ignore at SemEval-2024 Task 5: A Legal Classification Model with Summary Generation and Contrastive Learning

Binjie Sun
School of Information Science and Engineering
Yunnan University
sunbinjie@stu.ynu.edu.cn

Xiaobing Zhou
School of Information Science and Engineering
Yunnan University
zhouxb@ynu.edu.cn

Abstract

This paper describes our work for SemEval-2024 Task 5: The Legal Argument Reasoning Task in Civil Procedure. After analyzing the task requirements and the training dataset, we used data augmentation, adopted the large model GPT for summary generation, and added supervised contrastive learning to the basic BERT model. Our system achieved an F1 score of 0.551, ranking 14th in the competition leaderboard. Our system achieves an F1 score improvement of 0.1241 over the official baseline model.

1 Introduction

In Task 5 of SemEval-2024: The Legal Argument Reasoning Task in Civil Procedure (Bongard et al., 2022), we expect to reason about legal arguments in civil actions, as shown in Table 1. The dataset for this task comes from a textbook for law students, and we believe it is a complex task that can be benchmarked against modern legal language models. Task 5 proposes a novel NLP task from the US civil procedure domain that is beneficial to the quest to improve modern legal language models.

The foundation model we choose is Legal-BERT (Chalkidis et al., 2020), which collects different English LEGAL texts from multiple domains (e.g., legislation, court cases, contracts) for pre-training. Compared with other models such as LEGAL-RoBERTa (Chalkidis* et al., 2023), it can handle this task data better. Based on that, a great variety of strategies have been tested along with our exploration, such as summary generation, data augmentation (DA), and contrastive learning.

Data analysis for this task revealed that the dataset size was relatively small (only 666 entries), yet each data point contains substantial information. In such a language environment, we realize using and enriching data fully is very important. We used generative summarization, contrastive learning, and data augmentation to train the model, which led to our system ranking 14th in this task.

2 Related Work

Legal information is mostly expressed in the form of text, such as legal cases, bills, contracts, legislation, and so on. Therefore, legal text processing is an important area of NLP, including classifying legal topics (Nallapati and Manning, 2008), generating rulings based on what the court has already done (Ye et al., 2018), etc. In the past, some traditional machine learning methods like SVM bag of words (Aletras et al., 2016; Medvedeva et al., 2018) performed worse than neural models on legal tasks. The use of generic pre-trained models becomes the new paradigm, such as Legall Longformer (Chalkidis* et al., 2023) and Italian-Legal-BERT (Licari and Comandè, 2022). Data augmentation is a mature method for expanding a dataset when there is little training data, and in this case, we are not using external data but rather taking full advantage of the various fields of the provided data.

Task 5 is a small sample task, and we adopt contrastive learning to distinguish them from different samples by grouping similar samples together, hoping to learn from the intrinsic structure of the data. We use triples as a loss function(Schroff et al., 2015), and according to the characteristics of our task, we use a supervised contrast learning(Khosla et al., 2020) algorithm, where the triples are anchor points, positive samples, and negative samples.

3 System Overview

Our baseline system simply feeds Legal-BERT with two pieces of text, classifies its output [CLS] tokens, and scores their similarity with the human-annotated data by cross-entropy loss training. All the optimized strategies discussed below are based on this framework, and the overall framework of our final system is shown in Figure 1. After train-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 530–535

key	value
Introduction	My students always get confused about the relationship between removal to federal court and personal jurisdiction. Suppose that a defendant is sued in Arizona and believes that she is not subject to personal jurisdiction there .[...]Fed. R. Civ. P. 4(k)(1)(A). I've stumped a multitude of students on this point. Consider the following two cases to clarify the point.
Question	7. A switch in time. Yasuda, from Oregon, sues Boyle, from Idaho, on a state law unfair competition claim, seeking $250,000 in damages. He sues in state court in Oregon.[...] Five days after removing, Boyle answers the complaint, including in her answer an objection to personal jurisdiction. Boyle's objection to personal jurisdiction is
Answer Candidate	not waived by removal, but will be denied because the federal courts have power to exercise broader personal jurisdiction than the state courts.
Label	0

Table 1: An example in the training set.

ing with all positive policies, we ensemble the best model on each fold for the final prediction.

3.1 Data Augmentation

In this task, we augment the training data in two ways. First, we combine the explanation, the question, and the complete analysis corresponding to the answer to form new positive sample data by utilizing the fields of the complete parsing of the answer. Second, the analysis corresponding to the wrong answer is combined with the answers to other questions to form new negative sample data.

In the original training dataset, the data ratio of positive and negative samples is 505:161 (505 samples have a label of 0). Through the above data augmentation methods, the data is expanded and the data set is balanced.

3.2 Summary Generation

The task requires giving a question and possibly correct answers to determine whether the answer is correct or incorrect. We should also consider short introductions to the question topic rather than directly using the question and answer fields of the sample data. For legal texts, the same question will have different answers in different contexts, and the differences in the answers are often huge.

We plan to concatenate the explanation and the question together to form the text for the first input system and the answer as the text for the second input. We choose Legal-Bert to handle up to 512 tokens, while most of the training data have more than 512 tokens, and the distribution of sample lengths in the training data set is shown in figure 2. We tried different truncation methods (direct truncation, sliding window truncation) to improve the performance of the model and finally found that using GPT3.5 to generate a summary of the context can achieve a better result than truncation processing.

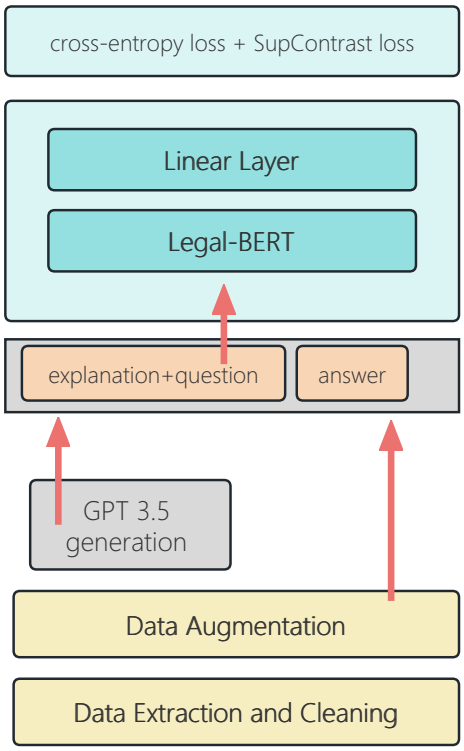

Figure 1: The overall framework of our system proposed for SemEval-2024 Task 5.

The specific treatment we adopt in direct and sliding window truncation is as follows. In direct truncation, we used the explanation and the question field 'l' space. Then, after the mosaics of the

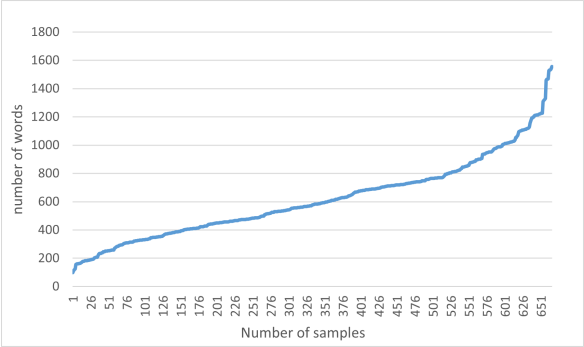

Figure 2: Sample length in the training dataset

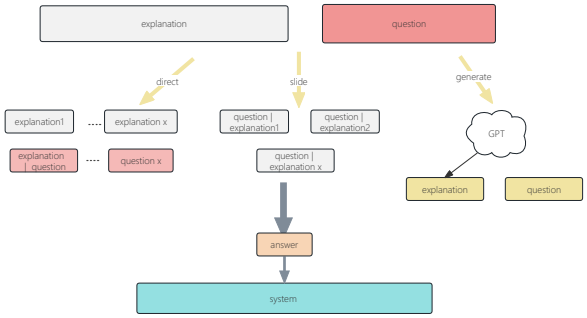

Figure 3: Explanation processing method

strings, separated by spaces counting more than 150 words, as a new sample data, the "id complete" field is used to identify the segmentation. In the sliding window, the basic strategy is the same as the above. Still, in each segmentation, the question's existence is guaranteed, and the part of the 150 words minus the question is explained on the concatenation. The specific process is shown in the figure 3.

However, we found their shortcomings in the above two processing methods. Directly truncating the simple truncated data will lead to the information in the question field with some sample data, either only the context or only the original question information. In sliding window truncation, although the original question field is preserved, we believe that the key information of the explanation is not uniformly distributed in the sentence. Therefore, we adopt GPT3.5 to generate the corresponding summary explanation according to the question pair context.

We believe that the important information to be extracted from the introduction usually involves key sentences, general sentences, and important details, which will affect whether the candidate's answer to the question is correct or not. Abstract generation for introducing a problem uses large models' good generalization ability to extract and compress this general knowledge. This can preserve the integrity of the information and capture the information from a broader perspective than the segmentation method.

3.3 Supervised Contrastive Learning

Contrastive learning aims to learn a data representation by maximizing the similarity between relevant samples and minimizing the similarity between irrelevant samples. In order to better fit this classification task, we use the Supervised Contrastive Learning strategy (Khosla et al., 2020; Chen et al., 2020), in which points belonging to the same class are pulled together in their own space. In contrast, points belonging to different classes are separated.

In a batch input, we treat the samples containing the original answer field as anchors, the newly added complete analysis of the answer as positive samples, and the remaining samples under the same question as negative samples. The contrastive loss under this triplet is shown in Eq 1, which is:

$$L_A = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(f(x_i),f(x_i^+))}}{e^{s(f(x_i),f(x_i^+))}+\sum_{j=1}^{N}e^{s(f(x_i),f(x_j^-))}}$$

where x_i is the input anchor, x_i^+ is the positive sample, x_i^- is the negative sample, $s(f(x_i), f(x_i^+))$ is the similarity measure function, and the inner product is commonly used.

We want to evaluate from an overall point of view, so we combine the cross-entropy loss and contrastive loss as the loss function of the model to train, and the loss function of the model is shown in Eq 2:

$$L = \frac{1}{2}\cdot(L_{CE}(y_i,\hat{y}_i) + L_A(x_i, x_i^+, x_i^-))$$

Where y_i is the ground truth, $\hat{y}_i$ is the predicted value, and x_i, x_i^+, x_i^- are the anchors, positive samples, and negative samples in the upper segment. Each data in the dataset has y_i and $\hat{y}_i$ after training, but only one kind of sample corresponds to the contrastive loss.

4 Experimental setup

4.1 Dataset Description

The training and validation sets contain 666 and 84 samples, respectively. Each sample contains a question, answer, label, analysis(excerpt from complete analysis relevant to answer candidate), complete

analysis(Glannon's explanation for the solution of the question), and explanation(topical introduction, additional context for question, potentially empty) fields. The test set contains 98 examples and has only question, answer, and explanation fields.

The task purpose is, given a question with a likely correct answer and a short introduction to the question topic, to determine whether the answer candidate is correct or incorrect. Each of these sample data does not exist independently, and most of them are 4 to 6 samples in the same group. This means that the questions and contexts of these four data are consistent, and the answers and analyses are different. Specific examples are shown in Table 2.

The following are the specific available fields and what they represent for the samples in the dataset:

- <question> 6. Any port in a storm. Cullen, a Vermont citizen, has an accident with Barnabas, a citizen of California, and Tecumseh, a New Yorker, in California. She sues Barnabas and Tecumseh for negligence in state court in Albany, New York, alleging negligence. She serves Barnabas with process in the ...
- <answer> a motion to transfer the case to a California court under 28 U.S.C. §1404(a).
- <analysis> A isn't right either. Section 1404(a) is a federal statute, authorizing a federal court to transfer a case to another federal court. It does not govern the state courts. There is no transfer statute allowing state courts in one state to transfer cases to ...
- <complete analysis> This question provides a nice little recap of various jurisdiction and venue issues. Barnabas wants out of the New York state court. What motion is likely to do the trick? Removal seems like an option, though of course he'd still have to litigate in New York. Remember that you can only ...
- <explanation> So, venue is the "third ring" in choosing a proper court, along with personal jurisdiction and subject matter jurisdiction. If all three rings are satisfied, the court has the power to hear the case. However, it doesn't always do so. Sometimes a case is filed in a court that has subject matter jurisdiction over the case, personal jurisdiction over the defendant, and is a proper venue under ...

Column	Train	Dev	Test
idx	true	true	true
question	true	true	true
answer	true	true	true
label	true	true	false
analysis	true	true	false
complete analysis	true	true	false
explanation	true	true	true

Table 2: Components of the dataset.

- <label> False

4.2 Dataset Split

We split the processed training set and validation data set into 10 subsets without intersection and randomly split them into units of the same background-size, which ensures that each set has the same proportion of positive and negative samples as the original full set. Ten-fold cross-validation is used, and the results are shown as averages to ensure that the strategy used is maximally effective on the final test set.

4.3 Pre-processing

The legal data in all datasets were provided to us by email by the task organizers. After getting the original file in CSV format, we remove the file headers and re-add the file headers based on data splitting or summarization. After the initial processing of the data, we split the data into a mini-batch of 8 according to the needs of contrastive learning, where the first data is the anchor, the second data is the positive example, and the third to six data are the negative examples. In the cleaning process, we mainly remove some dirty format data, such as some missing field data.

4.4 Evaluation Metrics

Task 5 has two evaluation metrics which are F1 score and precision, The F1 score is common in evaluating binary classification tasks, especially when the classes are imbalanced, it is more representative than precision or recall. The F1 score can range from 0 to 1, with values closer to 1 indicating better performance.

4.5 Others

Hyperparameter tuning was not a critical point of our work. Still, we tested several values over a small range as they did influence our decisions

System	F1 score
pratice augmentation	
Baseline	42.69
+ DA	46.96
evaluation augmentation	
Baseline	42.96
+ DA	50.33
+ Summary Generation	53.59
+ SupContrast Learning	**55.10**

Table 3: Best results with training methods we used.

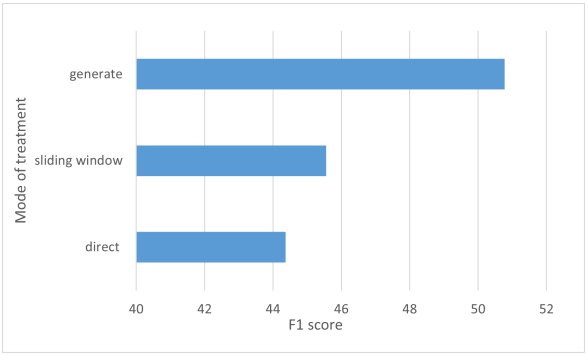

Figure 4: Summary generation effect comparison

about how well the policy worked (see Appendix). In addition, to help the reader replicate our experiments, details of tools and libraries are provided (see Appendix).

5 Results

5.1 Overall Performance

Finally, according to the official scoring system, our system got 0.551 on the test set and ranked 14th. As results are shown in Table 3, all the strategies presented in Section 3 produced positive effects, and we discuss the effects of these strategies one by one in the following subsections. For convenience, all the results from our experiments are multiplied by 100.

5.2 Data Augmentation

To verify whether the augmented dataset plays a positive role, we train with the augmented dataset in the Practice phase of the competition, which provides the official baseline, and this is the gap between the two baselines in Table 3.

As you can see from the top of Table 3, there is a significant increase, which is consistent with our inference that a richer training set is beneficial to build a more accurate system, and the way we augment the data is to some extent a multi-perspective supplement to the original data (from the analysis of the problem).

5.3 Summary Generation

As introduced in Section 3.2, we are aware of the importance of the corresponding explanation of the question. We propose several different ways to include segmentation fields and generate summaries. However, we are not sure which method is effective in collecting the characteristics of the data. Therefore, we tried each method, and the results are shown in Figure 4. Compared with direct truncation, the sliding window truncation method has an improvement of about 1 point, and the generated summary can be improved by about 4 points on this basis.

Obviously, through comparison, it is found that compared with direct truncation and sliding window truncation, the context summary generated by using a large model can better represent the features of the data. By comparing the direct truncation method and the sliding window truncation method, it can also be seen that the effect of the sliding window is better than the direct truncation to a certain extent, which conforms to our basic cognition that explanation is crucial in problem reasoning. Whether a candidate answer to a question is correct or not depends on the context of the question, that is, the relevant introduction.

5.4 Supervised Contrastive Learning

As mentioned in Section 3.3, contrastive learning is incorporated into our system. The loss function of our system is composed of a combination of cross-entropy loss and contrastive loss. We show the output of the cross-entropy loss and contrastive loss in some epochs of training and find that the contrastive learning function values are larger than the cross-entropy loss, and their magnitude is usually about double.

Through our final experimental results, as shown in the table, we can find that after the addition of contrastive learning, our system can learn more general features by reducing the distance between positive examples and away from the distance between negative examples, which increases the adversarial robustness of the model.

6 Conclusion

By deploying various optimization methods, including data augmentation, summary generation,

and supervised contrastive learning, we build a conceivably powerful system to reason about the task of legal argumentation in civil litigation. And ranked 14th in the evaluation stage competition with a 0.551 F1 score in the officially organized competition.

In future work, one is that law is a serious domain, and we plan to guide the model by prior knowledge. We also plan to incorporate domain-specific knowledge into the exercises and analyses of the law school textbooks under study. Second, we consider whether we can better model long texts by using tools external to the model to assist in processing long texts and optimizing the model.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Comput. Sci.*, 2:e93.

Leonard Bongard, Lena Held, and Ivan Habernal. 2022. The Legal Argument Reasoning Task in Civil Procedure. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207, Abu Dhabi, UAE. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis*, Nicolas Garneau*, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. LeX-Files and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.

Daniele Licari and Giovanni Comandè. 2022. ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. In *Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management*, volume 3256 of *CEUR Workshop Proceedings*, Bozen-Bolzano, Italy. CEUR. ISSN: 1613-0073.

Maria Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial decisions of the european court of human rights: looking into the crystall ball. In *Proceedings of the Conference on Empirical Legal Studies in Europe 2018*.

Ramesh Nallapati and Christopher D. Manning. 2008. Legal docket classification: Where machine learning stumbles. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 438–446, Honolulu, Hawaii. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

Table 4 and Table 5 provide the details of the corresponding hyperparameters and libraries.

Hyperparameter	Range/Value
Epoch	30 - 50
Batch Size	8
Warm-up-nums	10
Learning Rate	3e-5∼5e-5

Table 4: Main hyper-parameters tuned in our system.

Tools & Libraries	Version
NumPy	1.22.3
pandas	1.4.0
Python	3.7.10
PyTorch	1.13.0
Transformers	4.15.0

Table 5: Main tools and libraries used in our system.