BERT-based Ensemble Methods with Data Augmentation for Legal Textual Entailment in COLIEE Statute Law Task

Masaharu Yoshioka
yoshioka@ist.hokudai.ac.jp
Faculty of Information Science and Technology, Hokkaido University
Graduate School of Information Science and Technology, Hokkaido University
Sapporo-shi, Hokkaido, Japan

Yasuhiro Aoki
Youta Suzuki
yasu-a_01@eis.hokudai.ac.jp
suzuki@eis.hokudai.ac.jp
Graduate School of Information Science and Technology, Hokkaido University
Sapporo-shi, Hokkaido, Japan

ABSTRACT

The Competition on Legal Information Extraction/Entailment (COLIEE) statute law legal textual entailment task (task 4) is a task to make a system judge whether a given question statement is true or not by provided articles. In the last COLIEE 2020, the best performance system used bidirectional encoder representations from transformers (BERT), a deep-learning-based natural language processing tool for handling word semantics by considering their context. However, there are problems related to the small amount of training data and the variability of the questions. In this paper, we propose a BERT-based ensemble method with data augmentation to solve this problem. For the data augmentation, we propose a systematic method to make training data for understanding the syntactic structure of the questions and articles for entailment. In addition, due to the nature of the non-deterministic characteristics of BERT fine-tuning and the variability of the questions, we propose a method to construct multiple BERT fine-tuning models and select an appropriate set of models for ensemble. The accuracy of our proposed method for task 4 was 0.7037, which was the best performance among all submissions.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; *Ensemble methods.*

KEYWORDS

Textual entailment, Data augmentation, BERT, Ensemble method

ACM Reference Format:
Masaharu Yoshioka, Yasuhiro Aoki, and Youta Suzuki. 2021. BERT-based Ensemble Methods with Data Augmentation for Legal Textual Entailment in COLIEE Statute Law Task. In *Eighteenth International Conference for Artificial Intelligence and Law (ICAIL'21), June 21–25, 2021, São Paulo, Brazil.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3462757.3466105

ICAIL'21, June 21–25, 2021, São Paulo, Brazil

ACM ISBN 978-1-4503-8526-8/21/06…$15.00
https://doi.org/10.1145/3462757.3466105

1 INTRODUCTION

The Competition on Legal Information Extraction/Entailment (COLIEE) [3, 4, 10, 11, 15] serves as a forum to discuss issues related to legal information retrieval (IR) and entailment. There are two types of tasks in COLIEE. One is a task using case law (tasks 1 and 2), and the other is a task using Japanese statute law with Japanese bar exam questions (tasks 3, 4, and 5). Task 3 is an IR task that aims to retrieve (a) relevant law article(s) to judge whether the statement of the question is true, task 4 is an entailment task that judges whether a given relevant article entails a given question statement and task 5 is a combination of tasks 3 and 4.

Because a part of bar exam questions are based on real use cases, it is important to have a mechanism for semantic matching to discuss the relevance of words in the articles and those in the questions for entailment. At an earlier stage, machine-readable thesauruses such as WordNet[7] and distributed representation of words such as Word2Vec[6] were used. Recently, the deep learning-based natural language processing tool bidirectional encoder representations from transformers (BERT) [1] was introduced. One of the characteristics of BERT is that it provides a general semantic analysis system that can be fine-tuned for a particular task. For the last COLIEE [10], the best performance systems for tasks 3 [12] and 4 [9] used BERT as a core component of the system.

In this paper, we propose a method to use BERT-based ensemble methods for task 4. This method utilizes BERT with data augmentation that increases training examples by making article-and-question pairs systematically using sentences in the statute law articles. We also propose a system that ensembles the results from multiple BERT-based system outputs. The accuracy of the system for task 4 was 0.7037, which was the best performance among all the submitted runs for task 4 at COLIEE 2021.

2 RELATED WORKS

Because bar exam questions include questions about real use cases of articles, it is necessary to discuss the correspondence between the concepts used in the articles and real use cases. In the early stage of COLIEE, several attempts were made to utilize resources for discussing such semantic matching, such as a machine-readable thesaurus and data for the distributed representation of the terms. For example, Mi-Young et al. [5] used Word2Vec [6] as a resource for distributed representation, and Taniguchi et al. [14] proposed a method to utilize WordNet [7] as a machine-readable thesaurus. However, because those methods cannot handle the context to

estimate the meaning of such terms, they are not as effective for utilizing such resources.

Recently, Devlin et al. [1] proposed BERT, a deep learning-based natural processing tool pretrained for solving general tasks that require semantic information with larger corpora (such as the whole contents of Wikipedia). Based on this training process, BERT can handle the meaning (distributed representation) of words in a sentence by considering the context. In addition, BERT can be used for various tasks by employing a fine-tuning process that utilizes comparatively small numbers of training data. Because a pretrained model of BERT contains rich information about the semantics of the words, the fine-tuned models may be able to handle semantic information even though the words themselves are not included in the training data.

At COLIEE 2020, the BERT-based system achieved the best performance for legal textual entailment tasks (JNLP [9]). In that paper, they proposed a lawfulness classification approach that classified the appropriateness of legal statements by using many legal sentences that include bar exam questions provided by organizers without considering given relevant articles. This approach worked well for COLIEE 2020 because of the large number of training data. In addition, they also pointed out that it was difficult to select an appropriate model using validation data for the unseen questions because of the significant variability of the questions.

To increase the size of the training data, the data augmentation approach is widely used in the field of image recognition [13]. However, few studies related to the data augmentation method have been conducted for the legal textual entailment task. Min et al. [8] proposed a syntactic-based data augmentation method to increase the robustness of natural language inferences. They proposed a systematic method to create positive and negative data from the correct inference sentence by a syntactic operation such as passivation and the inversion of subject and object. Evans et al. [2] proposed a method for data augmentation for logical entailment. In this framework, their method increased negative and positive data by modifying logical inference rules using symbolic vocabulary permutation, which includes an operation to make implication rules that share the same contents for the condition and derived parts. Those approaches are useful to design data augmentation methods for legal textual entailment.

3 BERT-BASED ENSEMBLE LEGAL TEXTUAL ENTAILMENT SYSTEM

Based on a discussion of the previous best performance system (JNLP [9]), we propose a system with the following characteristics.

(1) Textual entailment approach with data augmentation
We assume that the reason why the lawfulness classification approach outperformed the textual entailment one in the last COLIEE is the size of the training data. Therefore, when we provide larger training data by data augmentation, the textual entailment approach may outperform the lawfulness classification approach because it uses the most important information (relevant articles).
(2) Ensemble results of multiple BERT-based model outputs
As discussed, it is difficult to select appropriate models for the task by only evaluating the validation model. From our preliminary experiment (the details are discussed in Section 3.3), we confirmed that the characteristics of the fine-tuned BERT-based model are different and that the accuracy of the validation data is not directly related to that of the test data. We assume that this result reflects the different characteristics of each model and that the appropriate selection of the generated models for ensemble may improve the performance for the unseen questions.

3.1 Data augmentation using articles

In the deep learning framework, it is common to enlarge training data by modifying the existing data (data augmentation). However, it is important to define the appropriate data augmentation method to obtain better results. Related to the legal textual entailment task, data augmentation methods have been used for natural language and logical inference, as introduced in Section 2. However, it is difficult to apply these methods to this legal textual entailment data.

In this research, we assume that there are two types of errors to judge whether an article entails a given question. One is semantic mismatch, and the other is logical mismatch (the appropriateness of the judicial decision).

For example, let us discuss the example of training data using the following article (a part of article 9): "A juridical act performed by an adult ward is voidable."

(1) "A juridical act performed by an adult is voidable."
The article does not entail this question because of semantic matching ("adult" is not "adult ward").
(2) "A juridical act performed by an adult ward is not voidable."
The article does not entail this question because of the inappropriateness of the juridical decision ("voidable" and "not voidable")
(3) "A juridical act performed by an adult is not voidable."
We cannot judge whether this question is true (it may require another article). However, a given article cannot entail the question.

For the semantic matching case (1), it is difficult to select appropriate pairs ("adult" and "adult ward") for replacement to make such a semantic mismatch sentence. For both cases (3), it is also difficult to make the data and to use these data for negative examples to identify types of errors to judge the entailment results.

By contrast, if we make the pair of correct answers with logical mismatch cases (2), the examples may help to explain the importance of comparisons between the judicial decision of the article and that of the question.

Based on this assumption, we create training data that characterize the logical mismatch between the articles and questions.

The procedures to make this augmented data are as follows.

(1) Extraction of (a) judicial decision parts from the articles.
If there are multiple decisions in an article, the sentences in the article are split into smaller sentences that contain one judicial decision (Figure 1). When the split sentence explains an exceptional case, a flipped judicial decision is complemented for the split sentence (underlined part of the split sentences). When the sentence does not contain any

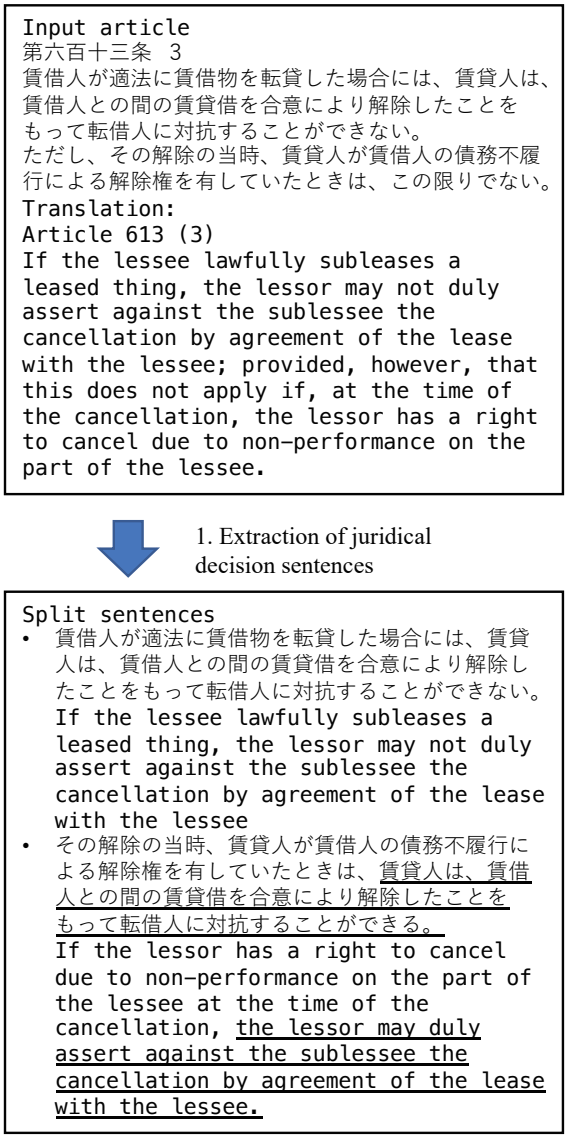

Figure 1: Example of splitting sentences (Article 613(3))

judicial decision (e.g., definitions of terms), those sentences are used only to generate positive pairs.

(2) Make positive and negative data by using extracted sentences.

We use a sentence extracted from the step 1 as an article part and the same sentence text as a question for a positive example pair. We made a sentence that is generated by flipping the judicial decision. For most of the cases, we add "ない" (not) or remove "ない" (not) for the judicial decision verb. We also use the antonym dictionary to make flipped sentences. Pairs of the original sentence and flipped sentence are used as negative example pairs. Figure 2 shows an example of making negative pairs.

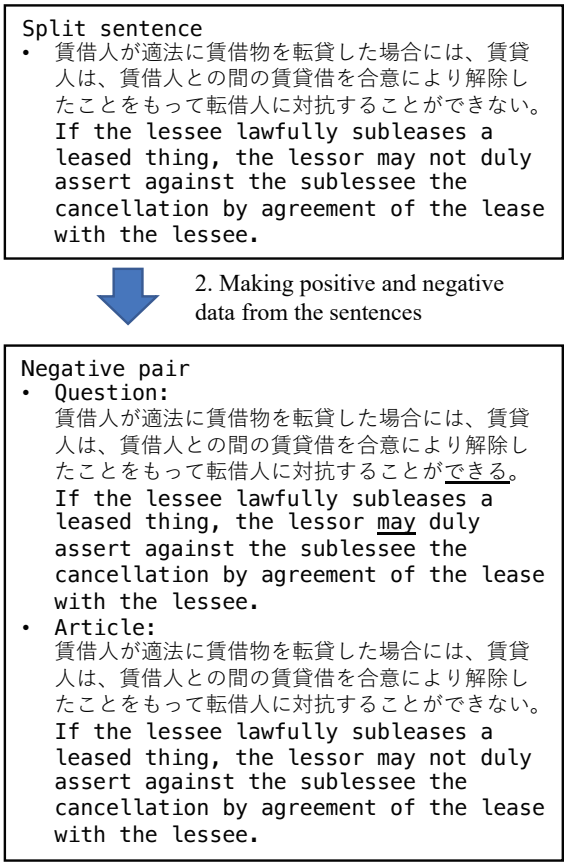

Figure 2: Example of splitting sentences (Article 613(3))

Based on this process, we construct 3,331 (positive: 1,677, negative: 1,654) training examples for data augmentation.

3.2 BERT-based entailment system

We implemented a BERT-based entailment system using the ordinal BERT fine-tuning process proposed in [1]. We concatenated the question and article using a sentence-separator token ([*SEP*]) and fed it into the BERT model to estimate whether the article entails a question (positive:1) or not (negative:0). We use the BERT-based model of BERT-Japanese [1].

Training and validation data are constructed from the training bar exam questions from H18–30 (13 years of data with 695 questions) by randomly splitting 90% (625) for training and 10% (70) for validation. All augmented data are merged with training data, and we use 3,956 examples for training and 70 for validation. We also made training sets without using augmented data (625 training and 70 validation examples) for comparing system performance without augmented data.

The fine-tuning of the BERT model is done using a max sequence length of 256, adam as an optimizer, a training batch size of 12, and a learning rate of 1e-5. The validation loss is calculated at each

[1]https://github.com/cl-tohoku/bert-japanese

Table 1: Evaluation results of the 10 models

Model No.	Validation Loss	Validation Accuracy	Test Accuracy
1	0.6935	0.4857	0.5946
2	0.7247	0.5286	0.6667
3	0.7566	0.6286	0.6486
4	0.6822	0.6286	0.6486
5	0.7347	0.5143	0.6486
6	0.7745	0.6143	0.6396
7	0.6913	0.5429	0.6126
8	0.7123	0.5857	0.6486
9	0.7504	0.6286	0.6486
10	0.7735	0.5857	0.6396

epoch and stop-training process when the validation loss increases. We use a model with minimal validation loss.

Fine-tuned models accept a pair of question statements and (an) article(s) as input and return whether the article(s) entail(s) the statement (positive) or not (negative), which is decided by comparing the score for the probability of being positive or negative.

3.3 Preliminary experiment

To evaluate the performance of the proposed BERT-based entailment system, we conducted a preliminary experiment using the R01 data (1 year of data with 111 questions) for evaluation. To discuss the effect of the variability of the training and validation questions set, we made 10 models using the same procedures. Because of the non-deterministic characteristics of the BERT fine-tuning process and different training sets selected randomly, we expected the system to construct different models that use different features for analyzing the texts. Table 1 shows the evaluation results for the validation and test data. As shown in the table, the validation accuracy and loss were not closely related to the test accuracy. We assume that these results reflect the variability of the question set.

We also made another 10 models without using augmented data. The average accuracy of these 10 models was 0.5108 (best: 0.5946, worst: 0.4505). From this comparison, we confirmed that data augmentation is effective to improve the performance of the BERT training process.

In this inference process, we can estimate the confidence of the BERT model output by comparing the probability of positive or negative. When the probability of positive is almost equal to 1(0), the system outputs positive (negative) results with higher confidence. By contrast, when the probability is close to 0.5, the system output can be interpreted as less confidence.

When we checked the distribution of such confidence for each question, the tendency of such confidence was not consistent among these models. From this observation, we assumed that these models may use different features to estimate the entailment results and that their characteristics may differ, even though we used the same model architecture for training. In such a case, there is a probability to increase the performance of the overall system by ensembling the results of different models.

Table 2: Evaluation results of the ensemble models

Model used	Accuracy
(1, 2, 3, 4, 5, 6, 7)	0.694
(1, 2, 4)	0.689
(1, 2, 3)	0.685
(1, 2, 3, 4, 7)	0.682
(1, 2, 6)	0.676
(1, 2, 5)	0.676
(1, 2, 3, 4, 5, 7)	0.676
(1, 2, 3, 4, 5, 6)	0.676
(1, 2, 3, 4, 5)	0.676
...	...
(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)	0.622

Table 3: Evaluation results of the ensemble models

Submission ID	Model used
HUKB-1	(1, 2, 3, 4, 5, 6, 7)
HUKB-2	(1, 2, 4)
HUKB-3	(1, 2, 4, 7 , 8)

To discuss the appropriate settings for making the ensemble model, we calculated the performance of the ensemble model using these 10 models. For making the ensemble model, we used the average probability of positive and negative from the target models. Table 6 shows the evaluation results of the ensemble models by accuracy. In this table, all combinations of the ensemble models are used (selecting three to 10 models from those introduced in Table 1).

There were large differences between the accuracies of the ensemble cases. The best accuracy system ensembled seven models, and the worst used all models. Most of the cases that used three to five models were adequate to estimate the results with better accuracy.

All of the highest rank sets contained the best performance system **model 2**. In addition, they also used **model 1**, even though the accuracy of **model 1** was the lowest among these 10 models. This suggests that it is important to use a complementary set of models that have different characteristics to improve the overall performance of the ensemble models.

3.4 Submitted results

Based on the results of the preliminary experiments, we submitted the following three results that used different model sets for the ensemble.

HUKB-1 and HUKB-2 were the best and second-best performance systems using R1 data as a kind of validation data. HUKB-3 selected the five best models using validation loss information.

Table 4 shows the final evaluation results of all submission runs, among which, HUKB-2 achieved the highest accuracy.

3.5 Discussion

To understand the effect of the ensemble method, we compared the performance of the ensemble results with one of each model.

Table 4: Final evaluation results

Submission ID	Correct	Accuracy
BaseLine	No 43/All 81	0.5309
HUKB-2	57	0.7037
HUKB-1	55	0.6790
HUKB-3	55	0.6790
UA_parser	54	0.6667
JNLP.Enss5C15050	51	0.6296
JNLP.Enss5C15050SilverE2E10	51	0.6296
JNLP.EnssBest	51	0.6296
OVGU_run3	48	0.5926
TR-Ensemble	48	0.5926
TR-MTE	48	0.5926
OVGU_run2	45	0.5556
KIS1	44	0.5432
KIS3	44	0.5432
UA_1st	44	0.5432
KIS2	43	0.5309
UA_dl	43	0.5309
TR_Electra	41	0.5062
OVGU_run1	36	0.4444

Table 5: Evaluation results of the 10 models for the test data

Model No.	Accuracy
1	0.6790
2	0.6666
3	0.5185
4	0.5555
5	0.6666
6	0.5308
7	0.6790
8	0.5925
9	0.5555
10	0.5308

Table 5 shows the evaluation results of the 10 models. This year, the basic model performed well and the best performance systems were almost equivalent to the ensemble ones. However, the appropriate selection of the models (HUKB-2) made the ensemble results better than one for each model.

These results justify the appropriateness of using the ensemble method by selecting an appropriate ensemble set using validation data.

Table 6 shows the number of questions classified by agreement level among the models used. "Agree," "Majority," and "Other" represent "all models return the same results," "final results are same as majority voting," and others, respectively. From these results, we can confirm that the average calculation ensemble method is better than majority voting because the number of correct questions for others is larger than the number of wrong ones. For the "Agree" questions, the best performance system (HUKB-2) had the largest numbers because of the small number of used models (three), but the accuracy of HUKB-1 (using seven models) for "Agree" was better than that of HUKB-2. However, the accuracy of HUKB-3 was lower than that of HUKB-2, which suggests that selecting an appropriate set of models for the ensemble is also effective for maintaining the accuracy of the "Agree" questions.

Table 6: Number of questions classified by the ensemble results

Submission ID	Agree		Majority		Other	
	Correct	Wrong	Correct	Wrong	Correct	Wrong
HUKB-1	17	4	27	14	11	8
HUKB-2	28	9	23	11	6	4
HUKB-3	19	9	17	12	19	9

Table 7: Topic difficulty analysis based on the number of correct runs

No. of correct runs	No. of questions	No. of correct answers by HUKB-2
1–3	7	0
4–6	11	1
7–9	12	9
10–12	19	15
13–15	19	19
16–18	13	13

Second, we analyze the characteristics of our system based on the difficulty estimated by the number of runs that return the correct answer provided by organizers. Table 7 shows the number of questions corresponding to the number of correct runs from the 18 submitted runs (Table 4). Questions with a smaller number of correct runs may be common difficult problems among all submitted methods.

From this table, we confirm that our method answers the easy questions consistently. These characteristics may come from our ensemble method reducing the effect of variability of the training data sets.

By contrast, our system performs poorly for difficult questions, suggesting common problems that nearly all submitted systems cannot handle at this moment.

We would like to discuss the characteristics of such difficult questions using examples.

The following question (Figure 3) is a difficult question that only one run can answer correctly. Because the main terms appear in both the question and the first sentence, the systems tend to say positive (entail) for this question. However, it also matches the last sentence that explains an exceptional case of the articles. As a result, the given article does not entail the question.

Because our data augmentation method splits the sentences and only handles flipped negative cases, as introduced in Section 3.2, our system cannot answer this question correctly either. However, because several articles have such exceptional cases, it may be better to propose a data augmentation method to handle such articles.

The following failure example (Figure 4: one run can answer correctly) is also related to the logical expression (quantifier). The article says "together with the obligee" (more than two), but the

Question: R02-25-E

賃借人が適法に賃借物を転貸し，その後，賃貸人が賃借人との間の賃貸借を合意により解除した場合，賃貸人は，その解除の当時，賃借人の債務不履行による解除権を有していたときであっても，その合意解除をもって転借人に対抗することはできない。

If the lessee lawfully subleases a leased thing, the lessor may not duly assert against the sublessee the cancellation by agreement of the lease with the lessee even if the lessor has a right to cancel due to non-performance on the part of the lessee at the time of the cancellation.

Article for entailment (answer is No)

第六百十三条３　賃借人が適法に賃借物を転貸した場合には、賃貸人は、賃借人との間の賃貸借を合意により解除したことをもって転借人に対抗することができない。ただし、その解除の当時、賃貸人が賃借人の債務不履行による解除権を有していたときは、この限りでない。

Article 613 (3) If the lessee lawfully subleases a leased thing, the lessor may not duly assert against the sublessee the cancellation by agreement of the lease with the lessee; provided, however, that this does not apply if, at the time of the cancellation, the lessor has a right to cancel due to non-performance on the part of the lessee.

Figure 3: Example of the failure of a difficult question

question says "independently" (single). For this case, it is not so easy to make a simple data augmentation method for handling this type of logical mismatch.

The following failure example (Figure 5: three runs can answer correctly) is related to the logical expression and semantic mismatch. The article says "the other party to the contract gives consent" (require consent), but the question says "regardless of whether A consents." Because there are no patterns for handling such logical mismatches in augmented data, it is comparatively difficult for the system to identify this type of logical mismatch. In addition, the vocabulary used for representing related persons is totally different; "A", "B", and "E" are used for the questions and "one of the party", "the other party" and "the third party", are used in the article. It is also difficult for the system to estimate the relationship among them.

4 SUMMARY

In this paper, we introduced our system for participating in task 4 (legal textual entailment) of COLIEE 2021. This system uses a BERT-based entailment system with data augmentation by flipping the judicial decision of the article sentences. We also proposed a method to make various BERT models and selected an appropriate ensemble model set using a validation data set. The effectiveness of the proposed system was evaluated by COLIEE 2021 task 4 (textual entailment task), and the accuracy of our system was 0.7037, which

Question: R02-19-I

保証人は，被担保債権の一部を弁済したが残債務がある場合，その弁済をした価額の限度において，代位により取得した被担保債権及びその担保権を単独で行使することができる。

If a guarantor has partially paid a secured claim but there is a remaining obligation, the guarantor may independently exercise the secured claim and the security right acquired through subrogation in proportion to the value of the subrogee's performance.

Article for entailment (answer is No)

第五百二条　債権の一部について代位弁済があったときは、代位者は、債権者の同意を得て、その弁済をした価額に応じて、債権者とともにその権利を行使することができる。

Article 502　(1)　If performance by subrogation occurs with respect to one part of a claim, the subrogee, with the consent of the obligee, may exercise the rights of the subrogee together with the obligee in proportion to the value of the subrogee's performance.

Figure 4: Example of the failure of a difficult question (2)

Question: R02-23-U

ＡＢ間においてＡの所有する中古の時計甲の売買契約が締結された場合、Ｂが，Ｅとの間で，売買契約における買主たる地位をＥに譲渡する旨の合意をした場合，Ａの承諾の有無にかかわらず，買主たる地位はＥに移転する。

If B makes an agreement with E to transfer contractual status of the buyer to E, regardless of whether A consents, the status of the buyer is transferred to E.

Article for entailment (answer is No)

第五百三十九条の二　契約の当事者の一方が第三者との間で契約上の地位を譲渡する旨の合意をした場合において、その契約の相手方がその譲渡を承諾したときは、契約上の地位は、その第三者に移転する。

Article 539-2 If one of the parties to a contract made an agreement with a third party to transfer that party's contractual status to that third party, and the other party to the contract gives consent to the transfer, the contractual status is transferred to the third party.

Figure 5: Example of the failure of a difficult question (3)

was the best among all runs. We also discussed the characteristics of the failure of our system for future development.

ACKNOWLEDGMENT

We thank the organizers of the COLIEE for their efforts in constructing this test data. This work was partially supported by JSPS KAKENHI Grant Number 18H0333808.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[2] Richard Evans, David Saxton, David Amos, Pushmeet Kohli, and Edward Grefenstette. 2018. Can Neural Networks Understand Logical Entailment?. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=SkZxCk-0Z

[3] Yoshinobu Kano, Mi-Young Kim, Randy Goebel, and Ken Satoh. 2017. Overview of COLIEE 2017. In *COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment (EPiC Series in Computing, Vol. 47)*, Ken Satoh, Mi-Young Kim, Yoshinobu Kano, Randy Goebel, and Tiago Oliveira (Eds.). EasyChair, 1–8.

[4] Mi-Young Kim, Randy Goebel, Yoshinobu Kano, and Ken Satoh. 2016. COLIEE-2016: Evaluation of the Competition on Legal Information Extraction and Entailment. In *The Proceedings of the 10th International Workshop on Juris-Informatics (JURISIN2016)*. Paper 11.

[5] Mi-Young Kim, Ying Xu, and Randy Goebel. 2017. Applying a Convolutional Neural Network to Legal Question Answering. In *New Frontiers in Artificial Intelligence*, Mihoko Otake, Setsuya Kurahashi, Yuiko Ota, Ken Satoh, and Daisuke Bekki (Eds.). Springer International Publishing, Cham, 282–294.

[6] 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*. 3111–3119.

[7] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[8] Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic Data Augmentation Increases Robustness to Inference Heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2339–2352. https://doi.org/10.18653/v1/2020.acl-main.212

[9] Ha-Thanh Nguyen, Hai-Yen Thi Vuong, Phuong Minh Nguyen, Binh Tran Dang, Quan Minh Bui, Sinh Trong Vu, Chau Minh Nguyen, Vu Tran, Ken Satoh, and Minh Le Nguyen. 2020. JNLP Team: Deep Learning for Legal Processing. In *The Proceedings of the 14th International Workshop on Juris-Informatics (JURISIN2020)*. The Japanese Society of Artificial Intelligence„ 195–208.

[10] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2020. COLIEE2020:Methods for Legal Document Retrieval and Entailment. In *The Proceedings of the 14th International Workshop on Juris-Informatics (JURISIN2020)*. The Japanese Society of Artificial Intelligence„ 114–127.

[11] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2020. A Summary of the COLIEE 2019 Competition. In *New Frontiers in Artificial Intelligence*, Maki Sakamoto, Naoaki Okazaki, Koji Mineshima, and Ken Satoh (Eds.). Springer International Publishing, Cham, 34–49.

[12] Hsuan-Lei Shao, Yi-Chia Chen, and Sieh-Chuen Huang. 2020. BERT-based Ensemble Model for The Statute Law Retrieval and Legal Information Entailment. In *The Proceedings of the 14th International Workshop on Juris-Informatics (JURISIN2020)*. The Japanese Society of Artificial Intelligence„ 223–234.

[13] Connor Shorten and T. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6 (2019), 1–48. https://doi.org/10.1186/s40537-019-0197-0

[14] Ryosuke Taniguchi, Reina Hoshino, and Yoshinobu Kano. 2019. Legal Question Answering System Using FrameNet. In *New Frontiers in Artificial Intelligence*, Kazuhiro Kojima, Maki Sakamoto, Koji Mineshima, and Ken Satoh (Eds.). Springer International Publishing, Cham, 193–206.

[15] Masaharu Yoshioka, Yoshinobu Kano, Naoki Kiyota, and Ken Satoh. 2018. Overview of Japanese Statute Law Retrieval and Entailment Task at COLIEE-2018. In *The Proceedings of the 12th International Workshop on Juris-Informatics (JURISIN2018)*. The Japanese Society of Artificial Intelligence„ 117–128.