Summary of the Competition on Legal Information, Extraction/Entailment (COLIEE) 2024

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

COLIEE

- 11th edition
- Initial editions had two tasks on Statute Law
- Since 2018, two tasks on Case Law have been added

Task 1 - Description

- Case Law Retrieval
- Competitors are given a set of cases and a mapping containing query cases and their respective noticed cases
- The test dataset does not contain noticed cases, and the task is to predict which cases from the dataset should be noticed with respect to each query case
- Submissions are evaluated by the F1-measure (harmonic mean of precision and recall)

Task 1 - Dataset

- The data is drawn from a collection predominantly comprised of cases from the Supreme Court of Canada
  - Training data: 5,616 cases, of which 1,278 were used as query cases
  - Test data: 1,734 cases, of which 400 were used as query cases
- Direct citations are removed from the case contents
- Training data consists of a list of query cases and their respective noticed cases
- Test data does not include the list of noticed cases for each query case. The challenge in Task 1 is to predict those noticed cases

Task 1 - Training Dataset

Task 1 - Training Dataset

Task 1 - Training Dataset

Task 1 - Training Dataset

Task 1 - Training Dataset

Task 1 - Participation

- Ten teams (total 26 submissions)
  - Each team could send at most three submissions
  - Eight teams sent papers describing their methods

Task 1 - Official Results

Team	File	F1	Precision	Recall
TQM	task1_test_answer_2024	0.4432	0.5057	0.3944
TQM	task1_test_answer_2024	0.4342	0.5082	0.3790
UMNLP	task1_umnlp_run1.txt	0.4134	0.4000	0.4277
UMNLP	task1_umnlp_run2.txt	0.4097	0.3755	0.4507
UMNLP	task1_umnlp_runs_comb	0.4046	0.3597	0.4622
YR	task1_yr_run1.json	0.3605	0.3210	0.4110
TQM	task1_test_answer_2024	0.3548	0.4196	0.3073
YR	task1_yr_run2.json	0.3483	0.3245	0.3758
YR	task1_yr_run3.json	0.3417	0.3184	0.3688
JNLP	64b7b-07f39.txt	0.3246	0.3110	0.3393
JNLP	07f39.txt	0.3222	0.3347	0.3105
JNLP	64b7b-48fe5.txt	0.3103	0.3017	0.3195
WJY	submit_1.txt	0.3032	0.2700	0.3457

Team	File	F1	Precision	Recall
BM24	task1_test_result.txt	0.1878	0.1495	0.2522
CAPTAIN	captain_mstr.txt	0.1688	0.1793	0.1594
CAPTAIN	captain_ft5.txt	0.1574	0.1586	0.1562
NOWJ	nowjtask1run2.txt	0.1313	0.0895	0.2465
NOWJ	nowjtask1run3.txt	0.1306	0.0957	0.2055
NOWJ	nowjtask1run1.txt	0.1224	0.0813	0.2478
WJY	submit_3.txt	0.1179	0.0870	0.1831
WJY	submit_2.txt	0.1174	0.0824	0.2042
MIG	test1_ans.json	0.0508	0.0516	0.0499
UBCS	run3.txt	0.0276	0.0140	0.7196
UBCS	run2.txt	0.0275	0.0140	0.7177
UBCS	run1.txt	0.0272	0.0139	0.7100
CAPTAIN	captain_bm25.txt	0.0019	0.0019	0.0019

Task 1 - Methods

Team	Description
TQM (winning team)	Used lexical matching and dense vector retrieval to generate features (plus some simple features such as case length) that were submitted to a learning to rank method. The authors also applied pre and post processing to avoid irrelevant information. Their method not only applies all of those techniques, but aims at a deeper understanding of the case trying to capture the main facts described in the case.
JNLP	Proposed a three-phase approach: the first stage performs retrieval after splitting the query document into paragraphs and using a BM25 model with top-k cutout to retrieve candidate documents. Phase two is a re-ranking stage. The last stage is where prediction actually happens: after the re-ranking stage, for each query document, the authors select the top-k candidate documents from the re-ranked list as prediction with k selected, using grid-search on the validation set. They also developed an ensemble strategy by concatenating the prediction results of the re-rankers before selecting the top-k to boost the recall metric of the system.
BM24	Organize each case into segments summarized by gpt-3.5. Among them, one segment is selected to represent the case. An embedding of that segment is stored in FAISS. A segment of the query case is used to query the FAISS vector store to retrieve similar cases. AnglE is used as the sentence embedding model, trained from an open source model on the Task 1 training set pre-processed in the same way.
CAPTAIN	Performed some heuristic pre-processing steps, then used TF-IDF and BM25 to extract keywords and retrieve relevant documents, and then applied LLMs to summarize the decisions and perform fine-tuning of a retrieval model based on such summaries

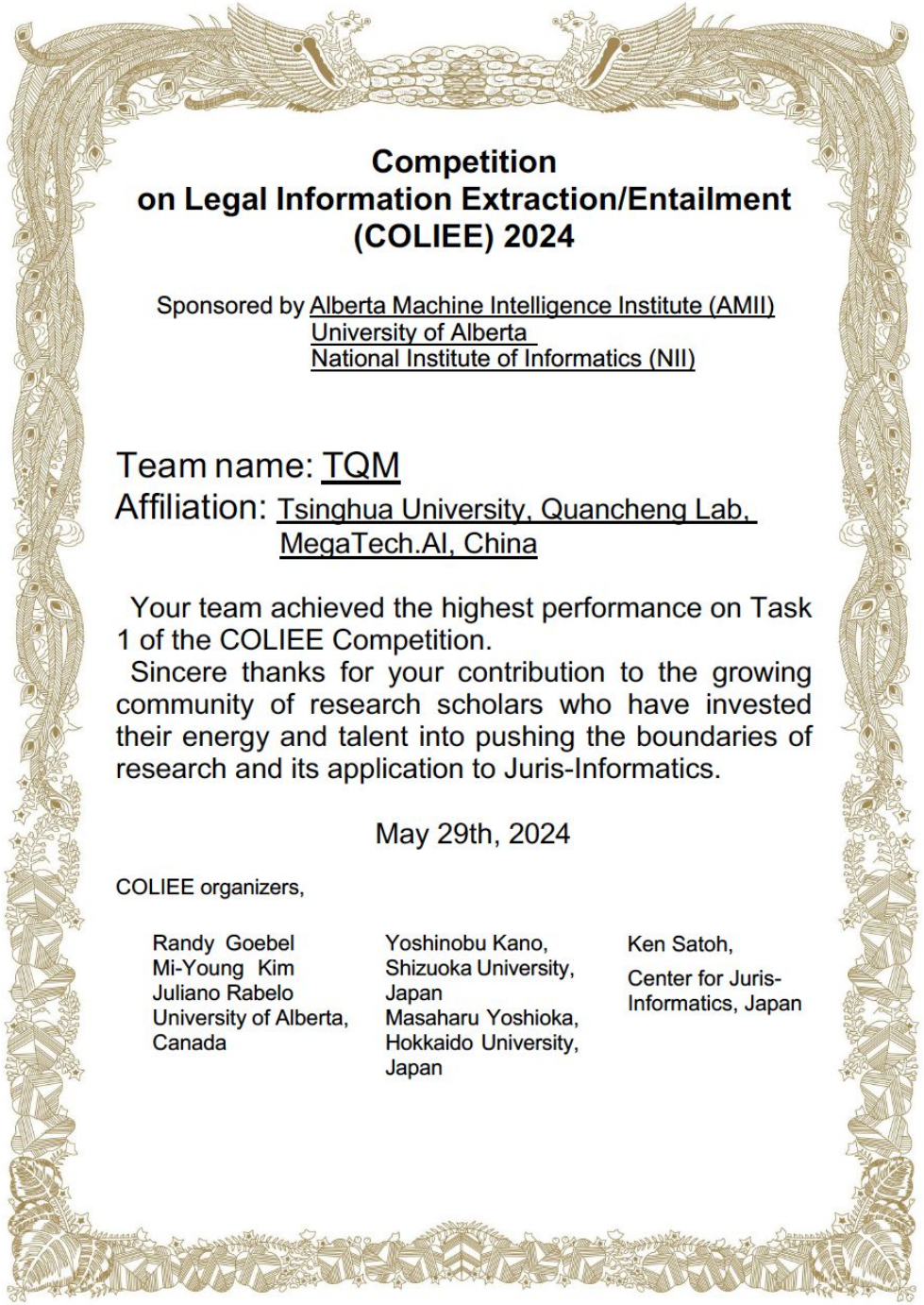
Task 1 - Methods

Team	Description
NOWJ	Developed an approach based on a combination of BM25 and a pre-trained Longformer. After an initial pre-processing step, BM25 is used to calculate the similarity between each pair of query case and candidate case. The result is used as a pre-ranking input to the Long-Former model. Scores from BM25 and LongFormer are then combined, with parameters being defined after a grid search is conducted.
MIG	Developed an informative baseline for Task 1 that does not apply any LLMs. The authors vectorized the cases with a tool on BERT-base and BERT-large. After vectorizing the cases, they compute the cosine similarity between the candidate cases and the given new case using FAISS. Then, for a new case, the authors ranked the candidate cases by their cosine similarity with the new case, and chose 20 candidates that were most similar to the new case. Then the difference between the cosine similarity between the i-th most similar case and the $(i + 1)$-th most similar case (d_i) is calculated, and the first i cases are recommended if $d_i > 2d_1$
UBCS	Applied TF-IDF to rank cases varying how the model is used. Their first approach is a baseline, with vanilla TF-IDF weighting model being used to retrieve and rank noticed cases for each given query case. The second approach applies summarization only on the query cases before using TF-IDF for retrieval. The third approach applies summarization for both the query and candidate cases.
UMNLP	Developed a pairwise similarity ranking framework. The authors train a feed-forward neural network to perform a binary classification task based on several features from each query-candidate case pair. Those features include the extraction and similarity matching for a novel feature which the authors call a “proposition” (a short summary of the basis upon which a noticed case has been cited), as well as the name of the judge deciding the case, verbatim quotations from the text, and several other novel features

Task 1 - Discussion

- Scores improved steadily since the new format for this task was introduced. Compared to last year, almost a 50% increase (0.30 to 0.44)
- Most competitors applied some transformer-based approach (at least as a component of their method)
- Usually transformer-based methods are combined with traditional IR methods such as BM25, and augmented with heuristics

Task 1 - Winner Certificate



Competition
on Legal Information Extraction/Entailment
(COLIEE) 2024

Sponsored by Alberta Machine Intelligence Institute (AMII)
University of Alberta
National Institute of Informatics (NII)

Team name: TQM
Affiliation: Tsinghua University, Quancheng Lab, MegaTech.AI, China

Your team achieved the highest performance on Task 1 of the COLIEE Competition.
Sincere thanks for your contribution to the growing community of research scholars who have invested their energy and talent into pushing the boundaries of research and its application to Juris-Informatics.

May 29th, 2024

COLIEE organizers,

Randy Goebel
Mi-Young Kim
Juliano Rabelo
University of Alberta, Canada

Yoshinobu Kano,
Shizuoka University, Japan
Masaharu Yoshioka,
Hokkaido University, Japan

Ken Satoh,
Center for Juris-Informatics, Japan

Task 1 - Summary

- Legal Case Retrieval Task
- Requires identification of noticed cases with respect to a query case
- The answer cannot rely on simple pattern matching as the explicit case citations are removed from the case contents
- Ten teams and 26 submissions
- Approaches predominantly relied on some combination of transformer-based and traditional IR methods

Summary of the Competition on Legal Information, Extraction/Entailment (COLIEE) 2024

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