Law Retrieval with Supervised Contrastive Learning Using the Hierarchical Structure of Law

Jungmin Choi[1,2], **Ukyo Honda**[1,2], **Taro Watanabe**[1], **Kentaro Inui**[2,3], **Hiroki Ouchi**[1,2]
[1] Nara Institute of Science and Technology
[2] RIKEN
[3] Tohoku University
choi.jungmin.ce6@is.naist.jp, honda.ukyo.hn6@is.naist.jp,
taro@is.naist.jp, inui@ecei.tohoku.ac.jp, hiroki.ouchi@is.naist.jp

Abstract

We study the information retrieval task to identify the relevant law articles for a query on a legal issue in when the legal system in question is statute law. In recent years, the mainstream approach has been to calculate the similarity between the query and each article using pre-trained language models. However, such methods have a weakness in retrieving relevant articles that have low n-gram similarity scores with the query. In this work, we show that in such hard cases, the articles tend to be of the same class as articles with high n-gram similarity scores in the hierarchical structure of statute law, for instance, the Japanese Civil Code. From this observation, we hypothesize that by making articles of same class close to each other in the feature space, we could make it easier to retrieve the above mentioned hard articles. Our proposed method realizes this by supervised contrastive learning using the hierarchical structure. Experimental results show that the proposed method achieves higher performance in retrieving the correct articles with low n-gram similarity to the query.

1 Introduction

Law is one of the domains where application of natural language processing is expected to bring immense benefit to the society. According to a survey conducted by Japan Federation of Bar Associations[1], nearly half of those who visited law firms or legal support centers for consultation answered that they had hesitated consulting lawyers before visiting. Some of the most frequently cited reasons include unapproacheable image of lawyers, anticipated difficulty communicating with them, and concerns about whether the issues will be taken seriously. This indicates that people often have psychological obstacles to accessing legal services. It is an important societal task to lower the barrier to accessing legal services and facilitate function of law throughout the society.

A solution to this problem by natural language processing is to build an information retrieval system that suggests relevant laws to user-given queries regarding their legal issues. It will provide the user with an approximate idea about how their issues could be described in legal terms, which will enable them to further search for more refined information without necessarily having to consult legal professionals or ask better informed questions when they seek legal support.

When developing such a system, it is essential to take into account the characteristics of the legal system of interest. Depending on the legal system, laws are written in vastly different style and structure. In this regard, legal systems can be categorized into two broad categories, case law and statute law. Case law, which includes the legal systems of the United Kingdom, United States, Canada, etc., is law that is based on past judicial decisions, while statute law, examples of which are the legal systems of Germany, France, Japan, etc., is written law passed by a body of legislature. With case law, the task in question would be to retrieve relevant cases, and whereas

[1] https://www.nichibenren.or.jp/library/ja/jfba_info/publication/data/shimin_needs.pdf

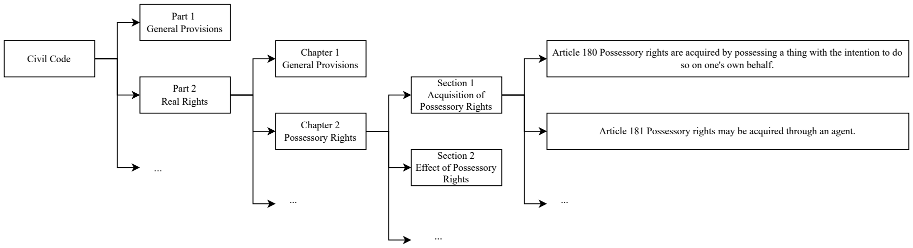

Figure 1: Hierarchical structure of Japanese Civil Code.

with statute law, it would be to retrieve relevant law articles. Note that a case document is written on actual legal disputes and therefore concrete in nature, and a law article is written in abstract legal terms and are organized in hierarchical structure.

In this paper, we focus on statute law. Our task here is to retrieve relevant articles to a given query from a set of candidate articles. As mentioned above, the language in which statute law is written is significantly different from ordinary language. For our purpose, which is to assist users who might not be familiar with those abstract legal terms, it is particularly important that our system can correctly retrieve relevant articles when the query has little to no overlap in terms of vocabulary with the relevant articles.

In this task, the mainstream approach has been to employ pretrained language models (Wehnert et al., 2021; Nguyen et al., 2021; Shao et al., 2021). However, it has been pointed out that these approaches perform poorly when detecting relevance between the query and article is semantically involved (Rabelo et al., 2021).

To address this problem we propose a method to use the hierarchical structure of statute law, which is not fully utilized in previous work. Statute law is typically organized in hierarchical structure. For example, articles in the Japanese civil code are classified on five levels: part, chapter, section, subsection, and division. See Figure 1. At each level, articles in the same class share the same topic. Within the framework of retrieving articles according to their similarity to the query in the embedding space, we hypothesize that by training the model to map articles in a same class closer in the embedding space and articles in different classes apart, we would be able to leverage the structure information and obtain better embeddings. This point will be elaborated in Section 5, Supervised Contrastive Learning for Law Retrieval.

We have conducted experiments using as the benchmark the data set created for Task 3 of the workshop, Competition on Legal Information Extraction/Entailment (COLIEE), where participants compete building systems that automatically answer Japanese bar exam problems. The experimental results show that our approach outperforms previous approaches which supports the effectiveness of contrastive learning as a way to incorporate hierarchical structure in the embeddings.

2 Related Work

2.1 Document Retrieval

The task of retrieving relevant documents to queries has been a central component of information extraction and question answering (Narasimhan et al., 2016; Kwok et al., 2001; Voorhees, 2001). While early research on this task has focused on sparse bag-of-words representations, recent advances in computational resources has inspired a plethora of research using neural network (Mitra and Craswell, 2017).

In general, neural models for document retrieval use vector representations of text, and contain a large number of parameters to be tuned. Therefore, they typically require a large training data set.

To mitigate the computational burden, a body of recent work has adopted a two-stage retrieval and ranking pipeline. At the first stage, they retrieve a large number of documents using sparse high dimensional query/document representations, and at the second stage, they rerank the documents with learned neural models (Nogueira and Cho, 2019; Yang et al., 2019). While this approach has achieved state-of-the-art results on information retrieval benchmarks, it suffers from the upper bound imposed by any recall errors in the first-stage retrieval model (Luan et al., 2020).

A strong alternative is to perform first-stage retrieval using learned dense low-dimensional encodings of queries and documents. Reimers and Gurevych (2019) has shown tha their dual encoder model which scores each document by the inner product between its encoding and that of the query perform well and efficiently. Karpukhin et al. (2020) outperformed traditional sparse vector space models such as TF-IDF and BM25 for retrieving answers from open-domain context by a simple dual-encoder framework.

2.2 Legal Information Retrieval

2.2.1 Case Retrieval

Searching through a large collection of previous cases (court decisions) for ones that apply to a particular situation is an important part of day-to-day work of legal professionals. Hence, there have been efforts to automate this task in jurisdictions from all over the world (Hafner, 1980; Parikh et al., 2021; Xiao et al., 2019; Rabelo et al., 2021; Chalkidis et al., 2020).

While this task can be formulated as a special case of document retrieval, it has been noted that document retrieval methods that perform well with general data sets do not transfer easily to the legal domain, for reasons such as the large number of candidate documents, verboseness of each case documents, the definition of relevance in the legal scenario being beyond the general definition of topical relevance (Shao et al., 2020; Alberts et al., 2021; Van Opijnen and Santos, 2017). Ma et al. (2021) has applied traditional language model and showed that it outperformed neural models. Rosa et al. (2021) has shown that their method of splitting case documents into segments and applying BM25 to rank cases by similarity to the query perform competitively against neural models.

2.2.2 Statute Retrieval

In recent years, the mainstream approach has been to use TF-IDF and pretrained language models.

Wehnert et al. (2021) computes the cosine similarity of each query-article pair based on Sentence-BERT (Reimers and Gurevych, 2019) representation and TF-IDF, sum the two cosine similarity values, and classify the pair as relevant if the value exceeds the predetermined threshold. If none of the articles has similarity higher than the threshold, the article with the highest similarity is selected as the relevant article. The threshold is determined so that the score will be highest if applied to the validation set. They use the English version, and employ a Sentence-BERT model pretrained with millions of paraphrase pairs. The problem with this approach is that it is hard to grasp the similarity between a query and article when it requires high-level semantic matching, e.g., when the query involves concrete examples of abstract concepts in the relevant article.

Nguyen et al. (2021) treats this task as a binary classification problem. They concatenate the query and the article with a SEP token to make a single sequence, applies linear transformation to the BERT features corresponding to the CLS token of this sequence and conduct binary classification whether or not the query and article are relevant. In order to mitigate the label imbalance problem, i.e., irrelevant query-article pairs far outnumbering relevant ones, they only use pairs whose article places in the top 150 among all articles in terms of TF-IDF similarity with the query. They use the Japanese version. Like Wehnert et al. (2021), it has difficulty in semantic matching. It also suffers from a strict upper bound imposed by any recall errors in the first stage where they limit the candidate to top 150.

2.3 Supervised Contrastive Learning

Contrastive learning is a method which aims to obtain effective representation of objects by gathering semantically similar ones in closer proximity in the embedding space and distancing dissimilar ones. Khosla et al. (2021) has introduced supervised contrastive learning for image classification task where they trained the model so that images that belong to same classes will have closer embeddings and those with different classes distant embeddings. Gao et al. (2021) applies such framework to sentence representations, and proposes simple contrastive learning method with a supervised setting. Using natural language inference (NLI) data sets, the method learns

Table 1: Number of Queries by Number of Relevant Articles

Number of Relevant Articles	1	2	3	4	5	6	Total
Number of Queries	567	125	25	5	2	1	725

to bring sentences with entailment relations closer together as positive example pairs and to move sentences with contradiction relations away from each other as negative example pairs. The method significantly outperforms previous methods on a variety of semantic textual similarity tasks.

3 Task

3.1 Overview

We focus on COLIEE Task 3 as our benchmark for the legal information retrieval task. In the data set, query statements regarding legal issues are each paired with sets of relevant law articles. The query statements mostly describe specific situations and tends to be written with ordinary vocabulary, while articles are written using abstract, legal terms, which makes it an appropriate benchmark for a system that identifies relevant articles to queries written by non-experts as described above.

The data consists of train and test data, and there the original version written in Japanese and a translated version in English. Every year, the latest bar examination problems become the test data, and the past problems, including the previous test data, form the train data. For the competition in 2021, the training data consisted of 725 examples of queries and corresponding sets of relevant articles, while the test data had 81 such examples. Table 1 shows the distribution of number of relevant articles to each query. Figure 2 is an example from the data.

3.2 Definition

Formally, the task is defined as follows. There is a set of N_A Japanese Civil Code articles $\mathbb{A} = \{a_1, a_2, \dots a_{N_A}\}$ and a set of N_Q queries $\mathbb{Q} = \{q_1, q_2, \dots, q_{N_Q}\}$. $\mathcal{D} = \{(q_i, \mathbb{A}_i)\}_{i=1}^{N_Q}$ is a set of pairs where each query q_i is paired with the set of its relevant articles $\mathbb{A}_i \subset \mathbb{A}$. Given a query q_i, the task is to find $\mathbb{A}_i$. It is assumed that for each q_i, exactly one such non-empty subset of $\mathbb{A}$ exists.

Query In cases where an individual rescues another person from getting hit by a car by pushing that person out of the way, causing the person’s luxury kimono to get dirty, the rescuer does not have to compensate damages for the kimono.

Relevant Article Article 698 If a manager engages in benevolent intervention in another’s business in order to allow a principal to escape imminent danger to the principal’s person, reputation, or property, the manager is not liable to compensate for damage resulting from this unless the manager has acted in bad faith or with gross negligence..

Figure 2: An example of Task 3. Note that here, ”getting hit by a car” is a concrete example of ”imminent danger” and ”kimono to get dirty” is that of ”damage resulting from this”

Table 2: R@10 scores of relevance prediction to query-article pairs by Wehnert et al. (2021). All represents the whole validation set, Easy represents Easy is the subset of All with higher n-gram similarity, Hard represents the subset with lower n-gram similarity. We show the recall score for each set.

Easy	Hard	All
97.30	42.67	60.71

4 Problems and Analysis of Previous Work

4.1 Problems

In previous studies, when the n-gram similarity between the query and the relevant article is low, it is difficult to classify the article as correct (Rabelo et al., 2021). This is the case, for example, when the query is a description of concrete facts, and it is necessary to correspond the norms of the article to the facts in order to determine the relevance of the relevant article to the query.

4.2 Analysis

In order to confirm this problem statistically, the validation data set was divided according to n-gram similarity, and the scores of each division is shown below. In the validation data set, there are a total of 112 pairs of query and relevant article. We apply TF-IDF vectorization to each query and article, and compute the cosine similarity of these vectors.

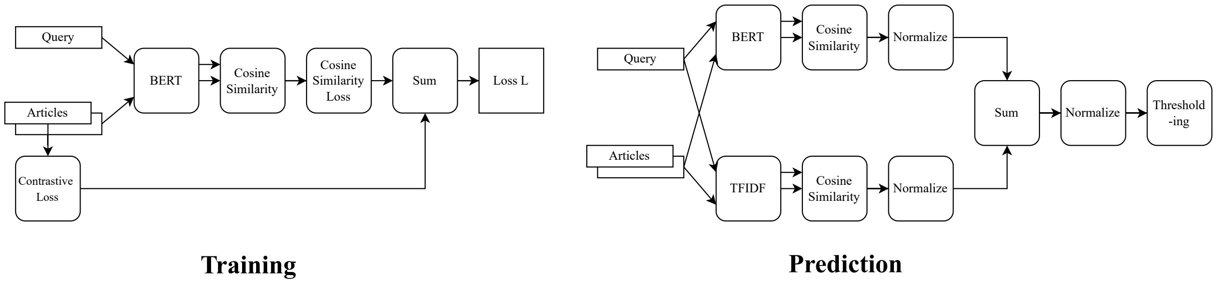

Figure 3: Overview of our proposed method.

There are 75 pairs where the normalized cosine similarity[2] between the query and the relevant article is less than 1. We call such cases Hard, and other cases Easy. The entire cases are called All. Table 2 show the accuracy score for each division from the model that replicates (Wehnert et al., 2021). Note that it is particularly low for Hard.

5 Supervised Contrastive Learning for Law Retrieval

Analyzing the 75 query-article pairs (q_i, a_j) with cosine similarity less than 1, we have found that there are 26 of cases where the article a_k with the highest cosine similarity to the query q_i and the correct article a_j are in the same section. This suggests that even a correct article a_j with low n-gram similarity can be pulled up to the top of the search results if the expressions of articles belonging to the same sections are made closer to each other. Based on this hypothesis, we propose contrastive learning using the section information of articles as labels. The proposed method uses the section information because almost every article belongs to some section and each section has on average approximately 11.3 articles, which makes it an appropriate way to divide articles into semantically similar groups.

Based on this analysis, we expect that it may be easier to obtain relevant articles with low n-gram similarity to the query by putting articles belonging to the same section closer to each other in the embedding space. Therefore, we propose a contrast learning method using sections as labels.

Previous methods have included hierarchical information (e.g., Part 1, Chapter2, Section3) in the input sentences, however, it is unlikely that these methods have successfully incorporated the hierarchical structure. Intuitively, simply including the hierarchical information in the input will not directly incentivize the model to pull same-section peers together, and might even hurt learning by forcing some tokens at the end of the input to be cut off because of sequence length limit of the model. Indeed, in our preliminary experiments, we found no advantage of including hierarchical information in the input compared to not including. We expect that by training the model to embed articles that are close to each other in the hierarchical structure also close to each other in the embedding space, we can more effectively incorporate hierarchical structure and therefore improve performance.

5.1 Overview of Model

In our proposed model, we compute two loss functions which we call basic and contrastive, respectively. We employ as the baseline a model which is similar to Wehnert et al. (2021) [3] except we perform fine-tuning using the COLIEE training set by cosine similarity loss. We sum the cosine similarity loss and the contrastive loss, introduced by the proposed model, and the sum is our final objective function. The overview of the proposed method is shown in Figure 3. During training, the model converts the query and article into BERT representations and cal-

[2] We normalize the values of cosine similarity so that for each query, the pair of this query and the most similar article has similarity value of 1, and the pair of this query and the least similar article 0.

[3] Strictly speaking, our baseline also differs from Wehnert et al. (2021) as they append to the articles commentaries obtained by web crawling and the queries that are entailed by the articles, whereas we omit them in this study because the performance without fine-tuning is almost the same as what is reported in Wehnert et al. (2021) if we do not include them

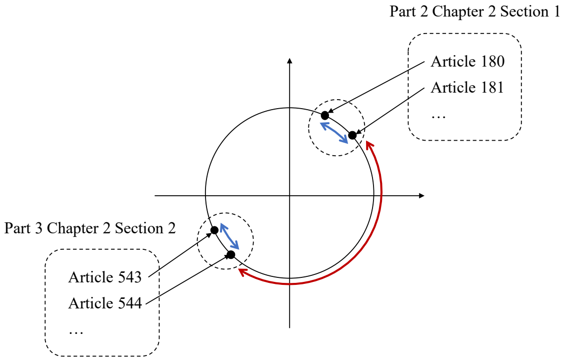

Figure 4: Idea of contrastive learning using hierarchical structure of law. The articles in the same section are pulled closer (blue arrows) while the those in different sections are pushed away (red arrow).

culates the cosine similarity loss from their cosine similarity. On the other hand, the contrastive loss is computed to bring the articles in the batch that belong to the same section closer together. The sum of the two losses is learned to be minimized. When making inference, cosine similarity scores between the query and the article are calculated using BERT and TF-IDF, which are each normalized. We sum them together, and then further normalize the scores to determine if they are relevant by comparing them to the threshold value.

5.2 Loss Function

We give the definitions of the loss functions below. First, for each query-article pair, $\text{pair}_k = (q_{k_q}, a_{k_a})$, we define the binary label indicating relevance of the pair l_k as

$$l_k = \begin{cases} 1 & \text{if } a_{k_a} \in \mathbb{A}_{k_q} \\ 0 & \text{otherwise} \end{cases} . \tag{1}$$

The query q_{k_q} and the article a_{k_a} are converted into features and denoted by $\mathbf{q_{k_q}} = \text{SBERT}(q_{k_q}) \in \mathbb{R}^d$ and $\mathbf{a_{k_a}} = \text{SBERT}(a_{k_a}) \in \mathbb{R}^d$, respectively, where $\text{SBERT}(x)$ is the feature corresponding to the CLS token when x is encoded by sentence-BERT, and d is the number of dimensions of the final layer of sentence-BERT.

5.2.1 Cosine Similarity Loss

Given N randomly sampled triplets of query representation, article representation, and the binary label indicating relevance of the two, cosine similarity loss is computed as follows:

$$\mathcal{L}^{cos} = \frac{1}{N} \sum_{k=1}^{N} \mathcal{L}_k^{cos}, \tag{2}$$

where

$$\mathcal{L}_k^{cos} = \begin{cases} 1 - \cos(\mathbf{q}_{k_q}, \mathbf{a}_{k_a}) & \text{if } l_k = 1 \\ \max(\cos(\mathbf{q}_{k_q}, \mathbf{a}_{k_a}), 0) & \text{otherwise} \end{cases} \tag{3}$$

Minimizing this loss function means penalizing positive cases for a lower similarity, and negative cases for a higher similarity.

5.2.2 Contrastive Loss

Figure 4 describes our idea for contrastive learning. Contrastive learning is performed with the pair of relevant article and an article in the same section as the relevant article as the positive example and the relevant article and an article randomly selected from all the relevant article as the negative example. The contrastive loss is calculated as follows. First, the set of representations of articles in the batch

$$\mathbb{A}_{\text{BATCH}}^{\text{EMB}} = \{\mathbf{a}_{b_1}, \mathbf{a}_{b_2}, \ldots, \mathbf{a}_{b_N}\}$$

is partitioned into groups

$$\mathbb{A}_{\text{BATCH}}^{\text{EMB}} = \bigcup_i \mathbb{S}_i$$

by section that they belong to. That is,

$$\mathbb{S}_i := \{\mathbf{a} \in \mathbb{A}_{\text{BATCH}}^{\text{EMB}} \mid \text{section}(\mathbf{a}) = i\},$$

where $\text{section}(\mathbf{a}) = i$ means article $\mathbf{a}$ belongs to section i.

Then, we construct triplets for contrastive learning. In summary, we (i) generate all possible pairs of articles that belong to the same section within the batch; then, (ii) to each pair, append a randomly selected negative example, which is an article from a different section, to form a triplet. Below is a more rigorous description. We let

$$\mathbb{C}_i := \{(\mathbf{a}_s, \mathbf{a}_t) \mid \mathbf{a}_s, \mathbf{a}_t \in \mathbb{S}_i; s < t\}$$

denote the set of all possible pairs of elements in $\mathbb{S}_i$. Also, let $\mathbb{C} := \bigcup_i \mathbb{C}_i$ and to each pair in $\mathbb{C}$, supply an article randomly chosen from a different section to form a triple, and denote the set of all these triple by $\tilde{\mathbb{C}}$. In other words

$$\tilde{\mathbb{C}} := \{(\mathbf{a}_s, \mathbf{a}_t, \mathbf{a}_{x(s,t)}) \mid (\mathbf{a}_s, \mathbf{a}_t) \in \mathbb{C}\}$$

where $x(s,t)$ is a random variable that takes a value in the set $\mathbb{I} := \{i \mid \mathbf{a}_i \in \mathbb{A}; \mathrm{section}(\mathbf{a}_i) \neq \mathrm{section}(\mathbf{a}_s) = \mathrm{section}(\mathbf{a}_t)\}$ according to the discrete uniform distribution. Hereafter, for simplicity, the notation of the elements of $\tilde{\mathbb{C}}$ will be changed as follows. Let $N_C := |\tilde{\mathbb{C}}|$ and for each $i \in \{1, 2 \ldots, N_C\}$, the i-th element of $\tilde{\mathbb{C}}$ is denoted by $(\mathbf{a}_i, \mathbf{a}_i^+, \mathbf{a}_i^-)$.

Then, the contrastive loss is

$$\mathcal{L}^{cont} = \\ -\log \frac{\mathrm{sim}(\mathbf{a}_i, \mathbf{a}_i^+)}{\sum_{j=1,\neq i}^{N_C} \mathrm{sim}(\mathbf{a}_i, \mathbf{a}_j^+) + \mathrm{sim}(\mathbf{a}_i, \mathbf{a}_j^-)} \quad (4)$$

where $\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \exp(\cos(\mathbf{u}, \mathbf{v})/\tau)$ and τ is the temperature parameter. Intuitively, minimizing this loss function means making the inner product of $\mathbf{a}_i$ with the positive example larger than the inner product with the negative example and other vectors in the batch.

The final loss function of the model is the sum of cosine similarity loss and scaled contrastive loss $\mathcal{L} = \mathcal{L}^{cos} + \alpha\mathcal{L}^{cont}, \alpha \in (0, 1)$.

6 Experiments

As in the COLIEE competition, we employed the recall, precision, F2 values, and the percentage of correct answers in the top ten articles predicted by the model (R@10), as evaluation metrics. In light of the practical goal of improving access to legal information for non-specialists, as mentioned in the introduction, it is more important that the top predictions include possible legal topics to which a query may relate than to present irrelevant topics. Therefore, we pay particular attention to R@10. These evaluation metrics are calculated for each query and averaged over all queries.

6.1 Settings

Wehnert et al. (2021) uses a Sentence-BERT model pretrained with general paraphrase data [4] to make predictions, without fine-tuning it with COLIEE training data. In order to make a fair comparison, we compare our method to a baseline that computes the cosine similarity loss in the same manner as Wehnert et al. (2021) and train it on COLIEE data set before making a prediction.

Preliminary experiments using the validation data showed that the baseline setting tended to improve scores almost steadily up to about the 9th epoch, but not much thereafter. Therefore, we expected that introducing control learning from this point would be effective, and compared (1) a model with 18 epochs of learning in the baseline setting and (2) a model with 9 epochs of learning with the baseline setting and then 9 epochs of learning with the proposed setting.

The number of epochs is 18, batch size is 256, optimization algorithm is Adam (Kingma and Ba, 2015), and the learning rate is $1e-6$, and hyperprameters regarding optimization other than the learning rate are set as recommended by (Kingma and Ba, 2015) for both the baseline and Ours. Learning rate is reset to the initial value every 3 epochs. As for parameters specific to contrastive learning in Ours, $\alpha = 1e-7, \tau = 20$. We report the average of 3 trials along with their standard deviations, in the form of (average value) ± (standard deviation).

6.2 Results

The results of experiments are shown in Table 3. As for F2 and precision, we do not observe a significant difference in the performance of Ours and the baseline. However, Ours performs better than the baseline in terms of R@10 and recall. R@10 and recall are the main focus of our study since our purpose is to build a system that provides non-expert users relevant legal topics to the query. Note that we have compared our method to a baseline stronger than previous state-of-the-art, so that we can differentiate the effectiveness of contrastive learning from that of the plain fine-tuning. The results have shown a solid evidence that contrastive learning is effective.

6.3 Discussion

We test the hypothesis that when a relevant article with a high similarity to the query and an relevant article with a low similarity to it are in the same section, then bringing the representations of these arti-

[4] sentence-transformers/paraphrase-distilroberta-base-v1

Table 3: Comparison with the baseline. Wehnert et al. (2021) indicates the scores reported in the paper. (Wehnert et al., 2021)† is replication by us. The scores of the baseline and Ours is the average of three trials and standard deviation.

model	F2	Rec.	Pre.	R@10
(Wehnert et al., 2021)	73.02	77.78	**67.49**	81.20
(Wehnert et al., 2021)†	72.25	**82.09**	48.81	83.33
Baseline	74.87 ± 0.87	79.22 ± 1.20	61.47 ± 2.22	87.04 ± 0.00
Baseline + Contrastive (Ours)	**75.53** ± 0.79	80.66 ± 0.71	60.26 ± 1.82	**88.48** ± 0.71

Table 4: R@10 for each division by TF-IDF similarity

model	Hard R@10	Hard-Anchored R@10	Easy R@10
(Wehnert et al., 2021)†	43.18	45.00	**100.00**
Baseline	56.82 ± 0.00	70.00 ± 0.00	**100.00** ± 0.00
Baseline + Contrastive (Ours)	**60.61** ± 1.31	**78.33** ± 2.89	**100.00** ± 0.00

cles closer together is effective in improving performance.

In the test data, there are 101 pairs of a query and relevant article, of which 44 are hard and 57 are easy. 20 of the hard articles are in the same section as the article that has the highest cosine similarity with the query (called hard-anchored). R@10 for each division is shown in Table 4.

In Hard-Anchored, the advantage of Ours over the baseline is especially pronounced. This explains why contrastive learning is effective. That is, a relevant article with a low similarity with the query becomes more similar to the query by being brought closer to a relevant article with a high similarity with the query.

Figure 5 shows hard-anchored cases that could not be obtained with the baseline method in the top 10 prediction but could be with the proposed method. The article called ”Relevant Article” is the one which became obtainable and ”Anchor Article” is its same-section peer with high n-gram similarity to the query; the n-grams with underline highly overlap between the Query and Anchor Article.

On the other hand, the cases which could not be obtained either by the baseline or proposed method involved complicated coreference resolution and hypernym detection. Examples are shown in Figure 6. This implies that a limitation of our method is that it still has not overcome the difficulty of capturing the correspondence between general concepts and specific examples, which we shall consider in future work. We saw no case where the baseline method successfully obtained a relevant article in its top 10 prediction but the proposed method failed. This indicates that the proposed method has achieved improvement in some hard cases without any sacrifice in terms of R@10.

7 Conclusion

In this study, to address the difficulty with previous work to classifying relevant articles with low n-gram similarity to the query as relevant in a legal information retrieval task, we focused on the fact that many such relevant articles are hierarchically close to relevant articles with high n-gram similarity to the query, and proposed supervised contrast learning using hierarchical information. Experimental results show that the proposed method outperforms previous methods, especially in classifying articles with low n-gram similarity as correct answers.

Acknowledgements

We thank the organizers of the COLIEE for constructing this dataset.

References

Houda Alberts, Akin Ipek, Roderick Lucas, and Phillip Wozny. 2021. Coliee 2020: Legal information retrieval and entailment with legal embeddings and boosting. In Naoaki Okazaki, Katsutoshi Yada, Ken Satoh, and Koji Mineshima, editors, *New Frontiers in*

Artificial Intelligence, pages 211–225, Cham. Springer International Publishing.

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, November. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*.

Carole D. Hafner. 1980. Representation of knowledge in a legal information retrieval system. In *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*, SIGIR '80, page 139–153, GBR. Butterworth amp; Co.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November. Association for Computational Linguistics.

Prannay Khosla, Teterwak Piotr, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning. In *NeurIPS*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Cody Kwok, Oren Etzioni, and Daniel S. Weld. 2001. Scaling question answering to the web. *ACM Trans. Inf. Syst.*, 19(3):242–262, jul.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2020. Sparse, dense, and attentional representations for text retrieval.

Yixiao Ma, Yunqiu Shao, Bulou Liu, Min Zhang, and Shaoping Ma. 2021. Retrieving legal cases from a large-scale candidate corpus. In *ICAIL/COLIEE*.

Bhaskar Mitra and Nick Craswell. 2017. Neural models for information retrieval.

Karthik Narasimhan, Adam Yala, and Regina Barzilay. 2016. Improving information extraction by acquiring external evidence with reinforcement learning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2355–2365, Austin, Texas, November. Association for Computational Linguistics.

Ha-Thanh Nguyen, Phuong Nguyen, Thi-Hai-Yen Vuong, Quan Bui, Chau Nguyen, Binh Dang, Vu Tran, Minh Nguyen, and Ken Satoh. 2021. Jnlp team: Deep learning approaches for legal processing tasks in coliee 2021. In *ICAIL/COLIEE*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert.

Vedant Parikh, Upal Bhattacharya, Parth Mehta, Ayan Bandyopadhyay, Paheli Bhattacharya, Kripa Ghosh, Saptarshi Ghosh, Arindam Pal, Arnab Bhattacharya, and Prasenjit Majumder. 2021. Aila 2021: Shared task on artificial intelligence for legal assistance. In *Forum for Information Retrieval Evaluation*, FIRE 2021, page 12–15, New York, NY, USA. Association for Computing Machinery.

Juliano Rabelo, Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, and Ken Satoh. 2021. Summary of the competition on legal information extraction/entailment (coliee) 2021. In *ICAIL/COLIEE*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP*.

Guilherme Rosa, Ruan Rodrigues, Roberto Lotufo, and Rodrigo Nogueira. 2021. Yes, bm25 is a strong baseline for legal case retrieval.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3501–3507. International Joint Conferences on Artificial Intelligence Organization, 7. Main track.

Hsuan-Lei Shao, Yi-Chia Chen, and Sieh-Chuen Huang. 2021. Bert-based ensemble model for statute law retrieval and legal information entailment. In Naoaki Okazaki, Katsutoshi Yada, Ken Satoh, and Koji Mineshima, editors, *New Frontiers in Artificial Intelligence*, pages 226–239.

Marc Van Opijnen and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. *Artif. Intell. Law*, 25(1):65–87, mar.

Ellen M. Voorhees. 2001. The trec question answering track. *Natural Language Engineering*, 7(4):361–378.

Sabine Wehnert, Viju Sudhi, Shipra Dureja, Libin Kutty, Saijal Shahania, and Ernesto W. De Luca. 2021. Legal norm retrieval with variations of the bert model combined with tf-idf vectorization. In *ICAIL/COLIEE*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2019. Cail2019-scm: A dataset of similar case matching in legal domain.

Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Simple applications of bert for ad hoc document retrieval.

Query If the contract of sale stipulates that F, who was not born at the time of the conclusion of the contract, is to acquire the ownership of X, the contract of sale is invalid.

Relevant Article Article 537 (1) If one of the parties promises in a contract to render a certain performance to a third party, the third party has the right to claim that performance directly from the obligor. ... (Omitted)

Anchor Article Article 548-4 (1) In the following cases, a preparer of the standard terms of contract may, by amending the standard terms of contract, modify the terms of the contract without making separate agreements with each of the counterparties and deem that the parties have agreed to the amended provisions of the standard terms of contract: ... (Omitted)

Query If a third-party collateral provider paid a secured claim, the third-party collateral provider may exercise the secured claim acquired through subrogation without requirement for perfection.

Relevant Article Article 500 The provisions of Article 467 apply mutatis mutandis in the case referred to in the preceding Article (unless a person with a legitimate interest in making performance is subrogated to the claim of the obligee).

Anchor Article
Article 501 (1) A person that is subrogated ... (Omitted) ... (i) a third party acquirer (meaning a person that has acquired from the obligor the property that is the subject of security; hereinafter the same applies in this paragraph) is not subrogated to the claim of the obligee in relation to any guarantors or third-party collateral providers;... (Omitted)

Figure 5: Hard-Anchored cases which became obtainable by contrastive learning

Query A took the jewelry that B had forgotten, believing without negligence that it belonged to A. In this case, A may not obtain the ownership of the jewelry by good faith acquisition.

Relevant Article Article 192 A person that commences the possession of movables peacefully and openly by a transactional act acquires the rights that are exercised with respect to the movables immediately if the person possesses it in good faith and without negligence.

Query If D owes C a debt (Y) of 300000 yen that is set off against the debt (X), and D demands payment of 600000 yen from A while C does not use a set-off for the debt (Y), A may refuse to pay 200000 yen of that debt.

Relevant Article Article 439 (1) If one of the joint and several obligors has a claim against the obligee and invokes a set-off, the claim is extinguished for the benefit of all joint and several obligors. (2) Until the joint and several obligor that has the claim referred to in the preceding paragraph invokes a set-off, other joint and several obligors may refuse to perform the obligation to the obligee only to the extent of that joint and several obligor's share of the obligation.

Figure 6: Hard case which is not obtainable either by baseline or proposed method