Generalizing over data sets: a preliminary study with BERT for Natural Language Inference*

Rubem G. Nanclarez[1], Norton T. Roman[1], Fernando J. V. da Silva[2]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)

[2]N2VEC Tecnologia

{rnanclarez, norton}@usp.br, fernando@n2vec.com

Abstract. *Natural language inference is the task of automatically identifying whether a given text (premise) implies another (hypothesis). Among multiple possible applications, it is especially relevant in the legal field to understand textual entailment between legal sentences, being the focus of recent research efforts. In this work, we evaluated the usage of BERT for natural language inference by conducting experiments and comparing results obtained by testing on a larger corpus with texts from multiple domains and a smaller corpus of legal sentences. Furthermore, we conducted a cross-experiment by training on the larger corpus and testing on the legal corpus. As a result, we obtained a mean accuracy of 88.91% in the corpus with multiple domains, a value comparable to related work. However, the same technique presented lower scores in the legal corpus and the cross-experiment.*

1. Introduction

Natural language inference is a classification task that seeks to determine an implication relationship between a given premise (p) and a hypothesis (h) [Ghuge and Bhattacharya 2014]. In other words, a system capable of performing such a task should be able to infer whether the given hypothesis is true based only on the given premise.

The natural language inference task is of great importance since, in addition to having great theoretical relevance, being even compared to the Turing test [Bos and Markert 2005], it is a task with several practical applications. In medicine, for example, [Saini et al. 2020] uses inference to summarize images in medical articles, and [Zhang et al. 2020] uses NLI in a patient triage system.

The legal field, in particular, can take great advantage of artificial intelligence models, especially models capable of performing natural language inference. With this in mind, the COLIEE (Competition on Legal Information Extraction/Entailment) has been promoted as a competition that seeks to elevate artificial intelligence models aimed at the legal environment to the state of the art[Goebel et al. 2021].

In recent editions, we have seen a significant increase in the use of large language models, especially those based on transformers[Vaswani et al. 2017]. Hence, this study

*The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and from the IBM Corporation.

aimed to understand the use of BERT [Devlin et al. 2019] for natural language inference. In order to achieve that, corpora with different sizes and complexities were used, and the impacts that such characteristics exert on the model's final performance were analyzed. In addition, an analysis was performed in order to understand whether or not it is possible to take advantage of the knowledge obtained by training the model in a more generic corpus, but with larger data, in a more specific and smaller one.

Our experiments showed us that, although the BERT could have great performance being trained on a large generic corpus like the SNLI, the direct use on small or from a domain-specific corpus like the legal domain may not bring the same results. It could indicate that, in these cases, external information may be necessary to enrich the model.

This article is structured as follows. Section 2 describes the corpora as well as the experiments performed. Section 3 discusses the results found. Section 4 discusses some approaches on the task. Finally, section 5 presents some conclusions and possible future works.

2. Materials and Methods

This study was performed using BERT [Devlin et al. 2019] both as a language model and a classifier since we tried to understand the impact that corpora with different characteristics, both in size and domain, can have on large distributed models like BERT, in the task of natural language inference. The choice for BERT as a model was due to its excellent results in several tasks, such as Sentiment Analysis[Hoang et al. 2019], question answering[Qu et al. 2019] and machine translation[Imamura and Sumita 2019], in addition to its wide use in the latest editions of COLIEE [Goebel et al. 2021].

BERT consists of a transformer-based encoder stack [Vaswani et al. 2017], coming in two reference models with different sizes, *i.e.* with different sets of parameters. Its smallest version, $\text{BERT}_{\text{base}}$, has 12 encoder blocks, an internal size (*i.e.* the size of the matrices in the attention layer) of 768 and 12 attention heads, totaling 110 million parameters. In its largest version, $\text{BERT}_{\text{large}}$ has 24 encoders, with an internal size of 1,024 and 16 attention heads, totaling 340 million parameters.

Due to the involved computational costs, and since our focus lied at comparing BERT's performance in different (although related) corpora, in this work, we elected $\text{BERT}_{\text{base}}$ as our classifier. The model was then created using the Pytorch library, with a linear layer for classification, in Google's Colab[1] computing platform. In all our experiments, we used a configuration with five epochs and a batch size of 32.

For our first testing corpus, we relied on SNLI[2] (the Stanford Natural Language Inference corpus), which was built to solve the problem of the scarcity of data sets for training natural language inference models [Bowman et al. 2015]. Comprising 570,152 sentence pairs in English, written and annotated by humans, this is a multi-class corpus, where each sentence pair is annotated with one out of three classes, to wit, Entailment, Contradiction, and Neutral. The corpus is well balanced, with 190,113 pairs at entailment, 189,218 pairs at neutral, and 189,702 at contradiction. The remaining 1,190 pairs were labeled with '-' at the validation phase and are not used in the experiments.

[1] https://colab.research.google.com/

[2] https://nlp.stanford.edu/projects/snli/

Table 1 shows some sample pairs from each class. In the first column, we see the final label assigned to the pair. The pair consists of a premise that must be taken as true and a hypothesis to be characterized according to one of the possible labels.

Table 1. Sample pairs at the SNLI corpus

Entailment	**Premise**: A soccer game with multiple males playing **Hypothesis**: Some men are playing a sport.
Neutral	**Premise**: An older and younger man smiling. **Hypothesis**: Two men are smiling and laughing at the cats playing on the floor.
Contradiction	**Premise**: A man inspects the uniform of a figure in some East Asian country. **Hypothesis**: The man is sleeping

To build the corpus, approximately 2,500 annotators were presented with premises extracted from captions from a public image bank. They were then asked to create a sentence corresponding to each of the labels. At the validation step, 10% of the corpus was presented to 4 other annotators (different from those who produced the original content). There were five labels for each of these pairs taken from the corpus: the label assigned when the pair was produced and four additional labels from the validators. If at least three annotators had chosen one of the three labels, then that label was deemed the gold standard for that pair. Otherwise, the pair was labeled with '-'. There was a consensus of 98% among three of the annotators and 58% among five of the annotators.

Our second corpus was the one distributed at the 2021 COLIEE[3] (Competition on Legal Information Extraction/Entailment). COLIEE is an event that takes place annually to improve artificial intelligence models for the legal domain and comprises a series of tasks involving information retrieval and natural language inference. Although all tasks can be approached with natural language inference techniques, in this work, we chose Task 4 because we believe it presents a better-suited corpus for natural language inference.

Task 4 is a task to determine textual entailment between a given problem sentence and article sentences. Competitor systems should then answer "yes" if the problem entails the article sentence, or "no" otherwise [Goebel et al. 2021]. As with SNLI, COLIEE's corpus is balanced, comprising sentence pairs extracted from a Japanese bar exam on Civil Law and translated into English. In this corpus, sentence pairs are assigned with one of two possible labels (as opposed to SNLI's three classes), *yes* and *no*, which correspond to the answers given at the exam. In COLIEE, a total of 806 sentence pairs build the corpus, 409 of which are labeled as "yes" and 397 as "no". Table 2 shows an example pair for each of the classes in COLIEE.

From Tables 1 and 2, one sees that the complexity of both corpora is considerably different. To start with, COLIEE corresponds to a binary classification problem, whereas SNLI has three classes. Moreover, the sentences at COLIEE are longer, also presenting a domain-specific vocabulary, with no mentions of external references, such as articles in the law. This is not the case with SNLI, where the pairs are shorter, free-domain, and

[3] https://icail.lawgorithm.com.br/workshop/coliee/

Table 2. Example pairs from COLIEE.

Yes	**Problem**: 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. **Article**: 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.
No	**Problem**: The provisions of the preceding three 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). Article 566 If the subject matter delivered by the seller to the buyer does not conform to the terms of the contract with respect to the kind or quality, and the buyer fails to notify the seller of the non-conformity within one year from the time when the buyer becomes aware of it, the buyer may not demand cure of the non-conformity of performance, demand a reduction of the price, claim compensation for loss or damage, or cancel the contract, on the grounds of the non-conformity;provided, however, that this does not apply if the seller knew or did not know due to gross negligence the non-conformity at the time of the delivery. **Article**: There is a limitation period on pursuance of warranty if there is restriction due to superficies on the subject matter, but there is no restriction on pursuance of warranty if the seller's rights were revoked due to execution of the mortgage.

reference-free. Finally, COLIEE, with its 806 sentence pairs, is considerably smaller than SNLI, which accounts for a total of 190,113 pairs.

In this research, the experiments were performed using 10-fold validation, where each corpus was divided into ten parts (folds), 9 of which were used for training and 1 for validation. Our BERT model was then trained at the nine training folds and validated at the remaining one, with this process being repeated ten times, one for each of the ten validation folds. To deal with the fact that one corpus has multiple classes and the other is binary, initially two experiments were carried out with SNLI. In the first experiment, all three classes were used, whereas in the second experiment "neutral" and "contradiction" were merged, thereby making up a binary corpus and allowing for a better comparison with COLIEE.

As a last experiment, we took binary SNLI's best accuracy instance of the trained model, amongst the ten folds, and tested it in all of COLIEE's validation sets (*i.e.* the same sets where the model instances trained at COLIEE's training folds were tested).

So that we could not only test this SNLI instance model with all versions produced at COLIEE, but also compare its performance in both corpora. Finally, it is worth recalling that both corpora are equivalent regarding their labels, with COLIEE's "yes" representing SNLI's "Entailment", and COLIEE's "no" covering the union of SNLI's "Neutral" and "Contradiction".

3. Results and Discussion

The 10-fold cross-validation execution of $\text{BERT}_{\text{base}}$ in SNLI resulted in a mean accuracy of 94.95% ($\sigma = 0.03\%$) at the training folds, and 88.91% ($\sigma = 0.12\%$) in the validation folds. Although the reduction in mean accuracy lies around 6% only, it might be an indication of some overfitting to the model. Also, accuracy values varied almost three times as much in the validation folds as in the testing folds, as measured by their standard deviation from the mean, indicating a greater variance in the validation data (although that might be caused by the reduced size of each validation fold, when compared to their training counterparts).

At COLIEE, $\text{BERT}_{\text{base}}$'s mean accuracy across the ten folds was 61.08% ($\sigma = 4.95\%$) in the training folds and 55.47% ($\sigma = 3.31\%$) in validation, representing a decrease of around 9% in accuracy, which might indicate a possible low generalization capability of the model (and its consequent overfitting to the training data). Interestingly, and contrary to what happened in SNLI, the standard deviation from mean accuracy was around 33% lower in the validation folds than in the training folds. This is a puzzling result, especially given that validation sets were 11% the size of their training counterparts.

This, however, could have been caused by the fact that COLIEE comprises translated real-life legal texts, meaning that they come with at least two layers of noise, to wit, the fact that texts are longer and more elaborated, and the possibility of something being lost in translation. Still, COLIEE is a binary-classed domain-specific data set, whereas SNLI is a three-classed and open domain, which would be expected to increase variation in it. Given that the same model (*i.e.* different instances of $\text{BERT}_{\text{base}}$) was run on both, this is an issue we believe deserves attention from future work.

To make both data sets more comparable, we also run $\text{BERT}_{\text{base}}$ in a binary-classed version of SNLI, where "neutral" and "contradiction" were merged to build a "Non-Entailment" class. Mean accuracy, across the 10 folds, at the training set was 94.15% ($\sigma = 0.95\%$), with 92.14% ($\sigma = 0.17\%$) at the validation set. As expected, turning the data set into a binary classification problem has reduced variance, as illustrated by the small reduction in accuracy (around 2%) between training and validation sets. Interestingly, there is almost no difference in accuracy, at the training sets, between SNLI's binary and ternary versions.

In our last experiment, we decided to investigate how a model trained at SNLI might perform when tested in COLIEE. To this end, we took the best accuracy instance of the model in binary SNLI and ran it in the ten validation folds of COLIEE. When doing so, mean accuracy across the ten folds was only 50.46% ($\sigma = 5.29\%$), only slightly above plain chance (recall that this is a binary problem). However bad, this result is nonetheless around 9% below the instance produced by training the same model in COLIEE (which, as mentioned, resulted in a 55.47% accuracy). This is yet another evidence of COLIEE's complexity and variance.

Obtained results are shown in Figure 1 and summarised in Table 3. In this table, one sees the mean accuracy of $\text{BERT}_{\text{base}}$, both in training and validation sets, across all tested data sets. As a final note, it is worth mentioning that observed differences were found to be of statistical significance ($ANOVA(df = 6) = 436, p \ll 0.001$, at the 95% confidence level). A pairwise comparison between conditions (*i.e.* all combinations of data set and training and validation accuracies) showed there to be no significant differences[4] only between training and validation sets, in binary SNLI, and between SNLI and its binary version (for both their training and validations sets), along with between SNLI's training set and binary SNLI's validation set. All other differences (including between SNLI's validation set and binary SNLI's training set) were found to be significant.

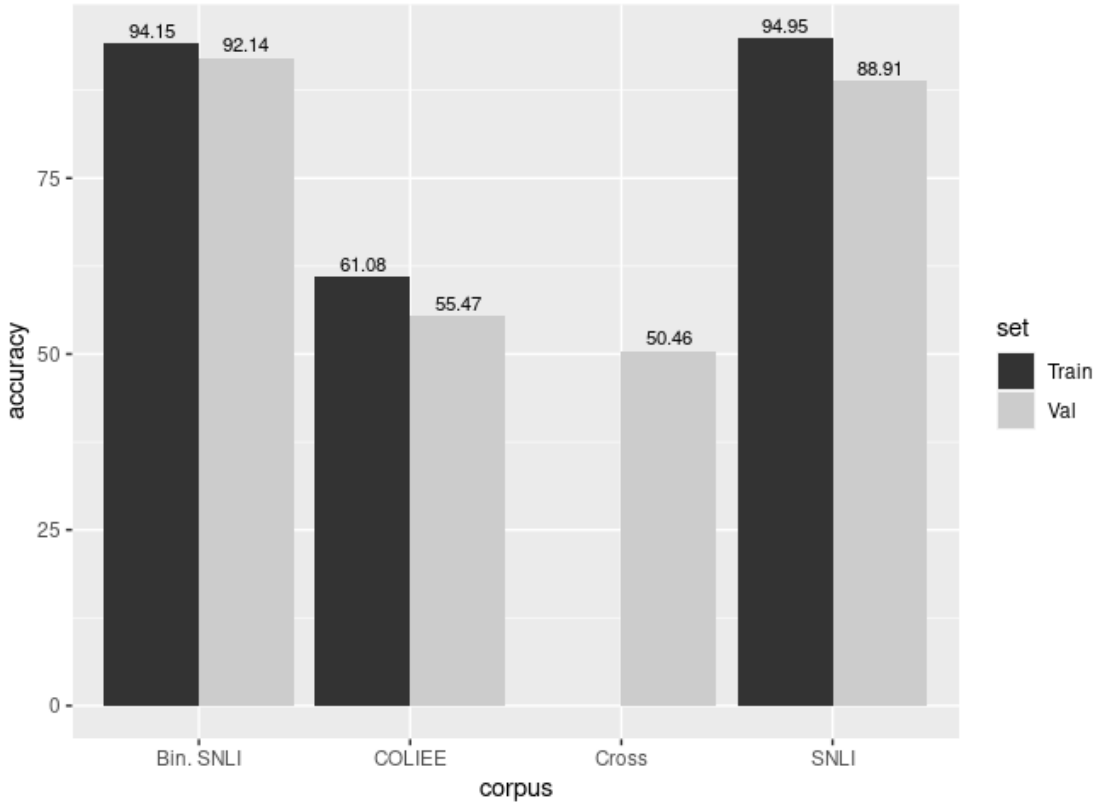

Figure 1. Mean accuracy for training and test sets across data sets.

Table 3. Mean accuracy across the 10 folds in each dataset, at the training and test sets

	Training	Validation
SNLI	$94.95\% \pm 0.03\%$	$88.91\% \pm 0.12\%$
COLIEE	$61.08\% \pm 4.95\%$	$55.47\% \pm 3.31\%$
Binary SNLI	$94.15\% \pm 0.95\%$	$92.14\% \pm 0.17\%$
Cross test	–	$50.46\% \pm 5.29\%$

[4] As indicated by a *post-hoc* Tuckey test, at the 95% confidence level, with correction for multiple testing.

As it turns out, although SNLI is significantly larger than COLIEE's corpus, having around 370k sentence pairs in the binary version we used for the cross-test, while COLIEE has around 800 pairs, the low complexity of SNLI, along with the fact that this is an open domain corpus (whereas COLIEE focus on law texts), may have played an important role in these results. Our cross-corpus experiment shows that these differences can have a great impact in the final performance of the model.

4. Comparison to Related Work

The SNLI is one of the most important corpora for the Natural Language Inference task. Therefore, a wide variety of models use SNLI to solve the NLI task. For example, [Du et al. 2020] seeks to explore the syntactic relationship between the hypothesis and the premise with a relation-head-dependency(RHD) model. These triples are processed by a neural model with attentional mechanisms reaching 0.875 accuracy in SNLI. [Quamer et al. 2021] uses a convolutional neural network to generate the representation of the premise and hypothesis.

These representations are submitted to internal and cross-attention devices, and finally, the final representation, used for classification, is generated by the fusion of these various intermediate representations, reaching an accuracy of 0.897. [Lian and Lan 2019] uses a model based on Matching Aggregation with the application of two layers of attention, one applied directly to the initial representation of the sentences and another after a layer that uses Bi-LSTM to capture contextual information of each sentence. They achieved an accuracy of 0.863.

One of the problems with the data from COLIEE is its size. To approach that [Yoshioka et al. 2021a, Yoshioka et al. 2021b] uses an ensemble of BERT with data augmentation to perform the task. They argue that there are two types of possible entailment, semantic and logical. The method takes the case where the logic determines the entailment relation and creates a new case by flipping the logical connectors. They achieve an accuracy of 0.7037

Although the BERT is a multilingual model, [Nguyen et al. 2021] argues that more explicit foreign knowledge could improve the model. In order to achieve this, they propose two approaches, Next Foreign Sentence Prediction (NFSP) and Neighbor Multilingual Sentence Prediction (NMSP). The NFSP is similar to the next sentence prediction task in BERT, where the model has to answer if two sentences are consecutive. The difference is that one sentence is translated, forcing the model to have a better generalization. The NMSP tries to generalize even more. In this task, same language sentences were used, and the model was asked if a sentence was next to another and if it was in normal or reverse order, non-contiguous, or just a random sampling. They achieve an accuracy of 0.6296.

[Schilder et al. 2021] uses an ensemble of models with transfer learning. In the first model, they implemented a multi-sentence Natural Language Inference model called Multee [Trivedi et al. 2019]. Second, they build Electra[Clark et al. 2020] model from the bottom up, assuming that, once the COLIEE corpus is relatively small compared to regular pre-training models, it would be better to train the model layer by layer. The Last model used T5[Raffel et al. 2020] with several different pre-training phases. Their best model achieves an accuracy of 0.5926.

Based on the premise that the usage of external information can enrich pre-trained models like BERT. [Kim et al. 2021] uses semantic information from the Kadokawa thesaurus to improve BERT. For this, the semantic category number was used as an additional feature. They achieved an accuracy of 0.6667.

5. Conclusion

In this article, we tried to understand the usage of a large language model like BERT in the natural language inference task. To achieve that, we first performed an experiment training the model on the SNLI. The experiment was performed using the original corpus and also a binary version of the corpus. Our experiments achieved 88.91% accuracy with the multi-class corpus and 92.14% with the binary corpus, which is comparable to related works.

After that, we used COLIEE's corpus, which is considerably more complex because it has longer sentences from a specific domain and also has a significantly smaller number of samples. This helped us understand how these differences can impact the model's performance. The resulting accuracy of 55.47% was lower than related works, which indicates that, for this type of corpus, the enrichment of the model may be necessary to improve its performance.

Our last experiment tried to understand if we could take advantage of training a larger corpus like SNLI in a smaller and more complex corpus like COLIEE. For that, we performed a cross-experiment in which we used the model instance with the best result presented in the training of the binary SNLI and tested it in the same validation corpora used in the training of the COLIEE. The drop in accuracy to 50.46% compared to the instance trained directly on the COLIEE's corpus may be a result of the great difference in complexity and size of the model. To address these issues, future work may analyze the use of pre-trained models in the legal domain or the enrichment of the model with external information, for example, syntactic or semantic information.

References

Bos, J. and Markert, K. (2005). Recognising textual entailment with logical inference. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05, pages 628–635, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs].

Du, Q., Zong, C., and Su, K.-Y. (2020). Conducting natural language inference with word-pair-dependency and local context. ACM Transactions on Asian and Low-Resource Language Information Processing, 19(3).

Ghuge, S. and Bhattacharya, A. (2014). Survey in Textual Entailment. Center for Indian Language Technology, page 28.

Goebel, J. R. R., Kano, Y., Kim, M.-Y., Yoshioka, M., and Satoh, K. (2021). Summary of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021. Proceedings of the Eigth International Competition on Legal Information Extraction/Entailment (COLIEE 2021).

Hoang, M., Bihorac, O. A., and Rouces, J. (2019). Aspect-Based Sentiment Analysis using BERT. In Proceedings of the 22nd Nordic Conference on Computational Linguistics, pages 187–196, Turku, Finland. Linköping University Electronic Press.

Imamura, K. and Sumita, E. (2019). Recycling a Pre-trained BERT Encoder for Neural Machine Translation. In Proceedings of the 3rd Workshop on Neural Generation and Translation, pages 23–31, Hong Kong. Association for Computational Linguistics.

Kim, M.-Y., Rabelo, J., and Goebel, R. (2021). BM25 and Transformer-based Legal Information Extraction and Entailment. Sao Paulo, page 6.

Lian, Z. and Lan, Y. (2019). Multi-layer attention neural network for sentence semantic matching. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence, ACAI 2019, pages 421–426, New York, NY, USA. Association for Computing Machinery.

Nguyen, H.-T., Tran, V., Nguyen, P. M., Vuong, T.-H.-Y., Bui, Q. M., Nguyen, C. M., Dang, B. T., Nguyen, M. L., and Satoh, K. (2021). ParaLaw Nets – Cross-lingual Sentence-level Pretraining for Legal Text Processing. arXiv:2106.13403 [cs].

Qu, C., Yang, L., Qiu, M., Croft, W. B., Zhang, Y., and Iyyer, M. (2019). BERT with History Answer Embedding for Conversational Question Answering. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, pages 1133–1136, New York, NY, USA. Association for Computing Machinery.

Quamer, W., Jain, P. K., Rai, A., Saravanan, V., Pamula, R., and Kumar, C. (2021). SACNN: Self-attentive convolutional neural network model for natural language inference. ACM Transactions on Asian and Low-Resource Language Information Processing, 20(3).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

Saini, N., Saha, S., Bhattacharyya, P., and Tuteja, H. (2020). Textual Entailment–Based figure summarization for biomedical articles. ACM Transactions on Multimedia Computing Communications and Applications, 16(1s).

Schilder, F., Chinnappa, D., Madan, K., Harmouche, J., Vold, A., Bretz, H., and Hudzina, J. (2021). A Pentapus Grapples with Legal Reasoning. page 9.

Trivedi, H., Kwon, H., Khot, T., Sabharwal, A., and Balasubramanian, N. (2019). Repurposing Entailment for Multi-Hop Question Answering Tasks.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. (2017). Attention is All you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Yoshioka, M., Aoki, Y., and Suzuki, Y. (2021a). BERT-based ensemble methods with data augmentation for legal textual entailment in COLIEE statute law task. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, ICAIL '21, pages 278–284, New York, NY, USA. Association for Computing Machinery.

Yoshioka, M., Suzuki, Y., and Aoki, Y. (2021b). BERT-based Ensemble Methods for Information Retrieval and Legal Textual Entailment in COLIEE Statute Law Task. page 6.

Zhang, X., Xiao, C., Glass, L. M., and Sun, J. (2020). DeepEnroll: Patient-trial matching with deep embedding and entailment prediction. In Proceedings of the Web Conference 2020, WWW '20, pages 1029–1037, New York, NY, USA. Association for Computing Machinery.