


Overview of COLIEE 2024 Japanese Statute Law Tasks

Task 3: The Statute Law Retrieval Task



Masaharu Yoshioka
Hokkaido University



Task settings

■ Target documents

- Japanese Civil law articles are provided in two languages (Japanese and English: 768 articles in total)
Data is updated because of the update of Japanese Civil Code.

■ Questions and gold standard data

- Questions are selected from Japanese Bar exam in two languages (Japanese and English)
 - Training 1,097 questions (extracted from exam conducted in 2006-2022 exam)
 - Test: 109 questions (extracted from exam conducted in 2023 exam)*
- Gold standard data
 - Relevant articles and entailment results (true/false) are judged by the person in legal domain



Regulation for using External Resources

- Concern about using GPT-4 at COLIEE 2023
 - Reproducibility problem
 - Contamination problem
- New regulation
 - Reproducibility
External resources used for the submission should be available for other participants.
 - Contamination
If the external resource (such as LLM) is machine learning based one, training data for the resource should be disclosed.

Evaluation Format

■ Evaluation measure

- Usage of macro average (calculate average over all evaluation measures for each question)
- Usage of F2 (put more emphasis on recall)
It is important to provide all related articles for the entailment process.

$$\text{F2-measure} = \frac{5 \times \text{Precision} \times \text{Recall}}{4 \times \text{Precision} + \text{Recall}}$$

■ Submission of long ranked list (top 100)

- We can analyze type of difficult question

Submitted Runs

- 20 runs from 8 teams
 - LLM prompt **1st**
 - AMHR(3 runs), CAPTAIN(3 runs), JNLP(3 runs)
 - Deep Neural LM (e.g., BERT, T5) or sentence embeddings **3rd** **2nd**
 - BM24 (1 run) NOWJ(3 runs), TQM (3 runs), UA (3 runs)
 - No information
 - PSI

JNLP and BM24 uses LLMs that are trained using undisclosed data

Evaluation results based

- Selected list for the best run of each team

Team	return	retrieved	F2	Precision	Recall	MAP
JNLP *	188	<u>99</u>	<u>0.807</u>	0.709	<u>0.870</u>	0.801
<u>CAPTAIN</u>	168	96	0.800	0.732	0.845	<u>0.815</u>
TQM	140	89	0.782	<u>0.785</u>	0.800	0.790
NOWJ	202	96	0.772	0.690	0.835	0.756
AMHR	185	95	0.749	0.651	0.825	0.740
UA	233	91	0.711	0.610	0.800	0.700
BM24 *	425	94	0.539	0.282	0.795	-
PSI	109	9	0.086	0.090	0.085	0.2

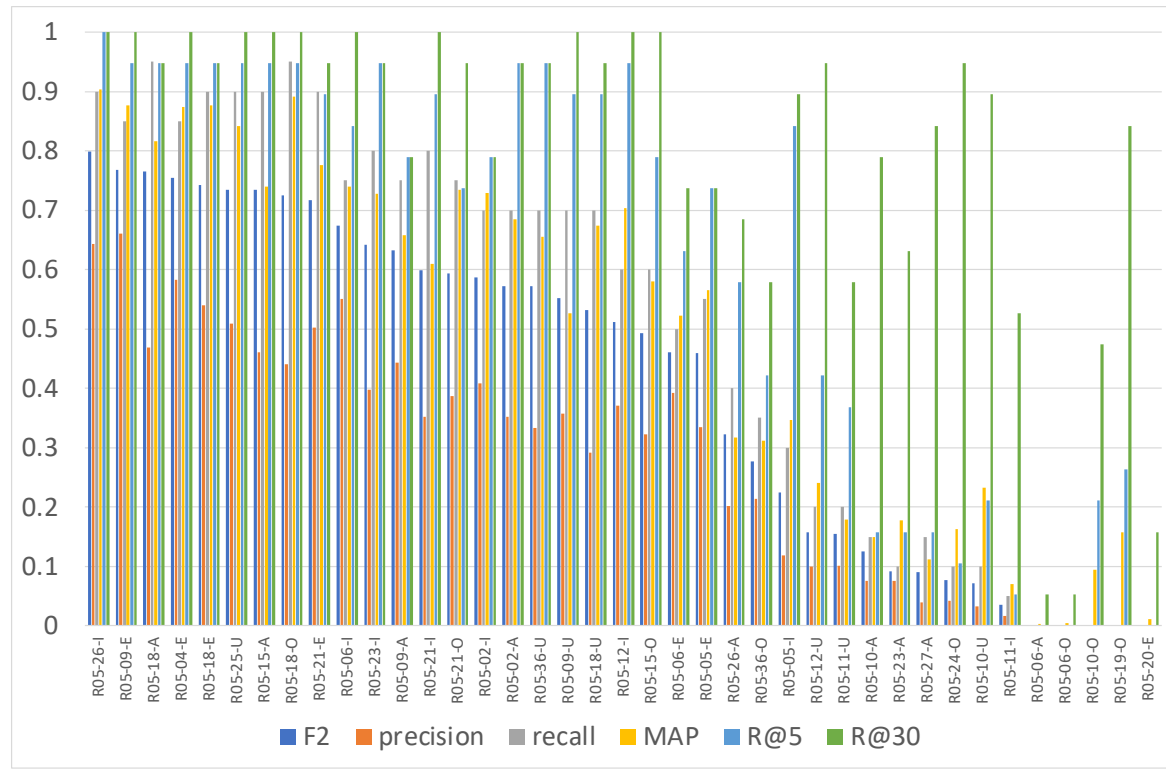
JNLP and BM24 uses LLMs that are trained using undisclosed data

Average Retrieval Results for Single Answer

■ 88 questions

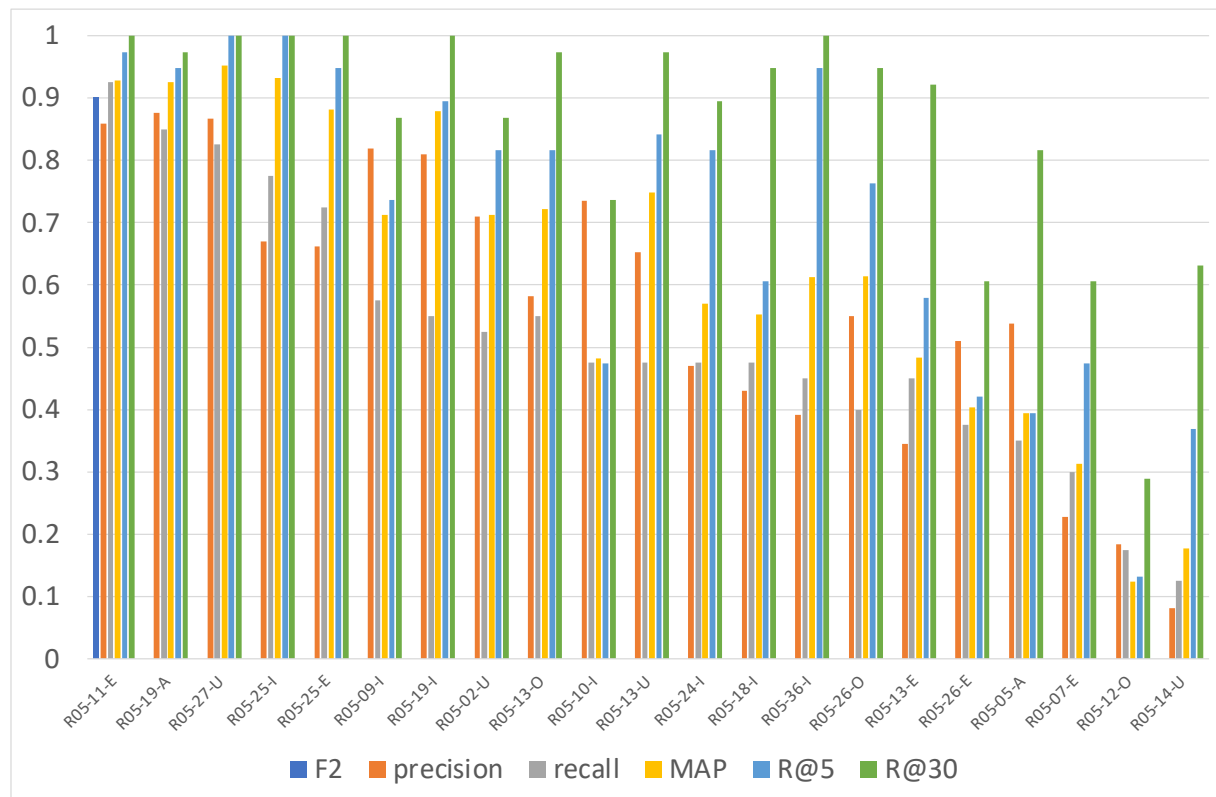
- Easy(Average of F2 is larger than 0.8): 49
- Difficult(Average of F2 is lower than 0.2): 13

Average performance for the difficult questions
(Average F2 is lower than 0.8)



Average Retrieval Results that have 2 Relevant Articles

- 21 questions: Recall is comparatively lower than ones for single question
 - However, performance is comparatively better than COLIEE 2023



Difficult questions

- 15 questions (average F2 is lower than 0.2)
 - 9 questions out of 15 use anonymized symbol (such as A and B)
- Question R05-6-A

If A concludes a **gift agreement** with B to the effect that A will give B 100,000 yen **if B does not commit theft for a certain period of time**, and if B, who has not **committed** theft for that period of time, demands A to pay 100,000 yen, A may refuse to do so.

- Relevant article

(Unlawful Conditions)

Article 132 A **juridical act** subject to **an unlawful condition** is void. The same applies to a **juridical act** subject to the condition that **an unlawful act** not be **performed**.

Performance for the anonymized question (45 questions)

- AMHR, JNLP, CAPTAIN that use LLM prompt have better performance than others (TQM is the best among others)
 - LLM prompt may be good to handle semantic matching.

Team	return	retrieved	F2	Precision	Recall	MAP
AMHR	87	<u>42</u>	<u>0.669</u>	0.561	<u>0.756</u>	0.726
JNLP	83	39	0.662	0.586	0.722	0.735
CAPTAIN	95	<u>42</u>	0.647	0.497	<u>0.756</u>	<u>0.742</u>
TQM	56	33	0.628	<u>0.678</u>	0.633	0.719



Summary

- There are many questions that can be retrieved by the all systems.
 - We still have problems to retrieve questions with anonymized symbols.
 - However, LLM prompt has slightly better performance than previous methods.