasks, the following measures are used for evaluation.

- Precision is a measure to analyze accuracy of the returned results using following formula, where N_{TP}, N_{FP} denote the number of true positive and false positive prediction respectively. $N_{TP} + N_{FP}$ equals to the number of all positive cases in the result.

$$Precision = \frac{N_{TP}}{N_{TP} + N_{FP}}, \tag{4}$$

- Recall is a measure to analyze the comprehensiveness of the returned results using following formula, where N_{TP}, N_{FP} denote the number of true positive and false positive prediction respectively. $N_{TP} + N_{TN}$ equals to the number of all true cases in the evaluation set.

$$Recall = \frac{N_{TP}}{N_{TP} + N_{TN}}, \tag{5}$$

- F1 is a measure that consider accuracy and comprehensiveness using harmonic mean of *Precision* and *Recall*.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{6}$$

- F2 is a variation of *F1* that puts more emphasis on *Recall*.

$$F2 = \frac{5 * Precision * Recall}{4 * Precision + Recall}, \tag{7}$$

- Precision@Rank, Recall@Rank are measures used for evaluating ranked list by selecting top *Rank* results as returned results for calculating *Precision* and *Recall* respectively.
- Mean Average Precision (MAP) is a measure to evaluate the quality of ranked retrieval results for document retrieval using following formula , where $Rel(i), N_{rel}, N_{ret}, N_q$ denote functions to check whether ith results is relevant or not, the number of relevant document for the query, returned documents and one of queries respectively.

$$AP = \frac{\sum_{i=0}^{N_{ret}} Precision@i * Rel(i)}{N_{rel}}, MAP = \frac{\sum i = 0^{N_q} AP}{N_q} \tag{8}$$

- Accuracy is calculated the accuracy of the returned results for the task that system returns one answer for each case. This is equivalent to *Precision* and *Recall*, because number of all positive cases from the system and the number of all true cases in the evaluation set are same.

$$Accuracy = \frac{N_{TP}}{N_{TP} + N_{FP}}, \tag{9}$$

We also use *micro-average* and *macro-average* for aggregating the evaluation measures with multiple cases. *Micro average* is calculated measure without considering the cases, but *macro average* is calculated as average of original evaluation measures for each query and calculate average of all cases.