Task 4: Textual Entailment in Statute Law

COLIEE 2024 Overview @ JURISIN 2024

狩　野　　　芳　伸
Yoshinobu Kano
静岡大学　情報学部
行　動　情　報　学　科
Kano Laboratory
国立大学法人 静岡大学
National University Corporation Shizuoka University

COLIEE Competition

- Competition for Legal Information Extraction and Entailment (COLIEE)
  - I am one of the organizers
  - COLIEE 2013, 2014, 2015, 2016, 2018, 2020, 2022, 2024 in JURISIN
  - COLIEE 2017, 2019, 2021, 2023 in ICAIL
- Case Law tasks (from 2018)
  - Canadian Federal Court database
- Statute Law tasks
  - Japanese Legal bar exam
  - Human applicants should pass the Legal Bar Exam to be a lawyer in Japan

Example of Statute Task

Question	A special provision that releases warranty can be made, but in that situation, when there are rights that the seller establishes on his/her own for a third party, the seller is not released of warranty.
Related Article (Task IR)	(Special Agreement Disclaiming Warranty) **Article 572** Even if the seller makes a special agreement to the effect that the seller will not provide the warranties set forth from Article 560 through to the preceding Article, the seller may not be released from that responsibility with respect to any fact that the seller knew but did not disclose, and with respect to any right that the seller himself/herself created for or assigned to a third party.
Label (Task TE/QA)	**Yes**

COLIEE 2024 Task 4 Dataset

- Dataset same as Task 3
- COLIEE 2024 training data
  - 1097 queries
  - built from the bar exam (short answer test) civil code part
  - published in 2006-2023
  - XML files, each corresponds to one year’s publication
- Japanese Civil Law Articles as knowledge base
- Both in original Japanese version and manually translated English version
- COLIEE 2024 test data
  - 109 queries from the latest bar exam of 2023
- Each team can submit up to three runs for each task
  - We asked to submit past formal run configurations as well
    - 2021 (R02), 2020 (R01), 2019 (H30)

Overview: Historical Development

- Linguistic structures specific to legal docs
  - ～COLIEE 2019: classic NLP
- Insufficient data size (pretrain/finetune)
  - ～COLIEE 2020: deep language model by transfer learning (pretraining)
  - ～COLIEE 2022: ensemble of different system outputs
  - COLIEE 2023～: LLM, generative AI
- General knowledge, evidence/explanations
  - ???: common sense, relationships, logic, etc.
  - COLIEE 2025 – new task/evaluation planned!

Clauses and Predicate Arguments

H24-2-1

Proposition Clause

Conditional Clause

制限行為能力者のした契約について，制限行為能力者及びその法定代理人が取消権を有するときは，契約の相手方も取消権を有する。

An act which may be rescinded on the grounds of the limited capacity to act of the person who performed such act may be rescinded only by the person whose capacity to act is limited, or its agent, successor, or a person who has the authority to give consent.

Subject

Predicate

契約の相手方も取消権を有する。

Object

Subject

Predicate

制限行為能力者及びその法定代理人が取消権を有するときは，

Object

PAS in proposition clause

{有する, 相手方, 取消権}

(has, the counterparty, the right to rescind)

PAS in conditional clause:

{有する, 法定代理人, 取消権}

(have, the statutory agent, the right to rescind)

Overview: Historical Development

- Linguistic structures specific to legal docs
  - ～COLIEE 2019: classic NLP
- Insufficient data size (pretrain/finetune)
  - ～COLIEE 2020: deep language model by transfer learning (pretraining)
  - ～COLIEE 2022: ensemble of different system outputs
  - COLIEE 2023～: LLM, generative AI
- General knowledge, evidence/explanations
  - ???: common sense, relationships, logic, etc.
  - COLIEE 2025 – new task/evaluation planned!

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: CAPTAIN
Affiliation: Japan Advanced Institute of Science and Technology, Japan

Your team achieved the highest performance on Task 4 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

Results

- Total 8 teams (24 runs)
- CAPTAIN2 winner!
  - LLM (flan-T5) with data augmentation and heuristic rules, fine-tune.
- *JNLP 2nd
  - different LLMs (Wqen (their original model), Mistral, Flan-Alpaca, and FlanT5) ensemble the results with majority voting, took the top-1 prompt from Flan-Alpaca

Team	Formal Run		Past Formal Runs		
	# Correct	R05	R02	R01	H30
BaseLine (Yes to all)	60	0.5505	0.5309	0.5315	0.5143
# Correct / # Total		60/109	43/81	59/111	36/70
CAPTAIN2	90	0.8257	0.7901	0.7568	0.8429
JNLP1 *	89	0.8165	0.7901	0.6937	0.7429
UA_slack	87	0.7982	0.7407	0.7117	0.7429
UA_encoder_decoder	87	0.7982	0.8395	0.7207	0.7571
CAPTAIN1	86	0.7890	0.8148	0.7748	0.8286
CAPTAIN3	86	0.7890	0.8395	0.7207	0.7286
JNLP2 *	86	0.7890	0.8272	0.7297	0.7857
UA_gpt	85	0.7798	0.7901	0.6847	0.7571
AMHR.ensembleA50	84	0.7706	0.8148	0.3784	0.6571
AMHR.single	84	0.7706	0.7901	0.3874	0.6714
HI1	82	0.7523	0.7284	0.6667	0.7000
NOWJ.pandap46 *	82	0.7523	N/A	N/A	N/A
AMHR.ensembleA0	80	0.7339	0.7778	0.4234	0.7000
JNLP3 *	80	0.7339	0.7901	0.6126	0.6571
NOWJ.flant5-panda *	80	0.7339	N/A	N/A	N/A
NOWJ.bagging *	78	0.7156	N/A	N/A	N/A
OVGU1 +	77	0.7064	0.7531	0.6937	0.6714
KIS2 +	76	0.6972	0.6543	0.6036	0.6429
OVGU3 +	76	0.6972	0.7654	0.6306	0.7000
OVGU2 +	70	0.6422	0.6790	0.6396	0.6000
KIS1	67	0.6147	0.6420	0.6847	0.6286
HI3	64	0.5872	0.6296	0.6306	0.6000
HI2	63	0.5780	0.7531	0.6937	0.7143
KIS3	62	0.5688	0.5926	0.6306	0.6429

Results

- UA_stack 3rd
  - used zero-shot learning on google/flant5-xxl with PromptSource8 fo finding potential good prompts, added positive and one negative example, chose the top 3 prompts, finally zero-shot inference with all three prompts and voting between them.
- ∗ indicates runs using not fully disclosed models
- + indicates runs with preprocessing by such models

Team	Formal Run		Past Formal Runs		
	# Correct	R05	R02	R01	H30
BaseLine (Yes to all)	60	0.5505	0.5309	0.5315	0.5143
# Correct / # Total		60/109	43/81	59/111	36/70
CAPTAIN2	90	0.8257	0.7901	0.7568	0.8429
JNLP1 ∗	89	0.8165	0.7901	0.6937	0.7429
UA_slack	87	0.7982	0.7407	0.7117	0.7429
UA_encoder_decoder	87	0.7982	0.8395	0.7207	0.7571
CAPTAIN1	86	0.7890	0.8148	0.7748	0.8286
CAPTAIN3	86	0.7890	0.8395	0.7207	0.7286
JNLP2 ∗	86	0.7890	0.8272	0.7297	0.7857
UA_gpt	85	0.7798	0.7901	0.6847	0.7571
AMHR.ensembleA50	84	0.7706	0.8148	0.3784	0.6571
AMHR.single	84	0.7706	0.7901	0.3874	0.6714
HI1	82	0.7523	0.7284	0.6667	0.7000
NOWJ.pandap46 ∗	82	0.7523	N/A	N/A	N/A
AMHR.ensembleA0	80	0.7339	0.7778	0.4234	0.7000
JNLP3 ∗	80	0.7339	0.7901	0.6126	0.6571
NOWJ.flant5-panda ∗	80	0.7339	N/A	N/A	N/A
NOWJ.bagging ∗	78	0.7156	N/A	N/A	N/A
OVGU1 +	77	0.7064	0.7531	0.6937	0.6714
KIS2 +	76	0.6972	0.6543	0.6036	0.6429
OVGU3 +	76	0.6972	0.7654	0.6306	0.7000
OVGU2 +	70	0.6422	0.6790	0.6396	0.6000
KIS1	67	0.6147	0.6420	0.6847	0.6286
HI3	64	0.5872	0.6296	0.6306	0.6000
HI2	63	0.5780	0.7531	0.6937	0.7143
KIS3	62	0.5688	0.5926	0.6306	0.6429

Task 4 Results (Textual En

- Comparison with previous formal run settings (training/eval)
  - 2021 (R02), 2020 (R01), 2019 (H30)
  - asked to apply with this year's same system
- Different year shows quite different charcteristics due to the datasets
  - any way to get more stable results?

Run	Past Formal Runs		
R05	R02	R01	H30
0.5505	0.5309	0.5315	0.5143
60/109	43/81	59/111	36/70
0.8257	0.7901	0.7568	0.8429
0.8165	0.7901	0.6937	0.7429
0.7982	0.7407	0.7117	0.7429
0.7982	0.8395	0.7207	0.7571
0.7890	0.8148	0.7748	0.8286
0.7890	0.8395	0.7207	0.7286
0.7890	0.8272	0.7297	0.7857
0.7798	0.7901	0.6847	0.7571
0.7706	0.8148	0.3784	0.6571
0.7706	0.7901	0.3874	0.6714
0.7523	0.7284	0.6667	0.7000
0.7523	N/A	N/A	N/A
0.7339	0.7778	0.4234	0.7000
0.7339	0.7901	0.6126	0.6571
0.7339	N/A	N/A	N/A
0.7156	N/A	N/A	N/A
0.7064	0.7531	0.6937	0.6714
0.6972	0.6543	0.6036	0.6429
0.6972	0.7654	0.6306	0.7000
0.6422	0.6790	0.6396	0.6000
0.6147	0.6420	0.6847	0.6286
0.5872	0.6296	0.6306	0.6000
0.5780	0.7531	0.6937	0.7143
0.5688	0.5926	0.6306	0.6429

Open Questions and Future Plans: What are the LLMs doing?

- LLMs could answer quite accurately
  - evidences are also fine in most cases
  - LLMs (implicitly) includes answers similar to our gold data in their training?
  - Humans can perform “symbolic processings” “logical calculations”
  - but LLMs should not perform “logical calculation” rather “compositions”
  - Needs precise analysis regarding what sort of issues are actually solved
- Can LLMs “logically” think?
- “Explainable AI” required in two meanings
  - explanation for humans
  - explanation of the internal process

Difficult Examples

- Temporal expressions, coreferences
  - *If person A donates a house that he/she is renting to person C with a provision for the payment of rent at the end of every month to person B midway through the month, if there is a special agreement between person A and person B, the rent for the month will be distributed between person A and person B in proportion to the number of the days.*
- Document structure, references, Negation (sometimes implicit in terms), Acronyms (domain dependent, vague), …

Overview: Historical Development

- Linguistic structures specific to legal docs
  - ～COLIEE 2019: classic NLP
- Insufficient data size (pretrain/finetune)
  - ～COLIEE 2020: deep language model by transfer learning (pretraining)
  - ～COLIEE 2022: ensemble of different system outputs
  - COLIEE 2023: LLM?
- General knowledge, evidence/explanations
  - ???: common sense, relationships, logic, etc.
  - COLIEE 2024 – new task/evaluation planned!

Look forward to see new participants in COLIEE 2025!