Overview of Benchmark Datasets and Methods for the Legal Information Extraction/Entailment Competition (COLIEE) 2024

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

Task 2

Mi-Young Kim, University of Alberta

Task 2 Description

- Legal Case Entailment Task
- Involves the identification of a paragraph from existing cases that entails the decision of a new case
- Given a decision Q of a new case and a relevant case R, a specific paragraph in R that entails the decision Q needs to be identified.
- The answer paragraph cannot be identified merely by information retrieval techniques
  - Because the case R is a relevant case to Q, many paragraphs in R can be relevant to Q regardless of entailment.

Task 2 Example

Base case	B232 arrived in Canada with 491 other persons aboard the MV Sun Sea...
Decision of the base case	Given that the Respondent remains a security risk whom the Minister has...
Noticed (relevant previous) case	[P1] Previous decisions to detain the individual must be...
	[P2] The Ministers are requesting an order...
	…
	[P39] THIS COURT ORDERS that the stay motion be granted until the final ...
Entailing paragraph to the decision of the base case	P27

Evaluation Measure and Dataset

- **F1-measure : harmonic mean of precision and recall**

- **The training data : 725 query cases and 25,783 paragraphs were provided for training.**
- **The test data: 100 query cases and 3,651 paragraphs**
- **The data is drawn from an existing collection of predominantly Federal Court of Canada case law.**
- **Training data consists of triples 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.**
- **Test data does not include the paragraph number of the noticed case. The goal of Task 2 is to identify this paragraph number.**

Participation in Task 2

- **Six teams (total 18 submissions)**
  - **3 submissions per team**

    (Each team was allowed maximum three submissions.)

Submitted methods

Team	Approaches
AMHR [8] (three runs)	proposed two approaches: (1) finetuning a legal-BERT model with triplet loss with labels as positive examples and all other paragraphs as negative examples on the train set provided for task 2. This approach resulted in overfitting. (2) finetuning a monoT5 model pre-trained on the MSMARCO dataset with hard negative mining examples chosen by BM25 and another version of the monoT5 model itself. They choose the top-2 predictions by this model as long as the ratio between their similarity score is less than 6.619 (a hyperparameter found by grid search); otherwise, they choose just the first prediction. The second approach got the best results on task 2, this year.
CAPTAIN [6] (three runs)	introduces a method that builds upon the state-of-the-art approach used in Task 2 of the 2023 competition. This method incorporates zero-shot and few-shot learning techniques to leverage the knowledge stored in large language models. Initially, they fine-tune a pre-trained monoT5 sequence-to-sequence model using hard negative sampling to produce an output. For each query paragraph, they select the top-k candidates with the highest scores to create zero-shot and few-shot prompting techniques for in-context learning with FlanT5 LLM.

[5]	Nguyen, C., Tran, T., Le, K., Nguyen, H., Do, T., Pham, T., Luu, S.T., Vo, T., , Nguyen, L.M.: Pushing the Boundaries of Legal Information Processing with Integration of Large Language Models. In: Proceedings of the Eighteenth International Workshop on Juris-informatics(JURISIN 2024) (2024)
[6]	Nguyen, P., Nguyen, C., Nguyen, H., Nguyen, M., Trieu, A., Nguyen, D., Nguyen, M.: CAPTAIN at COLIEE 2024: Large Language Model for Legal Text Retrieval and Entailment. In: Proceedings of the Eighteenth International Workshop on Juris-informatics(JURISIN 2024) (2024)
[7]	Nguyen, T.M., Nguyen, H.L., Nguyen, D.Q., Nguyen, H.T., Vuong, T.H.Y., Nguyen, H.T.: NOWJ@COLIEE 2024: Leveraging Advanced Deep Learning Techniques for Efficient and Effective Legal Information Processing. In: Proceedings of the Eighteenth International Workshop on Juris-informatics(JURISIN 2024) (2024)
[8]	Licato, J.: AMHR COLIEE 2024 Entry: Legal Entailment and Retrieval. In: Proceedings of the Eighteenth International Workshop on Juris-informatics(JURISIN 2024) (2024)
[11]	Wehnert, S., Murugadas, V., Naik, P.V., Luca, E.W.D.: Improving Robustness in Language Models for Legal Textual Entailment through Artifact-Aware Training. In: Proceedings of the Eighteenth International Workshop on Juris-informatics(JURISIN 2024) (2024)

Submitted methods

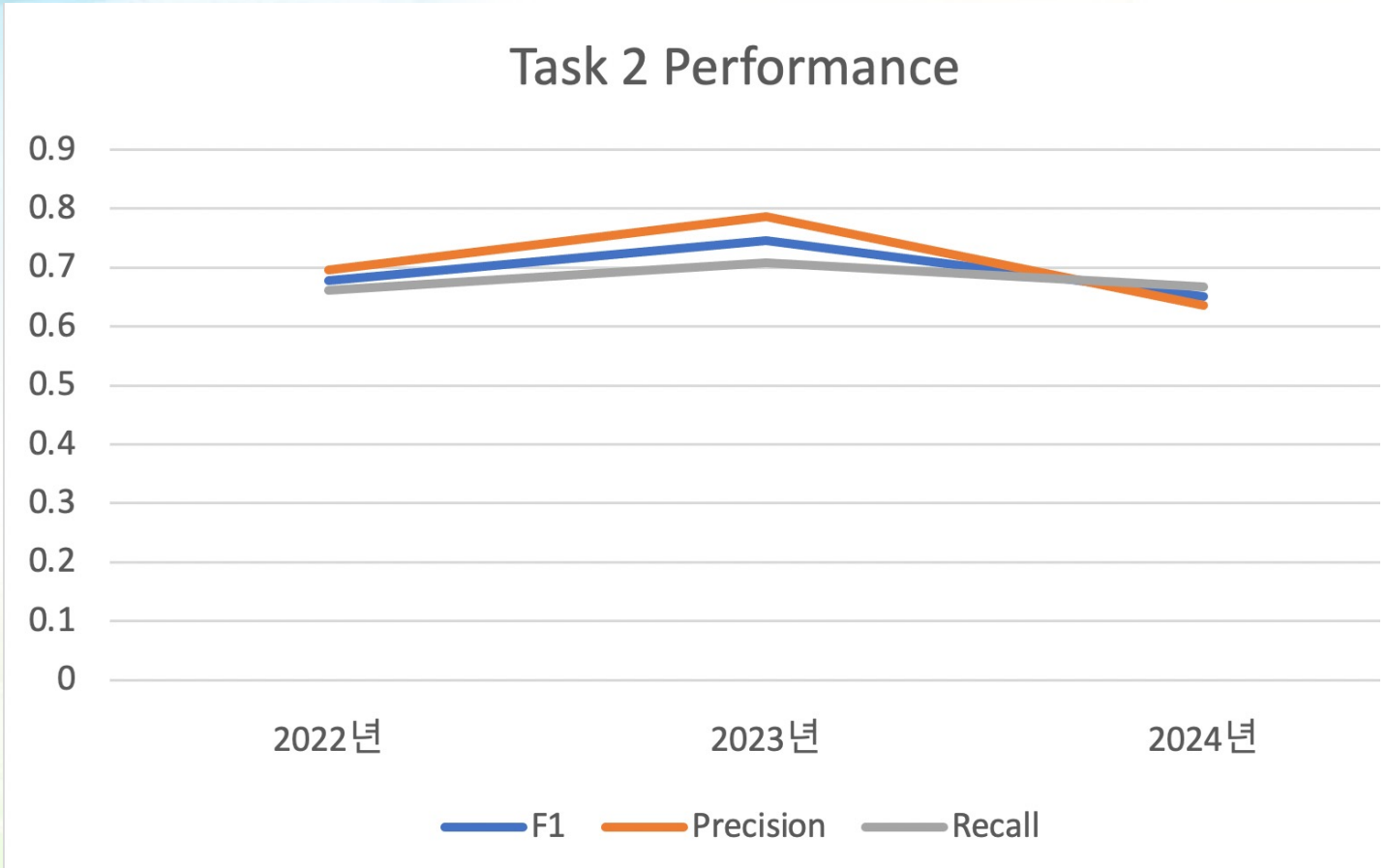
Team	Approaches
JNLP [5] (three runs)	fine-tuned MonoT5 on the training set of Task 2 with hard negative sampling. The model MonoT5 is a T5-3B reranker finetuned on the MS MARCO passage dataset for 10k steps. They used Flan-T5 and Mixtral for prompting.
NOWJ [7] (three runs)	proposes two approaches of entailment recognition, using multilingual BERT and monoT5 for the three runs. MonoT5 is a T5-based re-ranking model fine-tuned for the downstream task of classification, while mBERT is a traditional approach for document re-ranker. Multilingual BERT and training the mBERT model with weak labels [10] were their last year's solutions. Therefore, for the first two runs, they finetuned the models on this year's dataset.

Submitted methods

Team	Approaches
OVGU [11] (three runs)	team's proposed approach involves using a chain of pre-trained Custom Legal-BERT models that are fine-tuned on sub-datasets generated using BM25 and a Bi-Encoder to select the top-N candidate paragraphs. To enhance the models' robustness, a binomial test is employed for artifact detection. OpenAI's GPT-3.5-turbo model is used to create adversarial instances for selected training instances with annotation artifacts. The large language model was prompted to switch the previous negative entailment label into a positive one for balancing out the training examples with annotation artifacts. These instances, along with the top-N candidate paragraph dataset, are further used to fine-tune the models. A chained approach is applied during prediction: If the first model (specialized for high precision) fails to predict a hypothesis with at least one premise as 'Entailed,' the second model is used for that hypothesis. If any hypotheses are missed after using the second model, the BM25 top-ranked premise found for a given hypothesis is labeled as 'Entailed.'

Task 2 official results

Team	F1	Precision	Recall	Team	F1	Precision	Recall
AMHR	0.6512	0.6364	0.6667	CAPTAIN	0.6360	0.7281	0.5646
JNLP	0.6320	0.6967	0.5782	CAPTAIN	0.6235	0.7700	0.5238
CAPTAIN	0.6235	0.7700	0.5238	NOWJ	0.6117	0.6181	0.6054
JNLP	0.6045	0.6694	0.5510	OVGU	0.5962	0.5636	0.6327
NOWJ	0.5946	0.5906	0.5986	JNLP	0.5912	0.6378	0.5510
OVGU	0.5705	0.5506	0.5918	OVGU	0.5532	0.5000	0.6190
NOWJ	0.5197	0.5032	0.5374	MIG	0.4701	0.5673	0.4014
MIG	0.4696	0.5800	0.3946	AMHR	0.3542	0.3617	0.3469
AMHR	0.3320	0.4100	0.2789	MIG	0.1364	0.0979	0.2245


Task 2 Performance
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2022년
2023년
2024년
F1
Precision
Recall

Number of the answer paragraphs

<2024>

Number of the answer paragraphs	1	2	3	4
Number of queries	65	25	8	2

<2023>

Number of the answer paragraphs	1	2	3	4
Number of queries	86	9	4	1

Discussion

- The AMHR team attained the best results.
- CAPTAIN used last year's winner model, which is based on a fine-tuned monoT5, and their model was ranked second.
- The first ranked model also used fine-tuned monoT5, but they used a hyperparameter value as a threshold of the similarity score, and got the best result this year.

Certificate for Task 2 winner

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: AMHR
Affiliation: University of South Florida, USA

Your team achieved the highest performance on Task 2 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,
National Institute of Informatics (NII), Japan