An Empirical Study on Cross-Lingual and Cross-Domain Transfer for Legal Judgment Prediction

Anonymous ACL submission

Abstract

Cross-lingual transfer learning has proven useful in a variety of NLP tasks, but it is understudied in the context of legal NLP, and not at all on Legal Judgment Prediction (LJP). We explore transfer learning techniques on LJP using the trilingual Swiss-Judgment-Prediction (SJP) dataset, including cases written in three languages (German, French, Italian). We find that Cross-Lingual Transfer (CLT) improves the overall results across languages, especially when we augment the dataset with machine-translated versions of the original documents, using a 3× larger training corpus. Further on, we perform an analysis exploring the effect of cross-domain and cross-regional transfer, i.e., train a model across domains (legal areas), or regions. We find that in both settings (legal areas, origin regions), models trained across all groups perform overall better, while they also have improved results in the worst-case scenarios. Finally, we report improved results when we ambitiously apply cross-jurisdiction transfer, where we augment our dataset with Indian legal cases originally written in English.

1 Introduction

Rapid development in CLT has been achieved by pre-training transformer-based model in large multilingual corpora (Conneau et al., 2020; Xue et al., 2021), where these models have state-of-the-art results in multilingual NLU benchmarks (Ruder et al., 2021). Moreover, adapter-based fine-tuning (Houlsby et al., 2019; Pfeiffer et al., 2020) has been proposed to minimize the disalignment of multilingual knowledge (alignment) when CLT is applied, especially in a zero-shot fashion, where the target language is unseen during training. CLT is severely understudied in legal NLP applications with the exception of Chalkidis et al. (2021) who experimented with several methods for CLT on MultiEURLEX, a newly introduced multilingual legal topic classification dataset, including EU laws.

To the best of our knowledge, CLT has not been applied to the LJP task (Aletras et al., 2016; Xiao et al., 2018; Malik et al., 2021), where the goal is to predict the verdict (court decision) given the facts of a legal case. Following the work of Niklaus et al. (2021), we experiment with their newly released trilingual Swiss-Judgment-Prediction (SJP) dataset, containing cases from the Federal Supreme Court of Switzerland (FSCS), written in three official Swiss languages (German, French, Italian).

The dataset covers four core legal areas (public, penal, civil, and social law) and courts originated in eight regions of Switzerland (Zurich, Ticino, etc.), which poses interesting new challenges on model robustness / fairness and the effect of cross-domain and cross-regional knowledge sharing.

We examine three main research questions: (a) Is cross-lingual transfer beneficial across all or some of the languages?, (b) Do models benefit from cross-domain and cross-regional transfer?, and (c) Can we leverage data from another jurisdiction to improve performance? The contributions of this paper are threefold:

- We explore, for the first time, the application of cross-lingual transfer learning in the challenging task of Legal Judgment Prediction (LJP) in several settings. We find that pre-trained multilingual models trained in a multilingual fashion, outperform their mono-lingual counterparts, especially when we augment the training data with translated versions of the original documents (3× larger training corpus) with larger gains in a low-resource setting (Italian).
- We perform cross-domain and cross-regional analyses exploring the effects of cross-domain (and cross-regional) transfer, i.e., train a model across domains, with respect to the relevant legal areas (e.g., civil, penal law) or regions (e.g., Zurich, Ticino). We find that in both settings (legal areas, regions), models trained across all groups perform overall better and more robustly.

- We also report improved results when we apply cross-jurisdiction transfer, where we further augment our dataset with Indian legal cases originally written in English. The cumulative performance improvement is approx. 7% compared to the best reported scores in Niklaus et al. (2021).

2 Related Work

Legal Judgment Prediction (LJP) is the task, where given the facts of a legal case, a system has to predict the correct outcome (legal judgement). Many prior works experimented with some forms of LJP, however, the precise formulation of the LJP task is non-standard as the jurisdictions and legal frameworks vary. Aletras et al. (2016); Medvedeva et al. (2018); Chalkidis et al. (2019) predict the plausible violation of European Convention of Human Rights (ECHR) articles of the European Court of Human Rights (ECtHR). Xiao et al. (2018, 2021) study Chinese criminal cases where the goal is to predict the ruled duration of prison sentences and/or the relevant law articles.

Another setup is followed by Şulea et al. (2017); Malik et al. (2021); Niklaus et al. (2021), which use cases from Supreme Courts (French, Indian, Swiss, respectively), hearing appeals from lower courts relevant to several fields of law (legal areas). Across tasks (datasets), the goal is to predict the binary verdict of the court (approval or dismissal of the examined appeal) given a textual description of the case. None of these works have explored neither cross-lingual (i.e., models trained in multiple languages), nor cross-jurisdiction transfer, (i.e., from one jurisdiction to another), while the effects of cross-domain and cross-regional transfer are also not studied (analyzed).

Cross-Lingual Transfer (CLT) is a flourishing topic with the application of pre-trained transformer-based models trained in a multilingual setting (Devlin et al., 2019; Lample and Conneau, 2019; Conneau et al., 2020; Xue et al., 2021) excelling in NLU benchmarks (Ruder et al., 2021). Adapter-based fine-tuning (Houlsby et al., 2019; Pfeiffer et al., 2021a) has been proposed as an anti-measure to mitigate disalignment of multilingual knowledge when CLT is applied, especially in a zero-shot fashion, where the target language is unseen during training (or even pre-training).

Meanwhile, CLT is understudied in legal NLP applications. Chalkidis et al. (2021) experiment with standard fine-tuning, while they also examined the use of adapters (Houlsby et al., 2019) for zero-shot CLT on a legal topic classification dataset comprising European Union (EU) laws. They found adapters to achieve the best tradeoff between effectiveness and efficiency. Their work did not examine the use of methods incorporating translated versions of the original documents in any form, i.e., translate train documents or test ones. Other multilingual legal NLP resources (Galassi et al., 2020; Drawzeski et al., 2021) have been recently released, although CLT is not applied in any form.

3 Experiments

3.1 Experimental Set Up

We use the Swiss-Judgment-Prediction (SJP) dataset of Niklaus et al. (2021) containing cases from the FSCS. The dataset is not equally distributed; in fact, there is a notable representation disparity where Italian have far fewer documents (4.2k), compared to German (50k) and French (31k). Representation disparity is also vibrant with respect to legal areas and regions.

Since the dataset contains many documents with more than 512 tokens (90% of the documents are up to 2048), we use hierarchical BERT models (Chalkidis et al., 2019; Niklaus et al., 2021) to encode up to 2048 tokens per document (4×512 blocks). We follow Niklaus et al. and report macro-F1 score. We repeat each experiment with 3 different random seeds and report the average score and standard deviation across runs (seeds). Our code and additional resources will be publicly available.[1]

3.2 Cross-lingual Transfer

We first examine *cross-lingual transfer*, where the goal is to share (transfer) knowledge across languages, and we compare models in three main settings: (a) *Monolingual*: fine-tuned per language, using either the documents originally written in the language, or an augmented training set including the machine-translated versions of all other documents (originally written in another language),[2] (b) *Cross-lingual*: fine-tuned across languages with or without the additional translated versions, and (c) *Zero-shot cross-lingual*: fine-tuned across a subset of the languages excluding a target language at a time. We present the results in Table 1.

[1] Additional details on model configuration, training, and hyperparameter tuning can be found in Appendix A.2.

[2] We use the EasyNMT (`https://github.com/UKPLab/EasyNMT`) library to translate all documents using M2M (Fan et al., 2020). Additional details in Appendix A.3.

Model	**#M**	**de**	**fr**	**it**	**Avg**
A. *Fine-tune on the* **tgt training set** (**src** = **tgt**) — Baselines					
Niklaus et al. (2021)	N	68.5	70.2	57.1	65.2
NativeBERT	N	69.6	**72.0**	68.2	69.9
XLM-R	N	68.2	69.9	59.7	65.9
B. *Fine-tune on the* **tgt training set incl. translations** (**src** = **tgt**)					
NativeBERT	N	70.0	71.0	71.9	71.0
XLM-R	N	68.8	70.7	71.9	70.4
C. *Fine-tune on* **all training sets** (**src** ⊂ **tgt**)					
XLM-R	1	68.9	71.1	68.9	69.7
XLM-R + Adapt	1	66.0	66.3	67.0	66.4
D. *Fine-tune on* **all training sets incl. translations** (**src** ⊂ **tgt**)					
XLM-R	1	**70.2**	71.5	**72.1**	**71.3**
XLM-R + Adapt	1	70.0	70.5	69.8	70.1
E. *Fine-tune on* **all training sets excl. tgt language** (**src** ≠ **tgt**)					
XLM-R	1	58.4	69.1	68.4	65.3
XLM-R + Adapt	1	57.7	64.0	62.6	61.5

Table 1: Test results for all training set-ups (monolingual w/ or w/o translations, multilingual w/ or w/o translations, and zero-shot) w.r.t source (src) and target (tgt) language. Best overall results are in **bold**, and best per setting (group) are underlined. ***The multilingually trained model including translated versions*** (3× ***larger training corpus) have the best overall results***. #M is the number of models trained/used (1, or N=3).

We observe that the baseline *monolingually* pre-trained models (NativeBERT) have the best results compared to the multilingually pre-trained XLM-R (group A – Table 1). Augmenting the original training sets with translated versions of the documents (group B – Table 1), originally written in another language, improves performance in almost all (5/6) cases. Interestingly, the performance improvement in Italian, which has the least documents (less than 1/10 compared to German), is approx. 2%; making Italian the best performing language.

We now turn to the *cross-lingual transfer* setting, where we train XLM-R across all languages. We observe that cross-lingual transfer (group C – Table 1) improves performance across languages compared to the same model (XLM-R), fine-tuned in a monolingual setting. Augmenting the original training sets with the documents translated across all languages, further improves performance (group D – Table 1); translating the full training set provides a 3× larger training set (approx. 150k in total) that equally represents all three languages.

We also present results in a *zero-shot cross-lingual* setting (group E – Table 1), where XLM-R is trained in two languages and evaluated in the third one (unseen in fine-tuning). We observe that German has the worst performance (approx. 10% drop), which can be justified as German is a *Germanic* language, while both French and Italian are *Romance* and share a larger part of the vocabulary. Contrarily, Italian, the low-resource language in our experiments, strongly benefits from cross-lingual transfer, leading to approx. 10% improvement, compared to the monolingual XLM-R. A general negative result of our study is that the use of Adapters negatively affects results across all cross-lingual settings. Concluding, cross-lingual transfer with an augmented dataset comprised of the original and machine-translated versions of all documents, has the best overall performance with a vibrant improvement (approx. 3% compared to the baselines) in Italian, the least represented language.

Legal Area	**#D**	**Public**	**Civil**	**Penal**	**Social**	**All**
Public	15.2k	56.4	52.2	59.7	60.1	57.1
Civil	11.5k	44.4	64.2	45.5	43.6	49.4
Penal	11.8k	40.8	55.8	**84.5**	61.1	60.6
Social	9.7k	52.6	56.6	69.0	70.2	62.1
All (XLM-R)	59.7k	58.0	**67.2**	84.4	70.2	**70.0**
All (Native)	59.7k	**58.1**	64.5	83.0	**71.1**	69.2

Table 2: Test results for models trained per legal area (domain) or across all legal areas (domains). Best overall results are in **bold**, and in-domain are underlined. ***Cross-domain transfer is beneficial for 3 out of 4 legal areas and has the best overall results.*** #D is the number of training examples per legal area.

3.3 Cross-domain/regional Transfer Analysis

3.3.1 Legal Areas

In Table 2 we present the results for *cross-domain* transfer between legal areas. The results on the diagonal (underlined) are in-domain, i.e., fine-tuned and evaluated in the same legal area. Interesting to note is that the best results (**bold**) are achieved in the cross-domain setting, either by using XLM-R or NativeBERT in 3 out of 4 legal areas. Penal law poses the only exception where the domain-specific model outperforms the cross-domain model by a small margin. The shared multilingual model trained across all languages and legal areas (*All* – XLM-R) outperforms the NativeBERT models trained across all legal areas, giving another indication that the performance gains from cross-lingual transfer are robust domain-wise as well.

3.3.2 Origin Regions

In Table 3 we present the results for *cross-regional* transfer. In the top section of the table, we again present the region-specific multilingual models evaluated across regions (in-region on the diagonal, zero-shot otherwise). Surprisingly, in some cases the zero-shot model slightly outperforms the in-

Origin Region	**#D**	**#L**	**ZH**	**ES**	**CS**	**NWS**	**EM**	**RL**	**TI**	**FED**	**All**
Zürich (ZH)	8.8k	de	65.5	65.6	63.7	68.2	62.0	57.9	63.2	54.8	62.6
Eastern Switzerland (ES)	5.7k	de	62.9	66.9	62.8	65.2	62.2	60.2	57.8	55.1	61.6
Central Switzerland (CS)	4.8k	de	62.5	65.5	63.2	65.1	60.7	57.8	60.5	55.9	61.4
Northwestern Switzerland (NWS)	5.7k	de	66.0	68.6	65.2	67.9	61.6	57.0	57.1	55.5	62.4
Espace Mittelland (EM)	8.3k	de,fr	64.1	66.6	63.3	66.7	64.0	66.8	63.2	58.4	64.1
Région Lémanique (RL)	13.4k	fr,de	61.0	64.7	60.2	63.7	63.4	69.8	67.6	54.3	63.1
Ticino (TI)	2.3k	it	55.0	56.3	53.2	54.5	56.0	54.7	66.0	53.1	56.1
Federation (FED)	1.3k	de,fr,it	57.5	59.6	56.8	58.9	55.0	56.5	53.5	54.9	56.6
All (XLM-R)	59.7k	de,fr,it	**69.2**	**72.9**	68.3	**73.3**	**69.9**	71.7	**70.4**	**65.0**	**70.1**
All (Native)	59.7k	de,fr,it	69.0	72.1	**68.6**	72.0	**69.9**	**71.9**	68.8	64.8	69.6

Table 3: Test results for models trained per region or across all regions. Best overall results are in **bold**, and in-domain are underlined. ***Cross-regional transfer is beneficial for all regions and has the best overall results. The shared multilingual model trained across all languages and regions is comparable with the baseline (monolingual BERT models).*** #D is the number of training examples per origin region. #L are the languages covered.

domain model (e.g., NWS to ZH and vice-versa).[3] Similar to cross-domain transfer across legal areas, in cross-regional transfer, cross-lingual transfer is beneficial in 5 out of 8 origin regions. Also, even more audibly, the cross-regional models always outperform region-specific models.

3.4 Cross-Jurisdiction Transfer

We, finally, "ambitiously" stretch the limits of transfer learning in LJP and we apply *cross-jurisdiction* transfer, i.e., use of cases from different legal systems, another form of cross-domain transfer. For this purpose, we further augment the dataset of FSCS cases, with cases from the Supreme Court of India (SCI), published by (Malik et al., 2021).[4] We consider and translate all (approx. 31k) Indian cases ruled up to the last year (2014) of our training dataset, originally written in English, to all target languages (German, French, and Italian).[5]

In Table 4, we present the results for two cross-jurisdiction settings: *zero-shot* (Only MT Indian), where we train XLM-R on the machine-translated version of Indian cases, and *augmented* (+ MT Indian), where we further augment the (already augmented) training set of Swiss cases with the Indian ones. While zero-shot transfer clearly fails; interestingly, we observe improvement for all languages in the augmented setting. This opens a new fascinating direction for LJP research. The cumulative improvement from all applied enhancements adds up to approx. 7% macro-F1 compared to the XLM-R baseline and best-of Niklaus et al. (2021).

Dataset	**#D**	**de**	**fr**	**it**	**Avg**
XLM-R + MT Swiss	59.7k	70.2	71.5	72.1	71.3
+ MT Indian	90.9k	**70.5**	**71.8**	**73.5**	**72.0**
Only MT Indian	31.2k	50.4	47.9	49.5	49.3

Table 4: Test results for cross-jurisdiction transfer in both settings: *zero-shot* (Only MT Indian) and *augmented* (+ MT Indian). Best results are in **bold**. ***Augmenting with Indian cases is overall beneficial.***

Statistical Significance: Using Almost Stochastic Order (ASO) (Dror et al., 2019) with a confidence level $\alpha = 0.05$, we find the score distributions of the core models (NativeBERT, w/ and w/o MT Swiss, XLM-R w/ and w/o Indian MT) stochastically dominant ($\epsilon_{\min} = 0$) over each other in order. Results are presented in Table 5 in Appendix B.

4 Conclusions

We examined the application of CLT in Legal Judgment Prediction for the very first time. We found that a multilingually trained model including translated versions have the best overall results, especially in the low resource setting (Italian). We also examined the effects of cross-domain ((legal areas) and cross-regional transfer, which is overall beneficial in both settings, leading to more robust models. Cross-jurisdiction transfer by further augmenting the training set with machine-translated Indian cases improves overall performance. The cumulative improvement from all applied enhancements adds up to approx. 7% macro-F1.

[3] We consider the distributional similarity (or dissimilarity) w.r.t. legal areas as a plausible explanation for the minor performance differences, but the results in Table 6 in Appendix C does not fully justify all in/out of domain mismatches.

[4] Although SCI rules under the Indian jurisdiction (law), while the FSCS under the Swiss one, we hypothesize that both legal systems, primarily civil-based, share core standards, and thus transferring knowledge could potentially have a positive effect. We discuss this matter in Appendix E.

[5] We do not use the original documents written in English, as English is not part of our target languages.

Ethics Statement

The scope of this work is to study LJP to broaden the discussion and help practitioners to build assisting technology for legal professionals and laypersons. We believe that this is an important application field, where research should be conducted (Tsarapatsanis and Aletras, 2021) to improve legal services and democratize law, while also highlight (inform the audience on) the various multi-aspect shortcomings seeking a responsible and ethical (fair) deployment of legal-oriented technologies.

In this direction, we study how we could better exploit all the available resources (from various languages, domains, regions, or even different jurisdictions). This combination leads to models that improve overall performance -more robust models-, while having improved performance in the worst-case scenarios across many important demographic or legal dimensions (low-resource language, worst performing legal area and region).

Nonetheless, irresponsible use (deployment) of such technology is a plausible risk, as in any other application (e.g., content moderation) and domain (e.g., medical). We believe that similar technologies should only be deployed to assist human experts (legal scholars, or legal professionals).

The examined dataset, Swiss-Judgment-Prediction, released by Niklaus et al. (2021), comprises publicly available cases from the FSCS, where cases are pre-anonymized, i.e., names and other sensitive information are redacted. The same applies for the Indian Legal Documents Corpus (ILDC) of Malik et al. (2021).

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93. Publisher: PeerJ Inc.

Rich Caruana, Steve Lawrence, and C. Giles. 2001. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

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*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

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*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Branden Chan, Timo Möller, Malte Pietsch, Tanay Soni, and Chin Man Yeung. 2019. deepset - Open Sourcing German BERT.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]*. ArXiv: 1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Kasper Drawzeski, Andrea Galassi, Agnieszka Jablonowska, Francesca Lagioia, Marco Lippi, Hans Wolfgang Micklitz, Giovanni Sartor, Giacomo Tagiuri, and Paolo Torroni. 2021. A corpus for multilingual analysis of online terms of service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 1–8, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2785, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Andrea Galassi, Kasper Drazewski, Marco Lippi, and Paolo Torroni. 2020. Cross-lingual annotation projection in legal texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 915–926, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in*

Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. 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)*, pages 4046–4062, Online. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial decisions of the European Court of Human Rights: Looking into the crystal ball. In *Proceedings of the Conference on Empirical Legal Studies*, page 24.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. UmBERTo: an Italian Language Model trained with Whole Word Masking. Original-date: 2020-01-10T09:55:31Z.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021a. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021b. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Dennis Ulmer. 2021. deep-significance: Easy and Better Significance Testing for Deep Neural Networks. Https://github.com/Kaleidophon/deep-significance.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *CoRR*, abs/2105.03887.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. *arXiv:1807.02478 [cs]*. ArXiv: 1807.02478.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the Law Area and Decisions of French Supreme Court Cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria. INCOMA Ltd.

A Details on the Experimental Set Up

A.1 Hierarchical BERT

Since the Swiss-Judgment-Prediction dataset contains many documents with more than 512 tokens (90% of the documents are up to 2048), we use Hierarchical BERT models similar to (Chalkidis et al., 2019; Niklaus et al., 2021) to encode up to 2048 tokens per document (4×512 blocks).

We split the text into consecutive blocks of 512 tokens and feed the first 4 blocks to a shared standard BERT encoder. Then, we then aggregate the block-wise CLS tokens by passing them through another 2-layer transformer encoder, followed by max-pooling and a final classification layer.

We re-use the implementation released by Niklaus et al. (2021),[6] which is based on the Hugging Face Transformers library (Wolf et al., 2020). Notably, we improve the masking of the blocks. Specifically, when the document has less than the maximum number (4) of blocks, we pad with extra sequences of PAD tokens, without the use of special tokens (CLS, SEP), as was previously performed. This minor technical improvement seems to affect the model's performance at large (Table 1).

A.2 Hyperparameter Tuning

We experimented with learning rates in {1e-5, 2e-5, 3e-5, 4e-5, 5e-5} as suggested by Devlin et al. (2019). However, like reported by Mosbach et al. (2020), we also found RoBERTa-based models to exhibit large training instability with learning rate 3e-5, although this learning rate worked well for BERT-based models. 1e-5 worked well enough for all models. To avoid either over- or under-fitting, we use Early Stopping (Caruana et al., 2001) on development data.

We opted to use the standard Adapters of Houlsby et al. (2019), as the language Adapters introduced by Pfeiffer et al. (2020) are more resource-intensive and require further pre-training per language. We tuned the adapter reduction factor in {2×, 4×, 8×, 16×} and got the best results with 2× and 4×; we chose 4× for the final experiments to favor less additional parameters. We tuned the learning rate in {1e-5, 1e-4, 1e-3} and achieved the best results with 1e-5.

We additionally applied label smoothing (Szegedy et al., 2015) on cross-entropy loss. We achieved the best results with a label smoothing factor of 0.1 after tuning with {0, 0.1, 0.2, 0.3}.

We experiment with mono-lingually pre-trained BERT models and XLM-R (approx. 550M parameters) of Conneau et al. (2020), available at https://huggingface.co/models. Specifically, for monolingual experiments (Native BERTs), we use German-BERT (approx. 110M parameters) (Chan et al., 2019) for German, CamemBERT (Martin et al., 2020) (approx. 123M parameters) for French, and UmBERTo (approx. 123M parameters) (Parisi et al., 2020) for Italian, similar to Niklaus et al. (2021). These models are considered the best monolingual models in the respective languages.

A.3 Translating Documents with EasyNMT

We performed the translations using the EasyNMT[7] framework utilizing the *many-to-many* M2M_100_418M model of (Fan et al., 2020), since the OPUS-MT (Tiedemann and Thottingal, 2020) models did not have any model available from French to Italian. A manual check of some translated samples showed sufficient translation quality. However, we noted that a lack of legal-specific sentence splitters negatively affected translation quality. We also expect higher quality translations by designated (legal-oriented) NMT systems.

B Statistical Significance Testing

Using ASO (Dror et al., 2019) with a confidence level $\alpha = 0.05$, we found the score distributions of core models (NativeBERT, NativeBERT + MT Swiss, XLM-R + MT Swiss, XLM-R + MT Swiss/Indian) ranked worst to best are stochastically dominant over each other ($\epsilon_{\min} = 0$). We

[6] https://github.com/JoelNiklaus/SwissJudgementPrediction

[7] https://github.com/UKPLab/EasyNMT

Model Type	M1	M2	M3	M4
M1	1.0	1.0	1.0	1.0
M2	0.0	1.0	1.0	1.0
M3	0.0	0.0	1.0	1.0
M4	0.0	0.0	0.0	1.0

Table 5: Almost stochastic dominance ($\epsilon_{\min} < 0.5$) with ASO. Models are (M1: NativeBERT, M2: NativeBERT + MT, M3: XLM-R + MT, M4: XLM-R + MT Indian

compared all pairs of models based on three random seeds each using ASO with a confidence level of $\alpha = 0.05$ (before adjusting for all pair-wise comparisons using the Bonferroni correction). Almost stochastic dominance ($\epsilon_{\min} < 0.5$) is indicated in Table 5. We use the deep-significance Python library of Ulmer (2021).

C Distances Between Legal Area Distributions per Origin Regions

	ZH	ES	CS	NWS	EM	RL	TI	FED
ZH	.02	.02	.03	.02	.01	.02	.05	.12
ES	.03	.03	.04	.03	.02	.01	.06	.11
CS	.02	.01	.01	.02	.01	.04	.06	.13
NWS	.05	.04	.06	.04	.04	.03	.04	.09
EM	.03	.03	.04	.02	.03	.03	.04	.10
RL	.06	.05	.07	.05	.05	.05	.04	.07
TI	.07	.07	.08	.05	.07	.08	.02	.06
FED	.10	.10	.12	.09	.10	.10	.06	.02

Table 6: Wasserstein distances between the legal area distributions of the training and the test set per origin region across languages. The training sets are in the columns and the test sets in the rows.

In Table 6 we show the Wasserstein distances between the legal area distributions of the training and the test sets per origin region across languages. Unfortunately, this analysis does not explain why the NWS model (zero-shot) outperforms the ZH model (in-domain) on the ZH test set, as found in Table 3.3.2.

D Additional Results

In Tables 7, 8, 9, and 10 we present detailed results for all experiments. All tables include both the average score across repetitions, as reported in the original tables in the main article, but also the standard deviations across repetitions.

Model	**#N**	**de**	**fr**	**it**	**Avg**
Fine-tune on the **tgt training set** (**src** = **tgt**) — Baselines					
Linear (BoW)	N	52.6 ± 0.1	56.6 ± 0.2	53.9 ± 0.6	54.4 ± 0.3
(Niklaus et al., 2021)	N	68.5 ± 1.6	70.2 ± 1.1	57.1 ± 0.4	65.2 ± 0.8
NativeBERT	N	69.6 ± 0.4	**72.0** ± 0.5	68.2 ± 1.3	69.9 ± 1.6
XLM-R	N	68.2 ± 0.3	69.9 ± 1.6	59.7 ± 10.8	65.9 ± 4.5
Fine-tune on the **tgt training set incl. translations** (**src** = **tgt**)					
NativeBERT	N	70.0 ± 0.7	71.0 ± 1.3	71.9 ± 2.5	71.0 ± 0.8
XLM-R	N	68.8 ± 1.4	70.7 ± 2.1	71.9 ± 2.6	70.4 ± 1.3
Fine-tune on **all training sets** (**src** ⊂ **tgt**)					
XLM-R	1	68.9 ± 0.3	71.1 ± 0.3	68.9 ± 1.4	69.7 ± 1.0
XLM-R + Adapt	1	66.0 ± 3.7	66.3 ± 3.3	67.0 ± 2.0	66.4 ± 0.4
Fine-tune on **all training sets incl. translations** (**src** ⊂ **tgt**)					
XLM-R	1	**70.2** ± 0.5	71.5 ± 1.1	**72.1** ± 1.2	**71.3** ± 0.7
XLM-R + Adapt	1	70.0 ± 0.3	70.5 ± 1.0	69.8 ± 0.6	70.2 ± 0.5
Fine-tune on **all training sets excl. tgt language** (**src** ≠ **tgt**)					
XLM-R	1	58.4 ± 1.2	69.1 ± 1.2	68.4 ± 1.1	65.3 ± 4.9
XLM-R + Adapt	1	57.7 ± 1.9	64.0 ± 2.0	62.6 ± 1.2	61.5 ± 2.7

Table 7: Test results for all examined training set-ups w.r.t source (src) and target (tgt) language. Best overall results are in **bold**, and best per setting (group) are underlined. The mean and standard deviation are computed across random seeds.

Area	**Public**	**Civil**	**Penal**	**Social**	**All**
Public	56.4 ± 2.2	52.2 ± 2.0	59.7 ± 4.9	60.1 ± 5.8	57.1 ± 3.2
Civil	44.4 ± 7.9	64.2 ± 0.6	45.5 ± 13.1	43.6 ± 5.2	49.4 ± 8.6
Penal	40.8 ± 10.1	55.8 ± 2.9	**84.5 ± 1.3**	61.1 ± 7.5	60.6 ± 15.7
Social	52.6 ± 4.2	56.6 ± 2.0	69.0 ± 5.5	70.2 ± 2.0	62.1 ± 7.6
All (XLM-R)	58.0 ± 3.0	**67.2 ± 1.6**	84.4 ± 0.2	70.2 ± 1.3	**70.0 ± 9.5**
All (Native)	**58.1 ± 3.0**	64.5 ± 3.7	83.0 ± 1.3	**71.1 ± 4.3**	69.2 ± 9.2

Table 8: Test results for models trained per legal area (domain) or across all legal areas (domains). Best overall results are in **bold**, and in-domain are underlined. The mean and standard deviations are computed across languages per legal area and across legal areas for the right-most column. The number in brackets shows the number of examples in the train set per legal area.

E Motivation for Cross-Jurisdiction Transfer

Legal systems vary from country to country. Although they develop in different ways, legal systems also have some similarities based on historically accepted justice ideals. Switzerland has a civil law legal system, i.e., statutes (legislation) is the primary source of law, at the crossroads between Germanic and French legal traditions. Contrary, India maintains a hybrid legal system with a mixture of civil, common law and customary, Islamic ethics, or religious law within the legal framework inherited from the colonial era and various legislation first introduced by the British are still in effect in modified forms today.

Although the Supreme Court of India (SCI) rules under the Indian jurisdiction (law), while the Fed-

Dataset	de	fr	it	Avg
XLM-R + MT Swiss	70.2 ± 0.5	71.5 ± 1.1	72.1 ± 1.2	71.3 ± 0.7
+ MT Indian	**70.5 ± 0.4**	**71.8 ± 0.3**	**73.5 ± 1.4**	**72.0 ± 0.9**
Only MT Indian	50.4 ± 1.5	47.9 ± 1.0	49.5 ± 1.3	49.3 ± 1.0

Table 9: Test results for cross-jurisdiction transfer in both settings: *zero-shot* (Only MT Indian) and *augmented* (+ MT Indian). Best results are in **bold**. ***Augmenting with Indian cases is overall beneficial.***

eral Supreme Court of Switzerland (FSCS) under the Swiss one, we hypothesize that the fundamentals of law in two primarily civil law legal systems are quite common, especially in penal law, and thus transferring knowledge could potentially have a positive effect.

F Responsible NLP Research

We include information on limitations, licensing of resources, and computing foot-print, as suggested by the newly introduced Responsible NLP Research checklist.

F.1 Limitations

In this appendix, we discuss core limitations that we identify in our work and should be considered in future work.

Adapter under-performance Contrary to the literature (Pfeiffer et al., 2021a,b; Chalkidis et al., 2021), in our case, Adapters do not improve in the cross-lingual transfer setting over fine-tuning. Although we tuned both the learning rate and the reduction factor (see Appendix A.3), we did not manage to improve the performance. So far, we do not have a reasonable explanation for this behavior.

Data size flunctuations We did not control for the sizes of the training datasets, which is why we reported them in the Tables 2, 3 and 4. This mimics a more realistic setting, where the training set size differs based on data availability. However, we cannot completely rule out different performance based on simply more training data.

Mismath in in/out of region model perfomance As described in Section 3.3.2, certain zero-shot evaluations outperform in-domain evaluations. Although we try to find an explanation for this in appendix C, it remains an open question.

Re-use of Indian cases Although we have empirical results confirming the statistically significant positive effect of training with additional translated Indian cases, we do not have a thorough legal justification for this finding at the moment.

F.2 Licensing

The SJP dataset (Niklaus et al., 2021) we mainly use in this work is available under a CC-BY-4 license. The second dataset, ILDC (Malik et al., 2021), comprising Indian cases is available upon request. The authors kindly provided their dataset. All used software and libraries (EasyNMT, Hugging Face Transformers, deep-significance, and several other typical scientific Python libraries) are publicly available and free to use, while we always cite the original work and creators. The artifacts (i.e., the translations and the code) we created, target academic research and are available under a CC-BY-4 license.

F.3 Computing Infrastructure

We used an NVIDIA GeForce RTX 3090 GPU with 24 GB memory for our experiments. In total, the experiments took approx. 70 GPU days, excluding the translations. The translations took approx. 7 GPU days per language from Indian to German, French, and Italian. The translation within the Swiss corpus took approx. 4 GPU days in total.

Region	**ZH**	**ES**	**CS**	**NWS**	**EM**	**RL**	**TI**	**FED**	**All**
ZH	65.5 ± 0.0	65.6 ± 0.0	63.7 ± 0.0	68.2 ± 0.0	62.0 ± 2.9	57.9 ± 6.7	63.2 ± 0.0	54.8 ± 5.1	62.6 ± 4.1
ES	62.9 ± 0.0	66.9 ± 0.0	62.8 ± 0.0	65.2 ± 0.0	62.2 ± 1.1	60.2 ± 5.3	57.8 ± 0.0	55.1 ± 6.3	61.6 ± 3.6
CS)	62.5 ± 0.0	65.5 ± 0.0	63.2 ± 0.0	65.1 ± 0.0	60.7 ± 1.6	57.8 ± 3.7	60.5 ± 0.0	55.9 ± 0.5	61.4 ± 3.1
NWS	66.0 ± 0.0	68.6 ± 0.0	65.2 ± 0.0	67.9 ± 0.0	61.6 ± 1.7	57.0 ± 4.9	57.1 ± 0.0	55.5 ± 5.7	62.4 ± 4.9
EM	64.1 ± 0.0	66.6 ± 0.0	63.3 ± 0.0	66.7 ± 0.0	64.0 ± 0.7	66.8 ± 2.9	63.2 ± 0.0	58.4 ± 0.3	64.1 ± 2.6
RL	61.0 ± 0.0	64.7 ± 0.0	60.2 ± 0.0	63.7 ± 0.0	63.4 ± 3.3	69.8 ± 2.7	67.6 ± 0.0	54.3 ± 7.2	63.1 ± 4.4
TI	55.0 ± 0.0	56.3 ± 0.0	53.2 ± 0.0	54.5 ± 0.0	56.0 ± 0.4	54.7 ± 0.9	66.0 ± 0.0	53.1 ± 6.4	56.1 ± 3.9
FED	57.5 ± 0.0	59.6 ± 0.0	56.8 ± 0.0	58.9 ± 0.0	55.0 ± 1.0	56.5 ± 1.1	53.5 ± 0.0	54.9 ± 2.9	56.6 ± 1.9
All (XLM-R)	**69.2 ± 0.0**	**72.9 ± 0.0**	68.3 ± 0.0	**73.3 ± 0.0**	**69.9 ± 1.6**	71.7 ± 2.8	**70.4 ± 0.0**	**65.0 ± 3.9**	**70.1 ± 2.5**
All (Native)	69.0 ± 0.0	72.1 ± 0.0	**68.6 ± 0.0**	72.0 ± 0.0	**69.9 ± 1.6**	**71.9 ± 0.7**	68.8 ± 0.0	64.8 ± 7.0	69.6 ± 2.3

Table 10: Test results for models trainer per region (domain) or across all regions (domains). Best overall results are in **bold**, and in-domain are underlined. The mean and standard deviations are computed across languages per origin region and across origin regions for the right-most column. The regions where only one language is spoken thus show std 0.