Is Selective Masking A Key to Improving Domain Adaptation for Masked Language Model?

Anas Belfathi, Ygor Gallina, Nicolas Hernandez, Laura Monceaux, Richard Dufour

Is Selective Masking A Key to Improving Domain Adaptation for Masked Language Model?

Anas Belfathi, Ygor Gallina, Nicolas Hernandez, Laura Monceaux, Richard Dufour
firstname.lastname@univ-nantes.fr
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000
Nantes, France

ABSTRACT

Pre-training and fine-tuning language models have been effective for domain adaptation in various NLP tasks. A key aspect of BERT-based models is the word-masking strategy used during training. However, commonly used random word masking may not fully capture domain-specific linguistic nuances. This paper aims to determine whether domain-dependent masking strategies, informed by thematic relevance or genre specificity, would be more effective than the conventional random strategy in different training phases. Our experiments on legal domain benchmarks reveal that selective masking offers no significant advantage over random masking in continual pre-training. However, domain-specific masking proves more effective in from-scratch pre-training, despite all available pre-trained language models being trained using random masking.

CCS CONCEPTS

• **Applied Computing** → **Law**.

KEYWORDS

Language modeling, Masking strategy, BERT, Genre, Tf-Idf

ACM Reference Format:
Anas Belfathi, Ygor Gallina, Nicolas Hernandez, Laura Monceaux, Richard Dufour. 2025. Is Selective Masking A Key to Improving Domain Adaptation for Masked Language Model?. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Among pre-trained language models (PLMs), masked language models (MLMs), such as BERT, stand out for their effectiveness in various natural language processing (NLP) tasks [6, 10, 15]. MLMs are trained using a word-masking strategy, in which words in a sequence are masked and the model must predict them. Traditionally, these words are selected randomly.

In specialized domains, random masking can overlook important domain-specific terms, which are often under-represented in the training data, resulting in suboptimal model performance [7]. Previous research [8, 10, 12, 13, 19] has shown that modifying the masking strategy, known as selective masking, can improve performance.

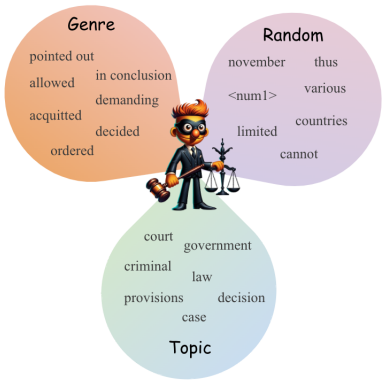

Figure 1: Highly weighted words from our topic and genre-based weighting strategies compared to random selection in the in-domain training corpus.

In the realm of specialized discourse, such as legal texts, genre and topics are crucial factors. As shown in Figure 1, genre refers to the structural and stylistic conventions that define text organization. For example, legal texts consistently use words like *ordered* and *allowed*, reflecting the formal and structured nature of legal discourse (details are provided in Section 3.1). This contrasts with general-purpose texts, which follow broader conventions. Although genre and topic information interact in ways that affect meaning and structure [17], this interaction has not been systematically studied in adapting MLMs like BERT.

In this paper, we address whether *selective masking* remains a key to adapting MLMs by proposing two weighting schemes that prioritize words directly related to topic- and genre-specific words. We also study the importance of selective masking methods in different learning strategies.

Our central hypothesis is that by modifying the masking distribution — that is, by controlling which tokens are masked — we can guide the model to learn more meaningful, domain-specific representations. The empirical results support this: models trained with topic-based masking consistently outperform those using random masking on five out of six evaluation tasks. This demonstrates that even simple modifications to the masking policy can significantly influence model behavior and quality.

To summarize, this paper makes the following contributions:

- We demonstrate that continual pre-training with domain-specific word masking strategies has minimal impact, suggesting that random masking remains a simple and robust baseline for adapting models.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

- We show that training MLMs from scratch using either topic-based or genre-based selective masking is effective, with topic-based masking yielding bigger improvements overall.

We provide open-source models and code for selective masking, facilitating MLM pre-training in specific domains based on our masking strategies[1].

2 RELATED WORK

Various strategies have been explored for selecting masked content, moving beyond random masking. Devlin et al. [6] introduced Whole-Word Masking, where entire words are predicted rather than individual tokens, allowing the model to capture meaningful word representations. Joshi et al. [10] expanded this approach with Span-BERT, masking spans of words instead of single tokens. In Span-BERT, the starting point of the span is randomly selected, and the span size is also varied. This technique improves performance on downstream NLP tasks by better capturing context. Other strategies move away from random selection entirely, such as masking phrases and named entities, as in ERNIE [19], which led to improvements in Chinese NLP tasks. Another variation of ERNIE [23] integrates structured knowledge bases into contextualized representations, which benefits tasks such as entity typing and relation classification.

In domain-specific contexts, Althammer et al. [1] proposed Linguistically Informed Masking for the patent domain, where tokens are masked based on their likelihood of being part of a noun chunk. EntityBERT [14] used a domain-specific pre-trained entity recognizer to mask entities, requiring a unique tagger for each domain but applying the same masking strategy regardless of the domain's context. Golchin et al. [7] improved generalization in downstream tasks by masking keywords using a pre-trained keyword extraction model. However, this approach focused on task-specific adaptation, fine-tuning models for individual tasks rather than creating a model capable of handling multiple tasks within a domain. Difference-Masking [22] dynamically adapts masking strategies using unlabeled task-specific data. It identifies domain-specific features via Tf-Idf weights and applies cosine similarity to determine masking, boosting task-specific performance. However, its adaptability across varied tasks within the same domain is limited, constraining its broader applicability.

In line with previous works, we aim to capture the rhetorical structure that characterizes domain-specific texts by prioritizing words using our proposed statistical-based approaches.

3 PROPOSED SELECTIVE MASKING METHOD

Our selective masking method is divided into two steps: 1) Weighting words by their relevance within a specialized domain (Section 3.1) and 2) Choosing words to mask (Section 3.2).

3.1 Word weighting strategies

We define two types of relevance weighting measures to capture the **topic** and **genre** of in-domain texts.

[1] github.com/ygorg/legal-masking

Topic. To identify words relevant to the document topic, we use the Tf-Idf (Term Frequency-Inverse Document Frequency) metric [16] as a proxy for topical specificity. This metric emphasizes content-bearing words that distinguish documents (e.g., "contract," "asylum," "tax"), aligning with standard information retrieval theory and its application in prior legal NLP work. The weighting score for each term is calculated as follows:

$$s^{topic}(t,d) = tf_{d,t} \cdot \log\left(\frac{1+|D|}{1+df_t}+1\right) \tag{1}$$

$$df_t = |\{d \in D : t \in d\}| \tag{2}$$

where d is a document, t is a term, and D is the corpus of documents. The term frequency $tf_{d,t}$ corresponds to the number of occurrences of t in the document d, while df_t is the number of documents containing the term t.

Genre. Inspired by Hyland [9], who described genre-specific terms as linguistic elements organizing discourse (e.g., adverbials, connectors, and relevance markers), we propose a *genre weight* that measures how representative a word or phrase is of a specific genre. Unlike traditional IDF, which does not account for distribution within documents, our formula captures word distribution consistency to identify genre-specific markers. The weight for each word is calculated as:

$$s^{genre}(t) = \frac{df_t}{tf_t} * \left(1 - \frac{std(dtf_t)}{max(dtf_t)}\right) * \frac{df_t}{D} \tag{3}$$

where df_t denotes the number of documents in which the term t appears, tf_t represents the total number of occurrences of t in the corpus, dtf_t is a list of the number of occurrences of t in every document. Intuitively, the first term in the equation prioritizes words that appear in a large number of documents. The second term evaluates the consistency of a word's occurrence across documents, assigning higher weights to words with uniform frequency distribution, normalized by their maximum frequency. Finally, the third term rewards terms that are prevalent across many documents. Examples of specific genre markers for different tasks can be found in Table 1.

3.2 Word selection

After scoring words using one of our weighting methods (topic or genre), we prepare the masked sequence by randomly sampling words based on their relevance weights until 15% of the sequence tokens are selected, as recommended in previous works [6, 21]. This approach ensures exposure to diverse words during training, allowing the model to learn across both high- and low-frequency words. The detailed algorithm is provided in Algorithm 1.

4 EXPERIMENTAL PROTOCOL

To evaluate our proposed selective masking method, we introduce an in-domain training corpus and evaluation tasks (Section 4.1) and detail the baseline models of our study (Section 4.2).

Table 1: Illustrating high-scoring genre markers across studied downstream tasks

Dataset	Input(s)
ECtHR	12. In 1987 the applicant association (...)
	13. The book was published in the second quarter of 1987 (...)
SCOTUS	329 U.S. 29 67 S.Ct. 1 91 L.Ed. 22
	CHAMPLIN REFINING CO. V. UNITED STATES et al. No. 21. (...)
EUR-LEX	(1) Article 18 of Regulation (EC) No 2038/1999 provides that (...)
	(2) Regulation (EC) No 2038/1999 provides that (...)
LEDGAR	The validity or unenforceability of any -provision or provisions of this
	Agreement shall not (...)
RR	IN THE COURT OF THE V ADDL SESSIONS JUDGE, MYSORE.
	Dated this the 23rd day of May 2013 (...)

Algorithm 1 Selective Masking Algorithm for Domain-Specific Language Models

function MaskTokens($tokens$)
  $\mathcal{M} \leftarrow \emptyset$ ▷ Initialize the mask set
  $W \leftarrow$ WholeWords($tokens$) ▷ Tokenize into whole words
  $S \leftarrow$ ScoreSequence(W) ▷ Assign scores to tokens
  while $|\mathcal{M}| < 0.15 \times |\text{tokens}|$ **do**
    $i \leftarrow$ Sample(S) ▷ Weighted sampling
    Remove($W[i], S[i]$) ▷ Remove from selection pool
    if $|\mathcal{M}| + |W[i]| \leq 0.15 \times |\text{tokens}|$ **then** ▷ If word's token fit in mask set
      $\mathcal{M} \leftarrow \mathcal{M} \cup \{W[i]\}$
    end if
  end while
  return $\mathcal{M}$ ▷ Return the final set of masked tokens
end function

4.1 In-domain corpus and evaluation tasks for legal domain

Our study focuses on the legal domain. To capture the linguistic diversity in legal texts, we use two corpora: opinions from the Supreme Court of the United States (SCOTUS)[2] and a subset of the LexFiles corpus [5], which includes documents from various legal frameworks, such as the European Union and India (see Table 2 for corpus statistics).

Table 2: Details of the in-domain legal dataset

Sub-Corpus	# Doc	# Tokens
EU Case Law	29.8K	178.5M (29%)
ECtHR Case Law	12.5K	78.5M (13%)
Indian Case Law	34.8K	111.6M (19%)
SCOTUS Opinions	104.7K	235.5M (39%)
Total	**181.8K**	**604.1M**

[2]kaggle.com/datasets/gqfiddler/scotus-opinions

For downstream evaluation, we conduct experiments for 6 legal classification tasks:

ECtHR Tasks A & B *[2, 4].* seek to identify articles from the European Convention on Human Rights that are either violated (ECtHR A) or allegedly violated (ECtHR B), based on a set of facts.

SCOTUS *[18].* involves categorizing U.S. Supreme Court opinions into one of 14 thematic areas, such as Criminal Procedure or Economic Activity.

EUR-LEX *[3].* relates to categorizing EU legislation, to assign relevant EUROVOC thesaurus concepts to legislative texts.

LEDGAR *[20].* focuses on classifying paragraphs of US contracts from the EDGAR database into one of 100 topics (Salary, Warranty, Forfeitures, ...).

Rhetorical Roles Labeling (RR) *[11].* involves segmenting Indian legal judgments into semantically coherent segments, each labeled with a rhetorical role, such as preamble, fact, ruling, or argument.

We report the macro F1 score averaged over 5 runs for each task.

4.2 Baseline models

We evaluate our approach against three baselines:

SpanBERT [10]. extends BERT to improve span-level representation and excels in tasks like question-answering. This baseline helps us assess how span-based masking strategies perform on various domain-specific tasks compared to our method.

ERNIE [23]. incorporates knowledge graphs through entity masking, focusing on entity-specific information. This comparison allows us to evaluate the effectiveness of integrating external knowledge in domain adaptation.

BERT-Random [6]. is the widely used random masking strategy in most pre-trained language models and serves as our primary baseline. We evaluate its performance under both continual pre-training and from-scratch settings to assess the impact of prior knowledge on domain-specific tasks. We use the BERT-base uncased model (110M parameters), rather than the large version, as preliminary experiments showed no significant performance difference between the two in our setup.

5 EXPERIMENTS AND RESULTS

We study the impact of selective word-masking strategies on the pre-training of BERT models. Table 3 reports the results for all benchmark tasks. The models BERT-Genre and BERT-Topic represent those trained using genre and topic-specific information, respectively, under both continual pre-training and training-from-scratch conditions.

Comparison with State-of-the-Art masking strategies. We begin by comparing our continual pre-training models with baseline models (Section 4.2), as all models benefit from the prior knowledge embedded in the pre-trained BERT. The results demonstrate that selective word-masking strategies consistently outperform the SpanBERT and ERNIE baseline models. These findings suggest that the span-boundary objective in SpanBERT and the knowledge graph

Table 3: Performance with random and selective masking strategies. Bold marks the highest Macro-F1 score per task. Gray shows the difference from the baseline. † denotes significance over BERT-Random at the 0.01 level

		Legal-Eval	Lex-GLUE					
		RR	ECtHR (A)	ECtHR (B)	SCOTUS	EUR-LEX	LEDGAR	Avg.
SOTA	**SpanBERT [10]**	47.39	51.76	59.90	54.58	54.43	81.78	58.30
	ERNIE [23]	46.31	45.26	52.29	42.35	52.64	80.54	53.23
Cont.	**BERT-Random**	49.32	55.65	61.93	59.13	56.56	**82.31**	60.81
	BERT-Genre	48.62 -0.70	56.15 0.50	63.14 1.21	58.79 -0.34	55.92 -0.64	82.29 -0.02	60.82
	BERT-Topic	**49.41** 0.09	**56.17**† 0.51	**64.14**† 2.21	**60.36** 1.23	**56.91** 0.35	82.24 -0.07	**61.54**
Scratch	**BERT-Random**	44.74	43.33	52.46	42.32	51.99	78.64	52.24
	BERT-Genre	**46.83**† 2.09	41.70 -1.63	50.03 -2.44	44.18 1.87	48.29 -3.7	**78.68** 0.04	51.62
	BERT-Topic	44.66 -0.08	**45.26** 1.93	**57.08**† 4.62	**51.74**† 9.42	**52.74** 0.75	78.64 0.01	**55.02**

integration in ERNIE are less effective at capturing the nuances of domain-specific legal language. Furthermore, these results highlight the potential of word-level information to better capture domain-specific text patterns, emphasizing the need for selective masking over more general strategies.

Importance of the PLMs' prior knowledge in continual pre-training. In the continual pre-training setting, BERT-Topic significantly outperforms the baseline, particularly in tasks such as the ECtHR (A & B) and the SCOTUS classification task, with gains of 0.51%, 2.21%, and 1.23%, respectively, over random masking. These results suggest that emphasizing topic-specific words slightly enhances the model's performance, even after it has already acquired general knowledge through random masking during the initial pre-training phase. In contrast, while BERT-Genre demonstrates positive gains in the ECtHR (A & B) tasks, with improvements of 0.51% and 1.21%, it underperforms in other tasks such as EUR-LEX and Legal-Eval RR. This performance gap is likely due to the original BERT model being pre-trained with random masking on large general-domain data, which may not capture the stylistic conventions or discourse patterns of legal texts, particularly genre-specific markers. This bias limits its adaptation to the legal genre without sufficient genre-focused pre-training data. For the LEDGAR task, neither genre nor topic masking showed notable improvement, likely due to the complexity of distinguishing between 100 classes, making selective masking less effective.

Evaluating the effectiveness of selective masking without the bias of pre-trained models. Given the limited impact and mitigated results of selective masking in the continual pre-training scenario, we then focused on training from scratch, eliminating the bias of prior knowledge embedded in pre-trained models. In this setting, all models are trained on the same data, and differences in performance can be attributed entirely to the masking strategy applied during training. Here, BERT-Topic consistently outperforms other models in 5 out of 6 tasks, with significant improvements particularly in SCOTUS and ECtHR (B), where it achieved gains of 9.42% and 4.62%, respectively. These results highlight the effectiveness of selective topic masking in adapting masked language models (MLMs) across a range of domain-specific tasks.

BERT-Genre also demonstrates notable improvements in the RR and SCOTUS tasks, with gains of 2.09% and 1.87%, respectively. This suggests that genre-specific words, which are indicative of task discourse structure, can be particularly beneficial for certain legal tasks where linguistic structure plays a central role in understanding document intent or categorization.

In contrast to the continual pre-training results, we observed performance declines with BERT-Genre in European legal tasks such as ECtHR (A & B) and EUR-LEX, with respective drops of 1.63%, 2.44%, and 3.7%. This suggests that models may need a stronger foundation in general language skills before specializing in European legal terms. Focusing solely on genre-specific information during training may fall short in capturing the full complexity of the legal language used in European jurisdictions, which combines technical terminology with highly structured legal discourse.

6 CONCLUSION

In this paper, we introduced a novel approach for adapting masked language models to specialized domains through selective masking rather than random word selection. We proposed two masking strategies that leverage genre and topical information to target domain-specific words during pre-training. Our results show that while continual pre-training with random masking remains effective for adapting pre-trained models, selective masking is more advantageous when training new models from scratch. For future work, we aim to investigate the complementary strengths of various state-of-the-art selective masking strategies to further enhance specialized language models. Additionally, we plan to conduct further studies by training models from scratch with large datasets using selective masking, to fully assess its advantages over random masking and to expand this work to other domains such as healthcare.

7 LIMITATIONS

Our study acknowledges limitations that could be addressed in future research. First, our exploration of selective masking strategies is limited to the BERT architecture and English-language data, due to resource constraints. This may affect the generalizability of our findings to other language models, languages, or domains. Second, although we introduced novel specificity scores aimed at capturing

topic and rhetorical structure, their evaluation remains partial. In particular, evaluating the genre-based score is especially challenging, as there are no standard quantitative benchmarks for genre-relevant features. While task-based performance offers indirect evidence of its utility, more targeted analyses — such as measuring correlation with genre-annotated corpora or using stylistic probing tasks — would provide stronger support. In the revised conclusion, we elaborate on this difficulty and outline concrete directions for future experimental validation. Lastly, our results show that the impact of selective masking during pre-training varies considerably across downstream tasks. Understanding why some tasks benefit more than others requires further investigation. Such insights could guide the development of more task-specific or adaptive masking strategies, ultimately improving the efficiency of domain-specific pre-training in legal NLP and other specialized areas.

Acknowledgments

This work was granted access to the HPC resources of IDRIS under the allocations 2023-AD011014882 and 2023-AD011014767, provided by GENCI.

This research was funded, in whole or in part, by l'Agence Nationale de la Recherche (ANR), project ANR-22-CE38-0004.

References

[1] Sophia Althammer, Mark Buckley, Sebastian Hofstätter, and Allan Hanbury. 2021. Linguistically Informed Masking for Representation Learning in the Patent Domain. In *2nd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech2021)*. Association for Computing Machinery, New York, NY, USA. http://ifs.tuwien.ac.at/patentsemtech/2021/fls/2021/3_Althammer.pdf

[2] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 4317–4323. doi:10.18653/v1/P19-1424

[3] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6974–6996. doi:10.18653/v1/2021.emnlp-main.559

[4] Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. 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*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 226–241. doi:10.18653/v1/2021.naacl-main.22

[5] Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 15513–15535. doi:10.18653/v1/2023.acl-long.865

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

[7] Shahriar Golchin, Mihai Surdeanu, Nazgol Tavabi, and Ata Kiapour. 2023. Do not Mask Randomly: Effective Domain-adaptive Pre-training by Masking In-domain Keywords. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, Burcu Can, Maximilian Mozes, Samuel Cahyawijaya, Naomi Saphra, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Chen Zhao, Isabelle Augenstein, Anna Rogers, Kyunghyun Cho, Edward Grefenstette, and Lena Voita (Eds.). Association for Computational Linguistics, Toronto, Canada, 13–21. doi:10.18653/v1/2023.repl4nlp-1.2

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009. https://ieeexplore.ieee.org/document/9879206

[9] Ken Hyland. 1998. Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics* 30, 4 (1998), 437–455. doi:10.1016/S0378-2166(98)00009-5

[10] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics* 8 (2020), 64–77. doi:10.1162/tacl_a_00300

[11] Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022. Corpus for Automatic Structuring of Legal Documents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 4420–4429. https://aclanthology.org/2022.lrec-1.470

[12] Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. 2021. PMI-Masking: Principled masking of correlated spans. In *International Conference on Learning Representations*. https://openreview.net/forum?id=3Aoft6NWFej

[13] Yian Li and Hai Zhao. 2021. Pre-training Universal Language Representation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 5122–5133. doi:10.18653/v1/2021.acl-long.398

[14] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. EntityBERT: Entity-centric Masking Strategy for Model Pretraining for the Clinical Domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii (Eds.). Association for Computational Linguistics, Online, 191–201. doi:10.18653/v1/2021.bionlp-1.21

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[16] Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation* 60, 5 (Oct. 2004), 503–520. doi:10.1108/00220410410560582

[17] Dmitri Roussinov and Serge Sharoff. 2023. BERT Goes Off-Topic: Investigating the Domain Transfer Challenge using Genre Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 468–483. doi:10.18653/v1/2023.findings-emnlp.34

[18] Harold J. Spaeth, Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh. 2023. Supreme Court Database, Version 2023 Release 01. http://supremecourtdatabase.org

[19] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced Representation through Knowledge Integration. arXiv:1904.09223 [cs.CL]

[20] Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 1235–1241. https://aclanthology.org/2020.lrec-1.155

[21] Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should You Mask 15% in Masked Language Modeling?. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2985–3000. doi:10.18653/v1/2023.eacl-main.217

[22] Alex Wilf, Syeda Akter, Leena Mathur, Paul Liang, Sheryl Mathew, Mengrou Shou, Eric Nyberg, and Louis-Philippe Morency. 2023. Difference-Masking: Choosing What to Mask in Continued Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13222–13234. doi:10.18653/v1/2023.findings-emnlp.881

[23] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association

for Computational Linguistics, Florence, Italy, 1441–1451. doi:10.18653/v1/P19- 1139