Enhancing Pre-Trained Language Models with Sentence Position Embeddings for Rhetorical Roles Recognition in Legal Opinions

Anas Belfathi[1], Nicolas Hernandez[1,*] and Laura Monceaux[1]

[1]*Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France*

Abstract

The legal domain is a vast and complex field that involves a considerable amount of text analysis, including laws, legal arguments, and legal opinions. Legal practitioners must analyze these texts to understand legal cases, research legal precedents, and prepare legal documents. The size of legal opinions continues to grow, making it increasingly challenging to develop a model that can accurately predict the rhetorical roles of legal opinions given their complexity and diversity. In this research paper, we propose a novel model architecture for automatically predicting rhetorical roles using pre-trained language models (PLMs) enhanced with knowledge of sentence position information within a document. Based on an annotated corpus from the LegalEval@SemEval2023 competition, we demonstrate that our approach requires fewer parameters, resulting in lower computational costs when compared to complex architectures employing a hierarchical model in a global-context, yet it achieves great performance. Moreover, we show that adding more attention to a hierarchical model based only on BERT in the local-context, along with incorporating sentence position information, enhances the results.

Keywords

Pre-trained language models, Discourse structure modeling, Legal Opinions, Sentence Positional Embeddings, Rhetorical Role, Sequence Labelling

1. Introduction

Pre-trained language models, such as BERT [1] and GPT3 [2], have shown significant improvements in performance across various Natural Language Processing (NLP) tasks. However, when it comes to apply these models to specific domains like legal documents, unique challenges arise. Legal documents are often lengthy and without explicit structure, requiring the identification of coherent parts, known as Rhetorical Roles (RRs) for tasks such as summarization, information extraction, and legal reasoning [3, 4, 5].

In this research work, we are interested in the task of rhetorical role prediction in legal judgements. In that context, examples of RRs are: PREAMBLE (meta-data related to the legal judgment document), FACTS (chronology of events that led to filing the case), RPC (final decisions ruled by the present court), etc. In particular, we work with the dataset provided by the organizers of the SemEval 2023 LegalEval competition for the rhetorical role prediction task[1]. For this task, Hierarchical Sequential Labelling Network (HSLN) [6, 4, 7, 8] and Pre-trained Language Models (PLM) which are able to handle long

word sequence [9, 10, 11] are prefered. However, simple text statistics on the data show that a text contains on average 4346.07 sub-words[2] ($\pm$2151.08) and therefore exceeds the maximum input length any Pre-trained Language Models can handle. In addition, the best current system does not exceed 87% of F1-score which justifies an interest in the task [12].

Our main contributions are the following:

- we enhance the pre-trained language model BERT with sentence position information at input;

- we study the sentence position information under various representations (absolute, normalized and K–quantile);

- we consider two architectures to contextualize the sentence representations: 1) a single BERT encoder and 2) a hierarchical model made of a BERT encoder layer to encode sentences and a shallow encoder (with two Transformer layers) to contextualize a sentence with its surrounding sentences;

- we evaluate these various models in the context of a rhetorical role sequence labelling task for legal judgments.

Our related work section (Section 2) covers various topics from the fusion of discourse information in language models to some rhetorical role prediction system

✉ anas.belfathi@edu.univ-paris13.fr (A. Belfathi);
nicolas.hernandez@univ-nantes.fr (N. Hernandez);
laura.monceaux@univ-nantes.fr (L. Monceaux)

[1]https://sites.google.com/view/legaleval.

[2]Computed with the BERT tokenizer.

architectures including a position embedding presentation. Then, in Section 3, we detail the models we propose. In Section 4, we present the methodology we use for our fine-tuning evaluation. Finally, we present our result and discuss our future work in the last sections.

Our code will be publicly available on MASKED_URL.

2. Related work

Injecting discourse information in language models Legal texts share linguistic characteristics specific to the legal domain (and often to a legal sub-domain). They have legal jargon, long sentences, unusual word order and long length [13, 14, 8, 15]. These characteristics do not allow to take full advantage of the state-of-the-art language models trained on the general domain and even show their limitations since most of them cannot handle a text length which goes beyond their maximum input length.

Transformers [16] suffer from a quadratic computational and memory complexity with respect to the sequence length. This lead most of the SOTA models (e.g. BERT [1], RoBERTa [17], LegalBERT [14]) to adopt 512 as their maximum sequence length.

Pre-training or retraining with discourse-based objectives can result in sentence and text representation that are more adapted to addressing NLP tasks at the discourse level [18, 19, 20] but this kind of approach does not address directly the limitation of the input length.

In terms of neural architecture adaptation, Hierarchical Attention Networks (HANs) have been proposed to model a sequence of sentences by stacking two layers of encoders: one to capture the word sequences and another (taking the former as input) to capture the sentence sequence. This architecture has been shown to perform significantly better than single layer encoders for text classification [21], text segmentation [22], recommendation [20] and sequence labelling tasks [8]. To avoid an important number of padded words in the first layer, [20] proposed to concatenate as many as natural sentences the input block can fit. Although extending the scope of single transformer encoder, the HAN architecture does not solve the complexity problems and has to process fixed length of sentence sequence and truncate too long documents. Recent architectures such as Longformer [9], BigBird [10] or Ernie [11] succeeded in extending to 4096 the maximum sequence length while reducing the Transformer's complexity[3] by introducing sparsity into attention layers (i.e. by allowing each token position to attend to a subset of token positions with respect to some sparse patterns). To improve such models, [15] suggested to consider the actual logical structure of documents.

Apart from these more complex architectures, [22] showed that, for a text segmentation task, looking at the local context around each candidate break (by taking the end of the previous sentences and the beginning of the following as input) is sufficient to obtain comparable performance to the HAN architectures.

Rhetorical Role Prediction The analysis of the rhetorical text[4] structure includes several tasks [24, 25, 26, 8, 12]: 1) text segmentation into text rhetorical units, 2) rhetorical role identification of each text unit, 3) structure prediction, which links the text units together and 4) relations labelling to name the connections. We focus here in the task of rhetorical role identification which has been considered in the literature as a sequence labelling task taking the sentence as a minimal text unit [12].

As reported by [12], the Hierarchical Sequential Labelling Network (HSLN) remains the most efficient architecture for this task (at least for the LegalEval dataset) [6, 4, 7, 8]. The model first encodes sentence representations (e.g. by using sent2vec [27] or any PLM like BERT or LegalBERT). Then it contextualizes the sentence representations (e.g. through a BiLSTM layer) and eventually predicts the label sequence thanks to a sequence labelling layer (like a CRF). On the LegalEval dataset, best performance were obtained by participants who used domain-adaptation techniques like a pre-train language model trained on Legal text, or augmented datasets. The baseline model were based on the HSLN architecture and had a performance of 79% F1 score. The proposed methods show an improvement over the baseline without ever succeeding to outperform by more than seven points. [28] showed that a single BERT can be sufficient to capture contextual dependencies without the need for hierarchical contextual encoding neither a CRF sequence labeler. The approach uses BERT to encode a concatenation of sentences (fixed at 10 sentences) and use a MLP over each encoded sentence separator token to predict the corresponding sentence label. Despite its low complexity, the approach is limited by the length of the input sequence and requires to tile the whole text to obtain the whole label sequence. [8] did not confirm the effectiveness of the method on the LegalEval dataset.

Position embeddings By nature, the Transformer architecture is not sensitive to the order of input tokens. To make the model position-aware, the position information of the input words is typically added as an additional embedding to the input token embeddings [16]. While absolute sinusoidal position encodings were utilized in the vanilla Transformer, some works showed that learned position embeddings can provide more flexibility in adapting to different tasks through back-propagation [1, 29], instead of using hand-crafted position represen-

---

[3]See [23] for a survey of techniques to address Transformer's limitations.

[4]We do not make a distinction here between the intra- and inter-sentencial levels.
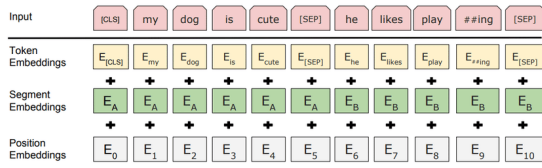
Figure 1: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings[1].

Figure 2: Average sentence position variation by label on the dataset.

tations [30, 31]. Multiple works explored also different token position information (absolute, relative) and ways to include it in Transformers (e.g. in the input or the attention matrix) [30, 32]. Very few were interested in sentence position information. [20] indicate to fuse sentence block representations with sentence block position embeddings but without mentioning precisely the nature of the position (relative, absolute...) and how the position table is built.

BERT Input Representation BERT is one of the models that utilizes three types of learned embedding layers (See Figure 1): Token Embeddings, Segment Embeddings, and Position Embeddings [1]:

- **Token Embeddings:** This layer is responsible for converting each word in the input text into a fixed-dimensional vector representation. In the BERT Base model, each word is represented as a 768-dimensional vector.

- **Segment Embeddings:** This layer has the task of distinguishing between the inputs in a given pair by assigning one of two vector representations to each token in the input.

- **Position Embeddings:** BERT takes into account the sequential nature of input sequences by learning a vector representation for each token position in the input. The Position Embeddings layer is a lookup table of size (512, 768), where each row represents the vector representation of a word at a specific position in the input sequence.

The representations from Token Embeddings, Segment Embeddings, and Position Embeddings are summed element-wise to produce a single representation with shape (1, n, 768), where *n* is the length of the input sequence. This combined representation captures contextual information of tokens in the input text. Although BERT has shown effective results for many tasks such as question answering and sentiment analysis of tweets, it is non-performant when working with lengthy documents. However, it does not incorporate any information about the position of sentences within a document, which can
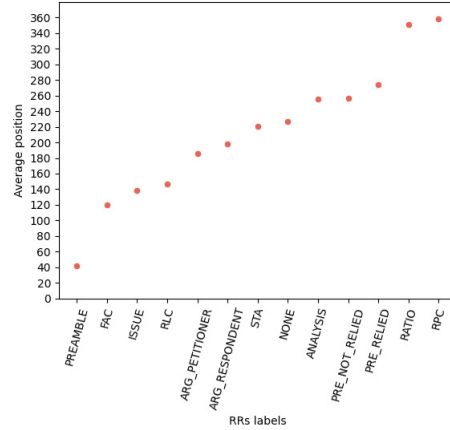
be crucial for identifying the RR of a particular sentence. To address this limitation, we propose the addition of sentence position embeddings to the BERT embedding, which aims to enhance the performance of RR prediction in legal opinions.

3. Fusion sentence position embeddings at input

In this section, we focus on the approach that we have developed for injecting discursive knowledge without pre-training through input embeddings. The use of input embeddings has been a popular approach in natural language processing (NLP) for representing text data in a high-dimensional vector space. By incorporating discursive knowledge into these embeddings, we aim to improve the performance of NLP models without the need for pre-training.

3.1. BERT-SentPos: BERT sentence encoder enhanced with sentence position embeddings at input

In the analysis we have conducted on legal documents, we have observed that each rhetorical role has a specific position within the document (See Figure 2). For example, the preamble role is found at the beginning, the analysis role in the middle of the document, and the RPC role at the end. To improve our model's performance, we decided to incorporate additional information that indicates the position of each sentence in the document. We achieved this by adding an extra embedding layer to the BERT embedding, which helped us capture the sequential nature of positions in vector space. Various Position

Embeddings (PEs) have been proposed in Transformer based architectures [16] to "capture the sequential nature of positions in vector space." These PEs range from fixed ad-hoc ways to fully learnable ones. BERT, in particular, uses fully learnable PEs. In the interest of simplicity, we decided to reuse the learned PEs to represent the sentence Position Embeddings. By doing so, we were able to create a more accurate and effective model for analyzing legal documents.

In this research, we employed various techniques to examine the positioning of sentences within legal documents. Specifically, we explored three different ways of analyzing sentence positions, including absolute position, normalized position, and k-quantile position :

Absolute position refers to the location of a sentence within a particular document in relation to other sentences in the same document. *For instance*, we may have a document that includes:

· Sentence 1: "The court hereby orders the defendant to appear for a hearing on Monday."
· Sentence 2: "The defendant shall provide all relevant documents to the plaintiff's attorney by Friday."

In this case, Sentence 1 has an absolute position of being the first sentence in the document, while Sentence 2 has an absolute position of being the second sentence.

Normalized position refers to the process of converting the position of a sentence in one document to a corresponding position in the largest document in a corpus that has the maximum length of sentences. This is done to ensure consistent normalization across different documents, by aligning sentence positions with respect to a common reference point. *For example*, let's consider Document A, which consists of 50 sentences, and Document B, the largest document in the corpus, containing 100 sentences. If we take Sentence 25 from Document A, we can calculate its normalized position as (25 multiplied by 100 divided by 50) = 50. This aligns Sentence 25 with the same relative position in Document B.

K–quantile position The analysis of legal documents often involves addressing variations in document length and unique writing styles of judges. This technique involves dividing the documents into k parts or quantiles to help control the absolute position of each rhetorical role. *For example*: We have an Document X (divided into 4 quantiles):

· Quantile 1: "The court presents the background of the case."
· Quantile 2: "The court discusses relevant legal precedents."
· Quantile 3: "The court analyzes the evidence presented by both parties."
· Quantile 4: "The court renders its final decision and issues the judgment."

By dividing the opinion into quantiles, each containing a specific role or aspect, it helps in addressing variations
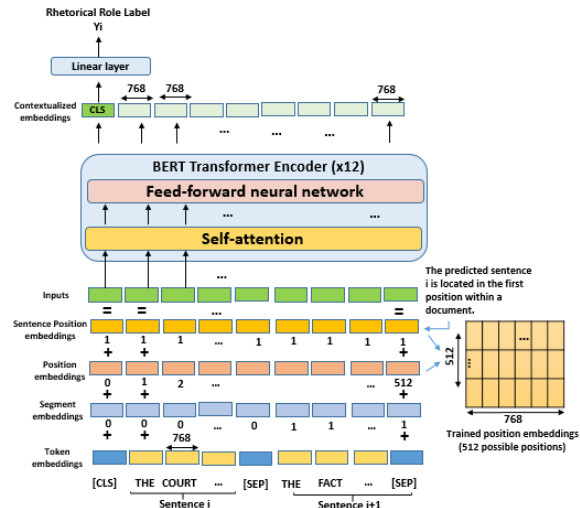

Figure 3: BERT-SentPos: A BERT Architecture with Sentence Position Embeddings fused at the BERT input layer using the Learnable Position Embedding of Tokens.

in document length and unique writing styles of judges.

Figure 3 illustrates our proposed architecture, which utilizes the BERT encoder. We input pairs of consecutive sentences denoted as sentence i and sentence $i + 1$, along with the corresponding sentence position of predicted sentence i, to enable our model to learn contextualized representations that consider the relationships between neighboring sentences. To capture positional information, we added BERT embeddings with sentence position embeddings. These embeddings provide the model with position information of each sentence within the document. The combined representation is passed through encoders and feed-forward layers for predicting the rhetorical role of sentence i. At the end we obtained three models: BERT enhanced with Absolute Sentence Position Embeddings (BERT-AbsPos), BERT enhanced with 8-Quantile Sentence Position (BERT-8-QuantilePos), BERT enhanced with Normalized Sentence Position (BERT-NorPos).

3.2. HiBERT: Hierarchical variant to contextualize sentence representations

In this section, we present HiBERT, a hierarchical variant of BERT-SentPos described in Section 3.1 (See Figure 4). The model is based on the [14, 33]'s hierarchical model. The model aims to label a sentence according to the sentences that precede and follow at a certain range. Each sentence representation is enhanced with sentence positional information before being encoded by a BERT en-
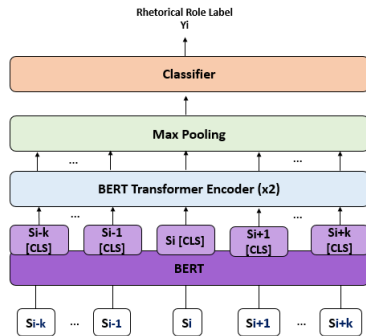
Figure 4: HiBERT: A shallow hierarchical encoder over a BERT-SentPos encoder to capture the sentence sequences coherence. The input layer takes the same input as the BERT-SentPos namely, for each sentence, an aggregation of token, segment, position and sentence position embeddings.

coder. The process generates a top-level representation, denoted as $S_i[CLS]$, for each sentence. These sentence representations are then fed into a 2-layered transformer encoder to contextualize a sentence sequence. This encoder follows the same specifications as BERT, including hidden units and the number of attention heads. Eventually we utilize a dense layer to predict the label of the sentence in focus in the current sentence sequence.

We have set a window of surrounding sentences to a maximum of ± 7 sentences. This range size corresponds to the average number of consecutive sentences with the same RR label in our train dataset. Four variations of HiBERT were experimented: HiBERT-AbsPos, which uses Absolute Sentence Positions with the maximum window size; HiBERT-NorPos, which uses Normalized Sentence Position with the maximum window size; HiBERT-AbsPosHalf, which uses Absolute Sentence Position with half of the maximum window size; and HiBERT-NorPosHalf, which uses Normalized Sentence Positions with half of the maximum window size. Overall, the hierarchical model with more attention attention is an effective solution for processing legal documents that consist of thousands of words. By using a hierarchical approach and taking into account the context of the text, the model is able to effectively process lengthy documents and make accurate classifications based on the document's content.

4. Experimental methodology

We evaluate our contributions in the fine-tuning phase of pre-trained models.

Table 1

Corpus Statistics: The corpus is split into training set, validation set and test set. The table shows number of documents, sentences and the maximum sentence position

Dataset	Docs	Sentences	Max Sentence Position
Train	220	25866	385
Validation	25	2875	348
Test	30	2849	208
Total	275	31590	

4.1. The LegalEval Dataset

We utilized the data supplied by Sub-task A "Rhetorical Roles Prediction" of the SemEval 2023 Task 6 "LegalEval - Understanding Legal Texts" challenge[5]. The dataset comprises Indian legal data extracted from court judgments and includes 13 different RRs, with the details and definitions for each RR provided in the article by Kalamkar et al. [8]. The average number of sentences per document is 117.31.

To prepare the dataset, we kept the same LegalEval split (train and validation data). We used the 90% of the train data to train and the remaining 10% to validate the model. Furthermore, we used the original validation data to evaluate the performance of the trained models. The statistics for splitting the corpus are shown in Table 1.

4.2. Hyperparameters

For the fine-tuning setup, hyperparameters are determined through experimentation and analysis. The batch size is set to 8, taking into consideration the available computational resources and model performance. The learning rate is set to 2e-5, which is a commonly used value for fine-tuning NLP models. The epoch number is chosen from {1, 2, 3}, with the final epoch number selected based on a balance between training time and model performance for each model.

4.3. Performance measures

The performance of the NLP models for the rhetorical roles task is assessed using Weighted-Precision ($_wP$), Weighted-Recall ($_wR$), Accuracy (A), Weighted-F1 ($_wF1$) and Macro F1 ($_MF1$) scores based on the hidden test set. The weighted F1 score considers both precision and recall, and it is calculated by taking into account the class-wise F1 scores weighted by the number of samples in each class. The Macro F1 score provides an overall assessment of model performance by calculating the F1 score for each class independently and then taking the average.

---

[5]https://sites.google.com/view/legaleval.

Table 2
Performance of Models on Test Data

Model	$_wP$	$_wR$	A	$_wF1$	$_MF1$
BERT-HSLN	**0.79**	**0.81**	**0.81**	**0.79**	0.57
BERT	0.69	0.67	0.67	0.65	0.47
BERT-8-QuantilePos	0.73	0.74	0.74	0.72	0.52
BERT-AbsPos	0.75	0.74	0.74	0.74	0.55
BERT-NorPos	0.76	0.77	0.77	0.75	**0.58**
HiBERT-AbsPos	0.76	0.74	0.74	0.75	0.56
HiBERT-NorPos	0.76	0.77	0.77	0.75	0.54
HiBERT-NorPosHalf	0.77	0.76	0.76	0.75	0.53
HiBERT-AbsPosHalf	0.77	0.79	0.79	0.77	0.54

5. Results

In this section, we present the experimental results of our study on injecting different types of sentence position embeddings in the BERT model (See Table 2). As mentioned earlier, we used the weighted F1 score as our primary performance metric. As a baseline, we report the score obtained by the BERT-HSLN[6] model [12] which achieved a performance of 79%. Our first experiment was conducted using BERT with three types of sentence position embeddings: Absolute position (BERT-AbsPos), Normalized position (BERT-NorPos), and K-quantile position(BERT-8-QuantilePos). Our results revealed that BERT performs comparatively poorly (65%) compared to the proposed models when considering three types of position. For the K-quantile position, we experimented with different numbers of parts and found that the best division based on performance is with 8 parts. However, we found that BERT with normalized position achieved a better score of 75%. We attribute this improvement to our efforts in controlling the variation in length across different legal documents.

To further enhance the performance of the recognition of RRs, we experimented with a hierarchical model that combines BERT with more attention and a window size equal to the average of consecutive sentences with the same RR label, while also injecting sentence position information (HiBERT-AbsPos and HiBERT-NorPos). Unfortunately, we did not observe any significant improvement of the results. Subsequently, we halved the average window size of sentences to take into consideration (HiBERT-AbsPosHalf and HiBERT-NorPosHalf). This led to an improvement in the results, particularly with Absolute position by 79%. We attribute this success to the fact that absolute position captures the specific position of a sentence within a document, providing crucial contextual information of the predicted sentence based on its surrounding sentences. This improvement with

---

a smaller context window can be explained by the fact that the semantics of a sentence are usually more dependent on local context than on knowing all sentences in a paragraph, for example.

Overall, our experiments demonstrated that using different types of sentence position information can significantly improve the performance of BERT on legal document classification tasks. Additionally, a hierarchical model that combines BERT with absolute position information and a window size of 4 sentences ($\lceil 7/2 \rceil$) can further enhance the performance of our proposed models. Furthermore, our approach also achieved low computational time, making it efficient and practical for real-world applications.

6. Conclusion and Future Work

The results of our study indicate that the inclusion of absolute, normalized, and k-quantile positional embeddings can significantly improve the performance of both BERT base and its hierarchical variant for the Rhetorical Roles Prediction Task. However, there is a room for improvement. It should be noted, for example, that one potential limitation of this approach is that the number of sentences in a given document may exceed the dimensions of the learnable embedding matrix. The framework proposed by [30] can help us to determine the interest of learning dedicated sentence position embeddings or using sinusoidal PEs. In addition, in the context of token position embeddings, [32] showed that encoding position to attention matrix per-head results in superior performance comparing to adding position embeddings to the input. A similar experiment could be conducted on the sentence position embeddings. Looking ahead, our future work aims to develop a new architecture with a greater number of layers that incorporates additional features such as metadata pertaining to each legal opinion. Furthermore, we plan to explore the potential benefits of pre-training instead of fine-tuning and fine-tuning with LegalBERT [14], a BERT-based model pre-trained on legal text. Our motivation for this direction is based on the fact that current state-of-the-art models may not be sufficient to achieve optimal performance for legal NLP tasks. By incorporating these improvements, we hope to develop a more robust architecture that achieves optimal performance for the Rhetorical Roles Prediction Task and other legal NLP tasks.

Acknowledgments

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[3] M. Saravanan, B. Ravindran, S. Raman, Automatic identification of rhetorical roles using conditional random fields for legal document summarization, in: Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I, 2008.

[4] P. Bhattacharya, S. Paul, K. Ghosh, S. Ghosh, A. Z. Wyner, Identification of rhetorical roles of sentences in indian legal judgments, in: The Thirty-second International Conference on Legal Knowledge and Information Systems (JURIX), volume 322, 2019, p. 3.

[5] V. Malik, R. Sanjay, S. K. Guha, A. Hazarika, S. Nigam, A. Bhattacharya, A. Modi, Semantic segmentation of legal documents via rhetorical roles, in: Proceedings of the Natural Legal Language Processing Workshop 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 153–171. URL: https://aclanthology.org/2022.nllp-1.13.

[6] D. Jin, P. Szolovits, Hierarchical neural networks for sequential sentence classification in medical scientific abstracts, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3100–3109. URL: https://aclanthology.org/D18-1349. doi:10.18653/v1/D18-1349.

[7] A. Brack, A. Hoppe, P. Buschermöhle, R. Ewerth, Cross-domain multi-task learning for sequential sentence classification in research papers, in: Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, JCDL '22, Association for Computing Machinery, New York, NY, USA, 2022. URL: https://doi.org/10.1145/3529372.3530922. doi:10.1145/3529372.3530922.

[8] P. Kalamkar, A. Tiwari, A. Agarwal, S. Karn, S. Gupta, V. Raghavan, A. Modi, Corpus for automatic structuring of legal documents, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 4420–4429. URL: https://aclanthology.org/2022.lrec-1.470.

[9] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, 2020. arXiv:2004.05150.

[10] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, A. Ahmed, Big bird: Transformers for longer sequences, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 17283–17297. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

[11] X. Ouyang, S. Wang, C. Pang, Y. Sun, H. Tian, H. Wu, H. Wang, ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 27–38. URL: https://aclanthology.org/2021.emnlp-main.3. doi:10.18653/v1/2021.emnlp-main.3.

[12] A. Modi, P. Kalamkar, S. Karn, A. Tiwari, A. Joshi, S. K. Tanikella, S. K. Guha, S. Malhan, V. Raghavan, Semeval 2023 task 6: Legaleval - understanding legal texts, 2023. arXiv:2304.09548.

[13] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, M. Sun, How does NLP benefit legal system: A summary of legal artificial intelligence, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5218–5230. URL: https://aclanthology.org/2020.acl-main.466. doi:10.18653/v1/2020.acl-main.466.

[14] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://aclanthology.org/2020.findings-emnlp.261. doi:10.18653/v1/2020.findings-emnlp.261.

[15] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, N. Aletras, LexGLUE: A benchmark dataset for legal language understanding in English, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for

Computational Linguistics, Dublin, Ireland, 2022, pp. 4310–4330. URL: https://aclanthology.org/2022.acl-long.297. doi:10.18653/v1/2022.acl-long.297.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[18] Y. Jernite, S. R. Bowman, D. Sontag, Discourse-based objectives for fast unsupervised sentence representation learning, 2017. arXiv:1705.00557.

[19] D. Iter, K. Guu, L. Lansing, D. Jurafsky, Pretraining with contrastive sentence objectives improves discourse performance of language models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4859–4870. URL: https://aclanthology.org/2020.acl-main.439. doi:10.18653/v1/2020.acl-main.439.

[20] L. Yang, M. Zhang, C. Li, M. Bendersky, M. Najork, Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching, in: Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1725–1734. URL: https://doi.org/10.1145/3340531.3411908. doi:10.1145/3340531.3411908.

[21] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1480–1489. URL: https://aclanthology.org/N16-1174. doi:10.18653/v1/N16-1174.

[22] M. Lukasik, B. Dadachev, K. Papineni, G. Simões, Text segmentation by cross segment attention, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4707–4716. URL: https://aclanthology.org/2020.emnlp-main.380. doi:10.18653/v1/2020.emnlp-main.380.

[23] Q. Fournier, G. M. Caron, D. Aloise, A practical survey on faster and lighter transformers, ACM Comput. Surv. (2023). URL: https://doi.org/10.1145/3586074. doi:10.1145/3586074, just Accepted.

[24] M. Lippi, P. Torroni, Argumentation mining: State of the art and emerging trends, ACM Trans. Internet Technol. 16 (2016). URL: https://doi.org/10.1145/2850417. doi:10.1145/2850417.

[25] H. Yamada, S. Teufel, T. Tokunaga, Building a corpus of legal argumentation in japanese judgement documents: towards structure-based summarisation, Artificial Intelligence and Law 27 (2019) 141–170. URL: https://doi.org/10.1007/s10506-019-09242-3. doi:10.1007/s10506-019-09242-3.

[26] J. W. G. Putra, S. Teufel, T. Tokunaga, Parsing argumentative structure in English-as-foreign-language essays, in: Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Online, 2021, pp. 97–109. URL: https://aclanthology.org/2021.bea-1.10.

[27] M. Pagliardini, P. Gupta, M. Jaggi, Unsupervised learning of sentence embeddings using compositional n-gram features, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 528–540. URL: https://aclanthology.org/N18-1049. doi:10.18653/v1/N18-1049.

[28] A. Cohan, I. Beltagy, D. King, B. Dalvi, D. Weld, Pretrained language models for sequential sentence classification, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3693–3699. URL: https://aclanthology.org/D19-1383. doi:10.18653/v1/D19-1383.

[29] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. N. Dauphin, Convolutional sequence to sequence learning, in: International conference on machine learning, PMLR, 2017, pp. 1243–1252.

[30] B. Wang, L. Shan, C. Lioma, X. Jiang, H. Yang, Q. Liu, J. Simonsen, On position embeddings in bert, 2021, pp. 1–21. 9th International Conference on Learning Representations - ICLR 2021 ; Conference date: 03-05-2021 Through 07-05-2021.

[31] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, AI Open 3 (2022) 111–132. URL: https://www.sciencedirect.com/science/article/pii/S2666651022000146. doi:https://doi.org/10.1016/j.aiopen.2022.10.001.

[32] P.-C. Chen, H. Tsai, S. Bhojanapalli, H. W. Chung,

Y.-W. Chang, C.-S. Ferng, A simple and effective positional encoding for transformers, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 2974–2988. URL: https://aclanthology.org/2021.emnlp-main.236. doi:10.18653/v1/2021.emnlp-main.236.

[33] I. Chalkidis, M. Fergadiotis, D. Tsarapatsanis, N. Aletras, I. Androutsopoulos, P. Malakasiotis, Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 226–241. URL: https://aclanthology.org/2021.naacl-main.22. doi:10.18653/v1/2021.naacl-main.22.