A Pentapus Grapples with Legal Reasoning

Frank Schilder,
Dhivya Chinnappa,
Kanika Madan,
Jinane Harmouche,
Andrew Vold,
Hiroko Bretz, &
John Hudzina

June 21, 2021

THOMSON REUTERS®

Overview

COLIEE 2021 Tasks

- Task 1: Legal Case Retrieval
- Task 2: Legal Case Entailment
- Task 3: Civil Code Retrieval
- Task 4: Civil Code Entailment
- Task 5: Civil Code Question Answering

Case Law Retrieval & Entailment

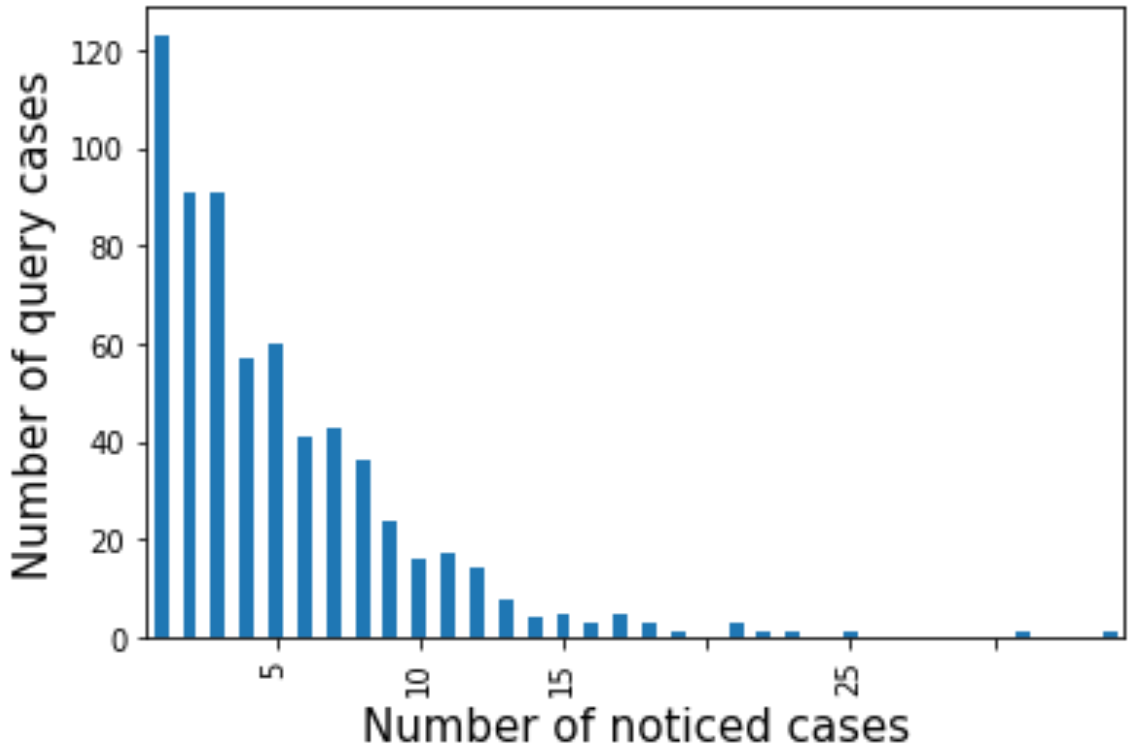
Task 1: Legal Case Retrieval: Task Description

For a given base case, identify supporting cases from the case law corpus (4415 cases)

- Training set - 650 query cases with noticed cases
- No candidate set given, thus entire corpus is the candidate set
- Most query cases have less than 10 supporting cases out of the 4414 cases in the case law corpus

Methodology:

- Generate candidate set
- Build classifiers to predict if a case is a noticed case for cases within candidate set



Task 1: Legal Case Retrieval: Task Approach

Generating candidate set:

For each query case, generate candidate set, a subset of the case law corpus including all noticed cases.

- TF-IDF similarity
- Jaccard similarity
- LDA similarity

Generating candidate set results:

Similarity technique	Rank Threshold					
	20	50	100	300	500	1000
TFIDF	.32	.47	.63	.72	.78	.87
Jaccard	.29	.38	.50	.57	.65	.75
LDA	.21	.35	.54	.66	.77	.86

Generated training set with equal number of noticed cases and unnoticed cases per query case. All unnoticed cases were from the candidate set.

Classification task:

Build classifiers over the candidate set obtained using TF-IDF

- Decision tree
- Adaboost
- Naïve Bayes
- Xgboost

Classification task Inference:

Results on validation set (F1):

Classifier	P	R	F
Decision tree	.56	.97	.71
Adaboost	.59	.56	.56
Naïve Bayes	.50	.56	.53
Xgboost	.60	.65	.62

Results on test set (Xgboost on TFIDF 1000 candidate set) F1: 0.0047

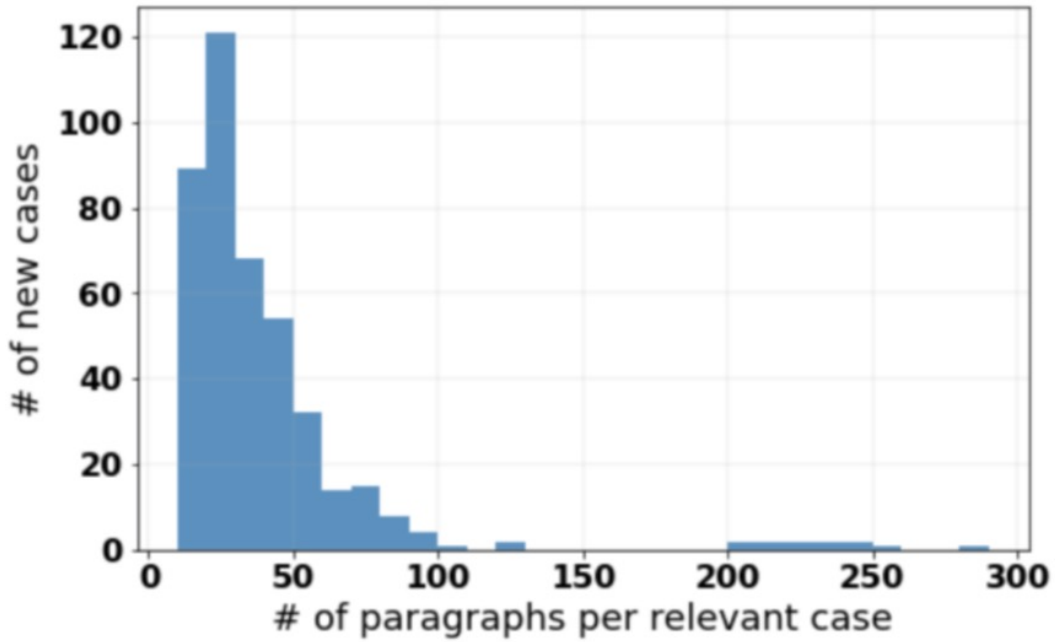
- Highly Imbalanced data distribution (validation set has balanced data)
- Might get better performance if transformer-based approaches are used

Task 2: Legal Case Entailment: Task Description

- Given a decision *Q* of a new case and a relevant case *R* with *N* paragraphs, a specific paragraph that entails the decision Q needs to be identified.

- The training dataset consists of 425 new cases

- Out of *N* paragraphs shown in the histogram, for a given query at most 4 paragraphs entail the decision paragraph



Task 2: Legal Case Entailment: Task Approach

- **Approach:** Use handcrafted similarity features to rank the different paragraphs in a case and reduce the number of potential candidates in order to train a classifier on a more balanced dataset

- **Paragraph vectorizers :** n-gram vectors, universal sentence encoder vectors, averaged word embedding vectors.

- **Classifier:** Random Forest with multi-fold cross validation for evaluation

- **Result :** F1-score of 0.56 obtained on development dataset (10% of the training set)

  F1-score of 0.54 obtained on unseen test set

Civil Code Retrieval & Entailment

Task 3 Description

Task: Retrieve most relevant civil code articles to answer given exam question

Language Used: Japanese

Strategies:

- Word Mover's Distance (WMD)
  - Alternative to TFIDF (Words do not have to be exact match)
  - Effective in selecting articles that are most similar to the question
  - The most similar articles are not always the most relevant
  - Need a better ranking strategy
- Transformer
  - Used pretrained Japanese BERT model
  - Finetuned on all premise-hypothesis pairs
  - Most likely underfit during pretraining and overfit on finetuning
  - Need data augmentation scheme and stricter regularization
  - General domain to legal domain transfer is difficult in Japanese

Team	F2 Score	Precision	Recall	MAP
OvGU_run1	0.6749	0.7778	0.7496	0.7525
JNLP. CrossLMul-tiLThreshold	0.6000	0.8025	0.7947	0.7822
BM25.UA	0.7092	0.7531	0.7037	0.7555
JNLP.CrossLBertJP	0.6241	0.7716	0.7783	0.8218
R3.LLNTU	0.6656	0.7438	0.7875	0.7921
R2.LLNTU	0.6770	0.7315	0.7893	0.7822
R1.LLNTU	0.6368	0.7315	0.7893	0.7822
JNLP.CrossLBert JPC15030C15050	0.5535	0.7778	0.7737	0.8119
OvGU_run2	0.4857	0.8025	0.7571	0.7525
TFIDF.UA	0.6790	0.6543	0.7306	0.7228
LM.UA	0.5460	0.5679	0.5432	0.6422
TR_HB	0.5226	0.3333	0.6173	0.6625
HUKB-3	0.5224	0.2901	0.6975	0.6100
HUKB-1	0.4732	0.2397	0.6543	0.6128
TR_AV1	0.3599	0.2622	0.5123	0.4653
TR_AV2	0.3369	0.1490	0.5556	0.4346
HUKB-2	0.3258	0.3272	0.3272	0.4167
OvGU_run3	0.1570	0.7006	0.5557	0.5743

Results:

Neither approaches were successful enough to rank high WMD approach ranked 14th place and the transformer Approach ranked 17 and 18th place of the 20 submissions.

Task 4 Description

Kim et al. (2019) model of the entailment task

Given a set of relevant articles as a premise $P = [P_1, P_{2, \ldots} P_n]$ and a hypotheses bar exam question *H*, determine if *P* entails *H*.

- Each article in *P* contains conditions, exceptions, and conclusions
- *H* contains facts and a conclusion
- The system must apply the facts to the articles conditions and determine if *H* can to the correct conclusion

Sent.	Text
P_1	A mandate shall terminate when the mandator or mandatary dies.
H	The mandate terminated upon the mandator's death.

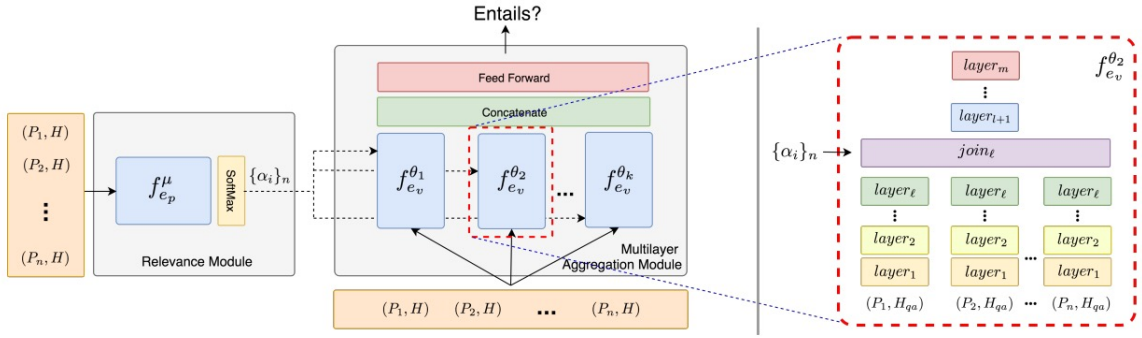
Task 4 Approach

Observations

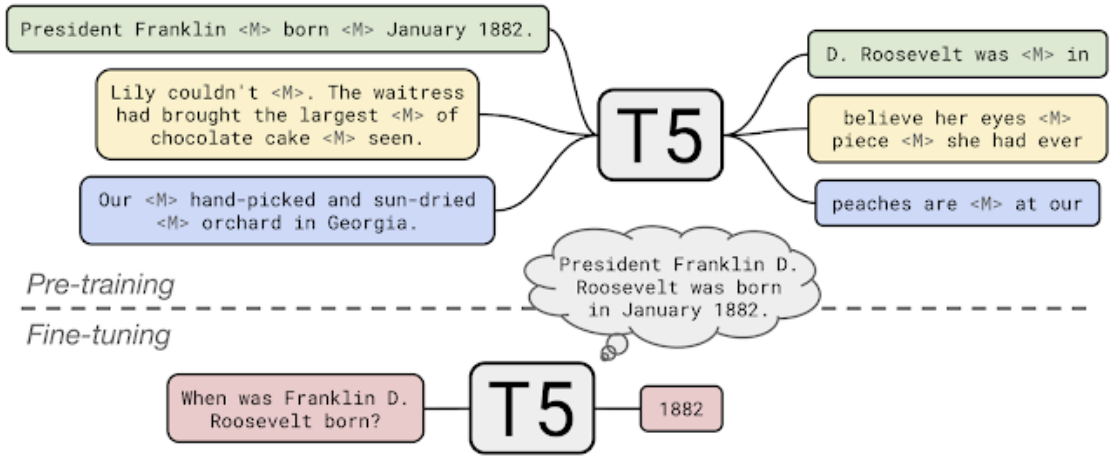
Approach must

- Match relevant phases in the *P* & *H*
- Deal with multiple conditions, exceptions and conclusions in P
- Leverage transfer learning if machine learning is used
- Handle legal terminology

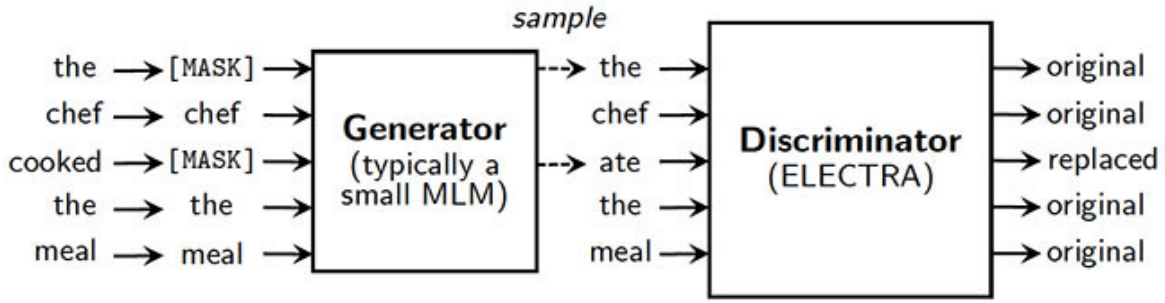
Multee: Multi-Hop Question & Answering



T5: Text to Text Transfer Transformer



ELECTRA



Task 4 - Training

No Pre-Training

Multee Training

- Pre-Trained Embeddings
  - GLoVe Embeddings
    - Common Crawl (860B,300D)
- Relevance Module
  - Multi-Genre NLI Corpus
  - Stanford NLI Corpus
- Multi-Level Aggregator
  - Modified OpenBookQA Dataset
    - Converted multiple choice to yes/no
  - COLIEE
    - Train: H18-H29
    - Val: H30

Japanese Electra

- Examined features from Japanese transformer embedding layers with respective tokenizers
  - Electra discriminator embedding distribution distances between labels is greatest
- Build classifier layer-by-layer
  - Start with one transformer layer
  - Train until validation F1 plateau
  - Add next layer, continue training
  - Repeat until validation F1 climax reached (3 layers)

T5 Experiments

Approach:

Augment COLIEE data set with similar tasks

Examine pre-training and fine-tuning approach

Pre-training & Fine-tuning Tasks

COLIEE Entailment – Reiwa 1 (R01) Validation Set

Span Corruption

- Japanese Civil Code
- Quebec & Louisiana Civil Code
- 796K US Legal Sentences

Natural Language Inference: MNLI & MultiRC

Span Extraction: Requisite and Effectuation Extraction

Training Runs

COLIEE Only Fine-Tuning (Baseline)

- T5 Base Embedding

COLIEE + Civil Code Fine Tuning

- T5 Base Embedding

Multi-Task Pretraining

- Pre-trained on non-task 4 dataset
- Fine-tuned on task 4
- Same Hyperparameters as T5-Base

Trail	Val. Acc.
COLIEE Only Fine-Tuning*	0.5405
COLIEE + Civil Code Fine Tuning	0.5135
Multi-Task Pretraining	0.46847

*Model used in ensemble

Task 4 Results

Run	Correct	Rank	Accuracy
Baseline	43	-	0.5309
HUKB2	57	1st	0.7037
TR-Ensemble	**48**	**5th (tied)**	**0.6250**
TR-MTE	**48**	**5th (tied)**	**0.6250**
TR_Electra	**41**	**9th**	**0.5062**

Discussion

- Multee
  - Weighing article relevance provide an improvement over last years attention-based model (from 0.5612 to 0.6250)
- T5
  - Multi-task fine tuning didn't preform as well compared to winning pretraining approach
  - Tasks only focused on unsupervised denoising and not explainability tasks
  - Future Work
    - Additional tasks related to statutory interpretation
    - Multi-task pretraining

Task 5: Question answering

- Approaches:
  - GPT-3: largest vs. smallest model:
    - Providing a prompt enabling few-shot learning
  - DistillRoBERTa: Sentence-based distilled model for paraphrase detection
    - Determining cut-off score for similarity score

Research questions:

- Does a large language model contain enough legal knowledge to answer questions without further specific training?
  - Is there a difference between the different models in terms of parameters?
  - Can GPT-3 produce good explanations?
- Would a smaller sentence-based BERT model fair well after a simple training phase?

GPT-3

Prompt:

This bot answers Japanese bar exam questions.

Question: *Is it true that 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?*

Answer: *Yes, because even if the seller makes a special agreement to the effect that the seller does not warrant in the case prescribed in the main clause of Article 562, paragraph (1) or Article 565, 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 personally created for or assigned to a third party.*

[…]

Results:

Team	sid	Correct	Accuracy
BaseLine	No 43/All	43	0.5309
BaseLine	Yes 38/All	38	0.4691
JNLP	JNLP.NFSP	49	0.6049
UA	UA_parser	46	0.5679
JNLP	JNLP.NMSP	45	0.5556
UA	UA_dl	45	0.5556
TR	TRDistillRoberta	44	0.5432
KIS	KIS_2	41	0.5062
KIS	KIS_3	41	0.5062
UA	UA_elmo	40	0.4938
JNLP	JNLP.task5.BERT_Multilingual	38	0.4691
KIS	KIS_1	35	0.4321
TR	TRGPT3Ada	35	0.4321
TR	TRGPT3Davinci	35	0.4321

Explanations

Question: In the cases where a lessor is the owner of the subject matter in a contract for lease, the ownership of the subject matter shall not be transfered to the lessee, on the other hand, in the cases where a lender is the owner of the subject matter in a contract for loan for consumption, the ownership of the subject matter shall be transfered to the borrower.

ada: Yes, because the provisions of the preceding two Articles apply mutatis mutandis if the right transferred by the seller to the buyer does not conform to the terms of the contract (including the case in which the seller fails to transfer part of a right that belongs to another person)

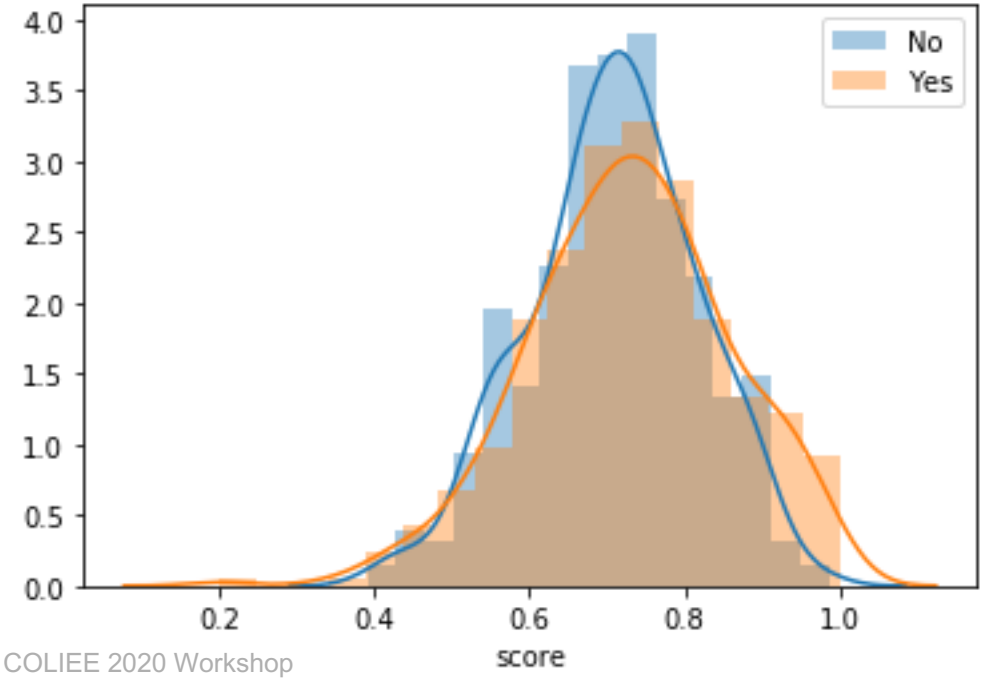
davinci: Yes, because in Article 579, paragraph (2), if a contract for lease is not terminated before the expiration of the term of the contract, the lessor has the right to cancel the lease and to reclaim the subject matter, however in the case of lender and borrower concerning a contract for loan for...

Article 579:
The buyer of immovables may cancel the sale by refunding the price (or any amount specified by agreement if otherwise agreed upon; the same applies in Article 583, paragraph (1)) and costs of the contract paid by the buyer in accordance with a special agreement on redemption executed simultaneously with the contract for sale. In such a case, unless a particular intention is manifested by the parties, it is deemed that the fruit of the immovables and the interest on the price have been set off against each other.

DistillRoBERTa

Training

- Computed the similarity score between the question and all articles
- Took the mean of all correct answers
- Used the mean as threshold for the test set.



Results

Team	sid	Correct	Accuracy
BaseLine	No 43/All	43	0.5309
BaseLine	Yes 38/All	38	0.4691
JNLP	JNLP.NFSP	49	0.6049
UA	UA_parser	46	0.5679
JNLP	JNLP.NMSP	45	0.5556
UA	UA_dl	45	0.5556
TR	TRDistillRoberta	44	0.5432
KIS	KIS_2	41	0.5062
KIS	KIS_3	41	0.5062
UA	UA_elmo	40	0.4938
JNLP	JNLP.task5.BERT_Multilingual	38	0.4691
KIS	KIS_1	35	0.4321
TR	TRGPT3Ada	35	0.4321
TR	TRGPT3Davinci	35	0.4321

Thank you