A performance study on fine-tuned large language models in the Legal Case Entailment task

Hiroaki Yamada
yamada.h.ax@m.titech.ac.jp
School of Computing, Tokyo Institute of Technology
Meguro, Tokyo

Takenobu Tokunaga
take@c.titech.ac.jp
School of Computing, Tokyo Institute of Technology
Meguro, Tokyo

ABSTRACT

Deep learning based approaches achieved significant advances in various Natural Language Processing (NLP) tasks. However, such approaches have not yet been evaluated in the legal domain compared to other domains such as news articles and colloquial texts. Since creating annotated data in the legal domain is expensive, applying deep learning models to the domain has been challenging. A fine-tuning approach can alleviate the situation; it allows a model trained with a large out-domain data set to be retrained on a smaller in-domain data set. A fine-tunable language model "BERT (Bidirectional Encoder Representation from Transformers)" [5] was proposed and achieved state-of-the-art in various NLP tasks. In this paper, we explored the fine-tuning based approach in the legal textual entailment task using the COLIEE task 2 data set. The experimental results show that the fine-tuning improves the task performance, achieving $F1 = 0.50$ with COLIEE task 2 dry run data (Our group ID: TTCL).

CCS CONCEPTS

• **Information systems** → **Decision support systems**; *Data mining*; *Information retrieval*; *Specialized information retrieval*; • **Applied computing** → **Law**.

KEYWORDS

textual entailment, legal information processing, natural language processing, neural networks, deep learning, text classification, case law, AI and law

ACM Reference Format:
Hiroaki Yamada and Takenobu Tokunaga. 2019. A performance study on fine-tuned large language models in the Legal Case Entailment task. In *Proceedings of COLIEE 2019 workshop: Competition on Legal Information Extraction/Entailment (COLIEE 2019).* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

COLIEE (Competition on Legal Information Extraction/Entailment) has been the most ambitious competitions in the area of statue law documents for years, and now it has the competition on the case law legal entailment task. The textual entailment in the legal domain is expected to be considerably more complex and challenging than other domains such as news texts and documents that are written in daily languages. The first case law legal textual entailment task in COLIEE'18 achieved at 0.26 in F1 measure [7]. The report of the best performing team in COLIEE'18 [14] concluded that the legal textual entailment task requires both deeper understanding of the problem and better embedding strategy of input sentences. The first point requires introducing domain-specific knowledge into models. The better understanding of the task domain enables to implement the better systems that utilized domain-specific characteristics such as, careful feature engineering, a citation network between cases, injecting external knowledge and building dedicated dictionaries. However, we leave this point for future research and devote this paper to the latter point – exploring the better way of embedding and encoding sentences in the legal domain.

The strategy of embedding and encoding sentences is a crucial part in neural network-based NLP. Word embedding like CBOW [10] and GloVe [13] has been a standard way to construct a vector representation of target input text for various classifiers in various NLP tasks. However, simple embedding is not sufficient for the legal entailment task to handle longer context and deeper semantics. We have to find a better way to encode entire sentences or paragraphs and to implement a classifier using the encoded features. One possible solution is a deep neural network-based model that can handle information beyond the word level, i.e., context. A bottleneck of this approach is that the model requires a large number of training instances. We can design a large model with many parameters that can handle deeper and longer context but it requires many instances to train the model properly, and it is hard to prepare large training data. The legal textual entailment is such a specialized task that annotator should have a certain level of legal knowledge to create stable and reliable training data. It makes building training data expensive. There are several solutions to the problem. Fine-tuning is one of the promising approaches. The fine-tuning retrains a model which is already trained with an out-domain data set with an in-domain data set. In COLIEE task 2, we can train a model with some large corpora (e.g., Wikipedia or Book corpus) in the first stage and later retrain it with the COLIEE task 2 data set. We can make the large model adapt to the smaller target data set by fine-tuning.

There have been several language models that are designed for a fine-tuning approach such as OpenAI GPT [15] and BERT [5] achieving state-of-the-art records in various kinds of NLP tasks including textual entailment. Our research question is to explore the effectiveness of fine-tuning in the legal textual entailment task and to identify its limitation.

COLIEE 2019, June 21, 2019, Montreal, Quebec

Table 1: Basic statistics on training data

	sentences	words
Avg. length of entailing paragraphs	3.8	104.1
Avg. length of entailed fragments	1.4	41.0

2 TASK AND DATA DESCRIPTION

We tackle COLIEE task 2, which is a legal textual entailment task in the Canadian case law system. Given a decision document Q (Base case) with an entailed fragment F and a relevant case document R (Noticed case) that includes n paragraphs as candidates ($P_R(n)$), a machine should identify paragraphs which entails Q (specifically F) from P_R. The number of entailing paragraphs is variable. We explain the task with an example in Figure 1 (Markers "ENTAILED FRAGMENT" and "ENTAILING PARAGRAPH" are added by author). The entailed fragment is Paragraph **[28]** of Base case 9, and the entailing paragraph was Paragraph **[11]** of Noticed case. This example has only one entailing paragraph.

The data source for the task is a collection of predominately Federal Court of Canada case law. In the provided corpus, there are the base case documents (Q), the entailed fragment (F), the paragraphs of noticed case documents ($P_R(n)$), and the paragraph IDs of the entailing paragraphs (answers) for the training data set. There is no entry for entailing paragraph IDs in the testing data set. Some parts of the base case documents are edited by the COLIEE organizer to replace some phrases with "FRAGNEBT_SUPPRESSED" to remove the obvious clues to resolve entailment. [11]

Table 1 shows the statistics of the data set. The average length of the entailed fragment was 104.1 in words and the average length of the entailing paragraph was 41.0 in words. Those numbers are calculated with the NLTK library [8]. The average number of the entailing paragraphs is 1.1 per case. Each entailing paragraph contains multiple sentences. The entailed fragments also can include multiple sentences though there is only one sentence in the example. As a model has to consider a broader context in this task, the problem is more difficult than conventional entailment tasks between sentences.

3 IMPLEMENTATION

We implemented baseline models with Support Vector Machine [4] and BERT-based models for COLIEE task 2. The provided base case documents include several editorial markers denoted by brackets, such as "[translation]", "[Emphasis added]", "[citations omitted]", "[my emphasis]", and "[End of document]". We removed those markers in the prepossessing stage.

3.1 Baseline models

Our baseline model is an SVM based implementation. The model takes the vector representations of both entailed fragment and target paragraph with auxiliary feature vectors and outputs a binary decision whether the pair has an entailing relationship. There are many ways of encoding sentences and paragraphs, and we prepared three ways of encoding with pre-trained embeddings and encoders — Skip-gram word2vec (W2V) [10], Neural Probabilistic Language

Base case (train-9)

...

This is an application for judicial review filed under subsection 72(1) of the Immigration and Refugee Protection Act, S.C. 2001, c. 27 (Act), against a decision by the Immigration Program Manager, Brian Ralph Hudson (Manager), of the Canadian Embassy in Beijing (embassy), People's Republic of China, dated November 18, 2004, denying the visa application of Yu Mei Zhang (applicant).

...

[27] In Mirzaii v. Canada (Minister of Citizenship and Immigration) , [2003] F.T.R. Uned. 100; 2003 FCT 213, Heneghan, J., states at paragraph 8 that a decision to issue a visa is a discretionary administrative decision and that it is therefore completely normal to rely on information gathered by an assistant. "In deciding whether to issue a visa, the Visa Officer is making an administrative decision involving the exercise of discretion. He was entitled to rely on information gathered by an assistant see Silion v. Canada (Minister of Citizenship and Immigration (1999), 173 F.T.R. 302. There is no evidence that Ms. Taheri did anything more than obtain information from the Applicant. The actual decision was made by the Visa Officer who was justified in relying on the facts obtained in the interview and recorded by Ms. Taheri in the CAIPS notes."

[28] in matters of administrative decisions, the rule of "he who hears must decide" does not apply. (ENTAILED FRAGMENT)

[29] I believe that the case law is clear on the fact that the manager need not personally conduct the interviews and the research. In conclusion, I do not think that in this case there was a violation of the principles of natural justice or of procedural fairness.

[30] THE COURT ORDERS THAT:
1. The application for judicial review be dismissed;
2. No question was submitted for certification.
Application dismissed.

Paragraphs from noticed case (14 paragraphs)

...

[10] This case concerns a decision to refuse a visa. The decision was clearly made by the visa officer, as he avers in his affidavit. This is supported by the affidavit of the IPO and further by the applicant's own affidavit which acknowledges that when she was advised of the decision to refuse her application she was told that the immigration officer made the decision, not the IPO.

[11] The decision is essentially an administrative one, made in the exercise of discretion by the visa officer. There is no requirement in the circumstances of this or any other case that he personally interview a visa applicant. There may be circumstances where failure to do so could constitute unfairness, but I am not persuaded that is the case here. Here the IPO did interview the applicant and reported on the results of that interview. That report was considered by the visa officer who made the decision. Staff processing and reporting on applications is a normal part of many administrative processes and it is not surprising it was here that followed. This is not a circumstance of a judicial or quasi-judicial decision by the visa officer which would attract the principle that he who hears must decide, or the reverse that he who decides must hear the applicant. (ENTAILING PARAGRAPH)

...

Figure 1: COLIEE Task2 query-answer sample

Model (NLM) [1], Universal Sentence Encoder(USE) [2]. Those pre-trained models are available online via Tensorflow hub[1].

W2V is a standard embedding method widely used in NLP tasks. The model is trained on the English Wikipedia corpus [9]. We tested two different models that have different vector sizes (250 and 500 dimensions). NLM is another token-based embedding technology which is built on a feedforward neural network language model. We used a model pre-trained on the English Google News corpus. We created a sentence embedding vector by normalizing weighted sum of the word embedding vectors. A sentence is represented by a 128-dimensional vector.

USE is designed to encode a sequence of more than one word. USE encodes a text into vector representations that can be used for various tasks including the text classification. We use the model to get the paragraph vectors of both entailed fragment and target paragraph. The pre-trained encoder models are available in two implementations from Tensorflow Hub. One is trained with a transformer encoder [16] and the other is trained with a deep averaging network [6] encoder. Both output a 512-dimensional vector.

As auxiliary features, we used the cosine similarity of the entailed fragment vector and the target paragraph vector, the relative position of the entailed fragment and the relative position of the target paragraph. Both relative positions are calculated as the position of the paragraph in the base case/noticed case documents divided by the total number of paragraphs in the documents.

We used the libsvm [3] based implementation provided by the scikit-learn machine learning library [12] for SVM.

3.2 BERT models

Our research question is whether language models trained with large corpora in general domains can be adapted to the legal text entailment task by the fine-tuning technique. We employ BERT in the present work. Several pre-trained models of BERT are available online[2] and they are trained on Wikipedia corpus and the BookCorpus [17]. Four pre-trained models are available with two kinds of options: two sizes (BERT-Base and BERT-Large) and two character casing options (Uncased and Cased). The Base models have 12 transformer blocks, hidden layers size of 768, 12 self-attention heads and 110M parameters while the Large models have 24 transformer blocks, hidden layers size of 1024, 16 self-attention heads, 340M parameters. The casing option means whether the text has been lowercased (uncased) or not. We use the uncased BERT-Base model in our fine-tuning experiments. The model which we used is a multi-layer Transformer encoder based architecture with approximately 110M parameters, which is similar to other Transformer based models like OpenAI GPT but employs the bidirectional self-attention.

BERT is capable of handling different language tasks such as sentence pair classification tasks, single sentence classification tasks, question answering tasks, and single sentence tagging tasks [5]. The COLIEE task 2 can be similar to a sentence pair classification task if we regard the entailed fragment and target paragraphs as "sentences". However, such straight-forward formalization is not appropriate since the pre-trained BERT model was trained for sequences up to 512 words. It means there are cases that we have to squeeze the fragments and paragraphs to that size, i.e. 512 words. According to our analysis, there are 54 cases that exceed 512 words. Instead of cutting the fragments and paragraphs, we break them into sentences and formalize the task as a simple sentence pairs task. In this new formalization, we firstly split fragments and paragraphs into sentences using NLTK [8] Punkt Sentence Tokenizer if there is more than one sentence, and make sentence pairs with every possible combination between sentences from entailed fragments and sentences from target paragraphs. Figure 2 visualizes the way of pairing. The BERT-based task 2 classifier takes each sentence pair and output whether the pair is in the relation of entailing. (Figure 3)

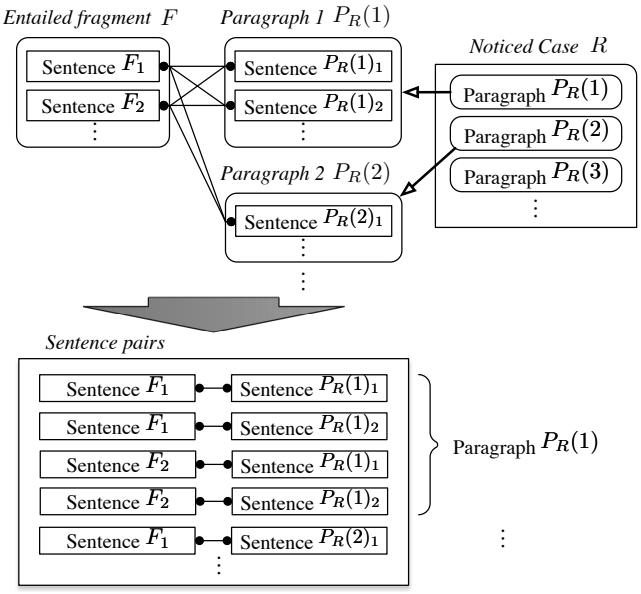

Figure 2: Visualized sentence pair preparation

To estimate an entailing relation between paragraph pairs from those of sentence pairs, we conduct the following post-processing. Given a paragraph pair, we regard the pair as an entailing pair if one or more sentence pairs across the paragraph pair are classified as "entailing."

4 EXPERIMENTS

4.1 Dry run experiments and results

Table 2: Grid search space for SVM hyperparameters

Hyper params	search space
C	1, 10, 100, 1,000
Kernels	Linear, RBF, Polynomial, Sigmoid
Gamma*	0.001, 0.0001
Degree*	2, 3, 4
Encoders	NLM, wiki500, wiki250, USE transformer, USE DAN

∗ If applicable

[1] https://tfhub.dev/
[2] https://github.com/google-research/bert

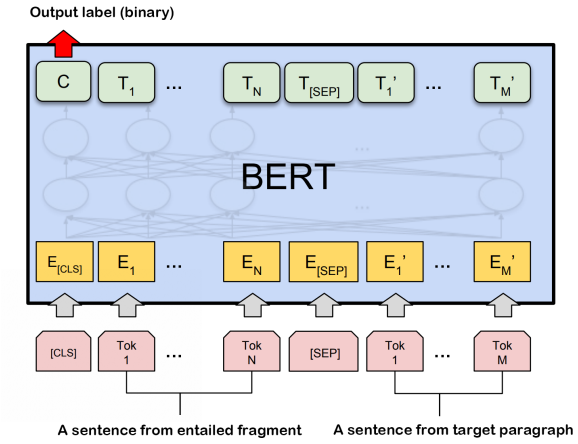

Figure 3: The model flow in COLIEE task 2 (the figure is based on and modified from [5])

We leave 20 base case documents out from the provided "train" data by the competition committee as a development set to tune hyperparameters and to select the optimal way of encoding. The remaining data is used as a training data set. Hyperparameter tuning is conducted using the grid search algorithm. Table 2 shows the search space. "NLM" stands for Neural Probabilistic Language Model based encoder. "wiki500" and "wiki250" are encoders that are based on W2V with different output vector sizes. The encoders output 500 and 250-dimensional vectors respectively. USE, universal sentence encoder has two implementations. "USE transformer" is a transformer based encoder and "USE DAN" is a deep averaging network based encoder.

As a result of tuning, we train SVM with the following parameters, $C = 10$, a linear kernel, and the USE DAN encoder.

We conducted experiments on the training data set with the five-fold cross validation to evaluate the performance of models. The data set includes 5,049 target paragraphs in total. We used an SVM model with tuned parameters and three BERT-based models with three different hyperparameter settings. Each BERT-based model allows different input sentence pair length (BERT-128, BERT-256 and BERT-512 take respectively 128, 256, 512 words as input). Also, we implemented an ensemble model BERT-vote using a simple voting method that the model will decide the output based on a majority prediction for each instance according to the three BERT-based models. Table 3 shows the results of the experiment. As for the BERT-based models, we report the results on sentence pairs based evaluation in addition to the standard paragraph based evaluation (results before merging) in Table 4. All BERT models show a significantly better result than the SVM-baseline model. BERT-vote was the best performing model among all model we implemented, achieving $F1 = 0.50$ while SVM-baseline was at $F1 = 0.22$.

4.2 Post-hoc analysis on dry run

Table 5 is a confusion matrix of the result from BERT-vote. We see that the model tends to produce more false positives than false negatives. Figure 4 shows a visualized comparison of predictions among the models in the experiment. Each square box represents a result of a prediction on each target paragraph. Blue square means true negative and green square means true positive while yellow square means false positive and red square means false negative. The figure includes only 51 cases of the provided "train" data and does not show the true negatives by all five models.

In the overall training data set (5,049 instances), 568 instances are correctly predicted by BERT-vote but not by SVM-baseline. On the other hand, 123 instances are correctly classified by SVM-baseline but not by BERT-vote. Neither SVM-baseline nor BERT-vote could correctly predict 109 instances, and we consider those instances were hard cases.

Figure 5 and Figure 6 are hard case examples from Base case 21. The true entailing paragraph is Paragraph 21, but all models failed to detect Paragraph **[21]** as entailing. Instead, BERT-256, BERT-vote, and BERT-512 predict Paragraph **[18]** and **[19]** as entailing. Interestingly, Paragraph **[18]** shares the same sentence, which is emphasized in boldface, with the entailed fragment of the Base case. The sentence should have been a misleading clue for the model in recognizing the entailment.

Table 3: Results on the training data set

Models	Precision	Recall	F1
SVM-baseline	0.14	0.52	0.22
BERT-128	0.37	0.54	0.44
BERT-256	0.36	0.66	0.46
BERT-512	0.36	0.59	0.45
BERT-vote	0.41	0.65	**0.50**

Table 4: Results on the training data set (sentence pairs)

Models	Precision	Recall	F1
BERT-128	0.43	0.12	0.18
BERT-256	0.41	0.14	**0.21**
BERT-512	0.42	0.12	0.19

Table 5: Confusion matrix for BERT-vote

		Predicted labels: Positive	Predicted labels: Negative	Total
True labels	Positive	118	64	182
	Negative	168	4,699	4,867
	Total	286	4,763	5,049

4.3 Formal run results

According to the results of the experiment on the training data set, we submitted three models, BERT-256, BERT-512 and BERT-vote to the formal run of COLIEE task 2. Those models are trained with all instances of the provided "train" data. Table 6 shows the results of the formal run. The best model was BERT-256 achieving $F = 0.53$.

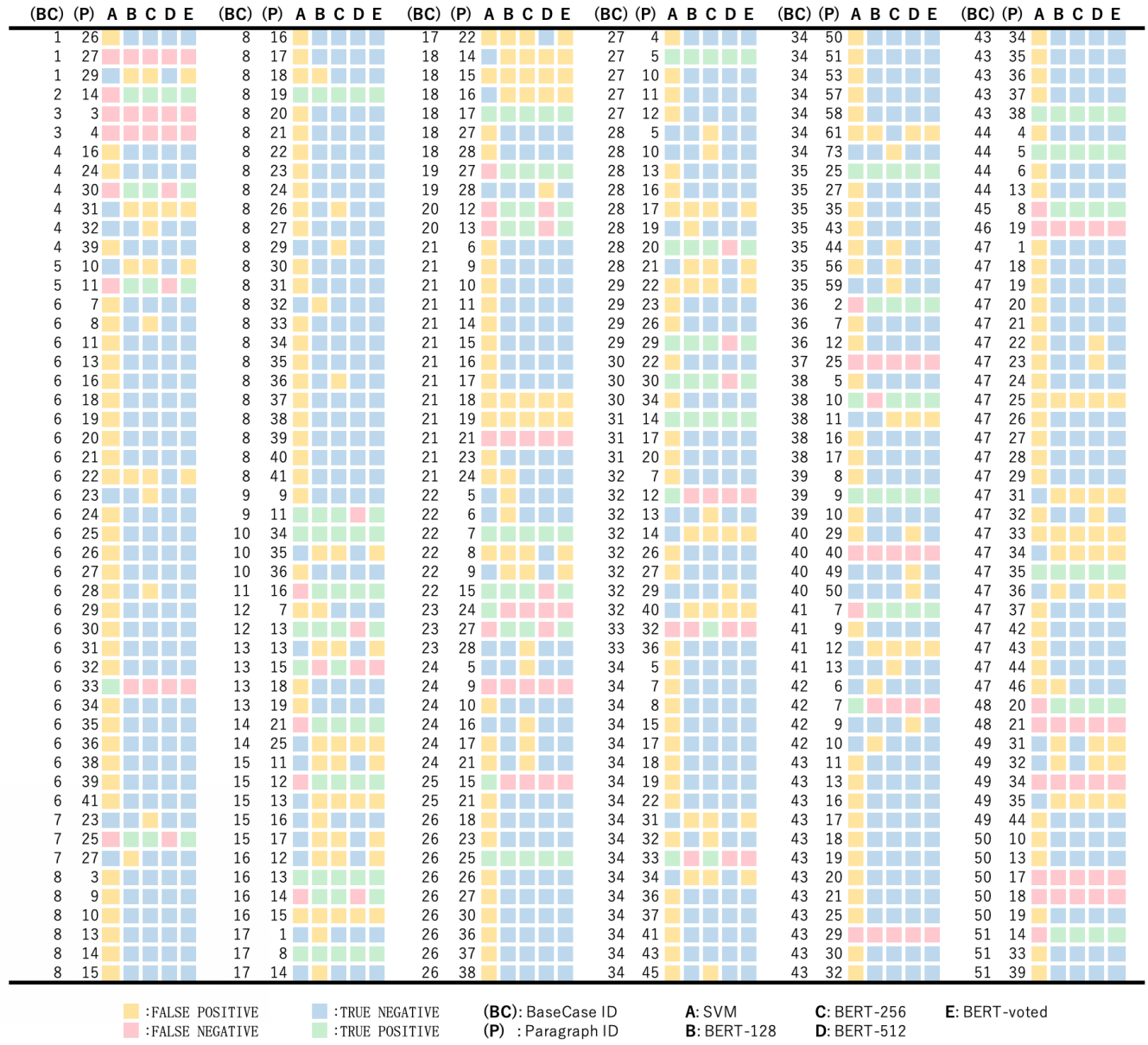

Figure 4: Visualized results of cross validation with training data (First 51 cases)

Table 6: Results of formal run

Models	Precision	Recall	F1
BERT-256	0.40	0.80	**0.53**
BERT-512	0.38	0.69	0.49
BERT-vote	0.39	0.73	0.51

5 CONCLUSION

In this paper, we proposed an SVM-based baseline model and four BERT-based models and compared their performance on COLIEE task 2. The SVM-based model adopted a linear kernel and inputs by the universal sentence encoder. The BERT-based models were pre-trained with the out-domain data (Wikipedia corpus and Book corpus) and fine-tuned with the COLIEE task 2 data. According to our dry run experiments, the BERT models performed significantly better than the baseline model and the ensemble model was the best among the four BERT-based models, achieving $F = 0.50$. The score is significantly improved from $F = 0.26$, which was the best score in the last competition although the data set is different. We successfully fitted the model to the legal textual entailment task by fine-tuning without any feature engineering. This result shows

Entailed fragment

Only where unreasonable omissions have been made, such as the failure to investigate obviously crucial evidence, is judicial review warranted

Entailing paragraph (correct answer, 021.txt)

[21] With respect to judicial review of decisions of the Canadian Human Rights Commission, Jerome, A.C.J., held in Lukian v. Canadian National Railway Co. (1994), 80 F.T.R. 38 (T.D.):

"Generally, when Courts are called upon to review the exercise of an administrative tribunal's discretionary power, they will be reluctant to interfere since tribunals, by virtue of their training, experience, knowledge and expertise, are better suited than the judiciary to exercise those powers. Provided the Commission's decision is within the discretion given to it, the Court will not interfere with the manner in which it was exercised, unless it can be shown the discretion was exercised contrary to law. What the law requires is the Commission to consider each individual case before it, to act in good faith, to have regard to all relevant considerations and not be swayed by irrelevant ones, and to refrain from acting for a purpose contrary to the spirit of its enabling legislation or in an arbitrary or capricious manner."

Figure 5: Hard case sample correct answer (train-21)

Entailing paragraphs (predicted by BERT-vote, 018.txt, 019.txt)

[18] In Slattery, supra, at 20 to 23, Nadon, J., required the elements of neutrality and thoroughness to be present in the Commission's investigation as these elements are necessary to provide a fair basis on which the Commission evaluates whether a tribunal should be appointed pursuant to s. 44(3)(a) of the Act. With regard to neutrality, it has been held that if the Commission simply adopts an investigator's conclusions without giving reasons, and those conclusions were made in a biased manner, a reviewable error occurs: Canadian Broadcasting Corp. v. Canadian Human Rights Commission et al. (1993), 71 F.T.R. 214 (T.D.). The requirement of thoroughness of investigation is discussed in Slattery and stems from the essential role that investigators play in determining the merits of particular complaints. In determining the degree of thoroughness of investigation required to be in accordance with the rules of procedural fairness, the interests of the complainant and the respondent must be balanced with the Commission's interest in maintaining an administratively workable system. Deference must be given to administrative decision-makers to assess the probative value of evidence and to decide whether to further investigate accordingly. **Only where unreasonable omissions have been made, such as the failure to investigate crucial evidence, is judicial review warranted.** In instances where parties have the legal right to make submissions in response to an investigator's report, parties may be able to compensate for minor omissions by bringing them to the attention of the decision-maker. Therefore, only where complainants are unable to rectify such omissions should judicial review be considered.

[19] In Slattery, supra, at 28, Nadon, J., set up a high threshold for judicial review of Commission decisions on the ground of thoroughness: "The fact that the investigator did not interview each and every witness that the applicant would have liked her to and the fact that the conclusion reached by the investigator did not address each and every alleged incident of discrimination are not in and of themselves fatal as well. This is particularly the case where the applicant has the opportunity to fill in gaps left by the investigator in subsequent submissions of her own. In the absence of guiding regulations, the investigators, much like the CHRC, must be master of its own procedure, and judicial review of an allegedly deficient investigation should only be warranted where the investigation is clearly deficient."

Figure 6: Hard case sample predictions (train-21)

the possibility of the fine-tuning with large pre-trained language models in the legal text entailment task. Although the BERT-based models showed good results, there is room for further improvement. We observed some cases where the models recognized non-entailing sentences as entailing due to the surface level matching. Introducing the domain knowledge and the appropriate representation of legal augmentations might remedy such errors. Furthermore, a separate feature extractor and a bypassed network for the legal dedicated knowledge could help to solve the problem.

REFERENCES

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3, Feb (2003), 1137–1155.

[2] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *CoRR* abs/1803.11175 (2018). arXiv:1803.11175 http://arxiv.org/abs/1803.11175

[3] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27.

[4] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (01 Sep 1995), 273–297. https://doi.org/10.1007/BF00994018

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 1681–1691.

[7] Mi-Young Kim, Yao Lu, Juliano Rabelo, and Randy Goebel. 2018. COLIEE-2018: Evaluation of the Competition on Case Law Information Extraction and Entailment. In *Proceedings of the Twelfth International Workshop on Juris-informatics (JURISIN 2018)*. 105–116.

[8] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1 (ETMTNLP '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 63–70. https://doi.org/10.3115/1118108.1118117

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[11] COLIEE 2019 organizing committee. 2019. COLIEE-2019. https://sites.ualberta.ca/~rabelo/COLIEE2019/. Accessed: 2019-04-10.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[13] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[14] Juliano Rabelo, Mi-Young Kim, Housam Babiker, Randy Goebel, and Nawshad Farruque. 2018. Legal Information Extraction and Entailment for Statute Law and Case Law. In *Proceedings of the Twelfth International Workshop on Juris-informatics (JURISIN 2018)*. 142–155.

[15] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Technical report* (2018).

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[17] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*. 19–27.