Realistic Zero-Shot Cross-Lingual Transfer in Legal Topic Classification

Anonymous ACL submission

Abstract

We consider zero-shot cross-lingual transfer in legal topic classification using the recent Multi-EURLEX dataset. Since the original dataset contains parallel documents, which is unrealistic for zero-shot cross-lingual transfer, we develop a new version of the dataset without parallel documents. We use it to show that translation-based methods vastly outperform cross-lingual fine-tuning of multilingually pre-trained models, the best previous zero-shot transfer method for Multi-EURLEX. We also develop a bilingual teacher-student zero-shot transfer approach, which exploits additional unlabeled documents of the target language and performs better than a model fine-tuned directly on labeled target language documents.

1 Introduction

Transformer-based (Vaswani et al., 2017) pre-trained models (Devlin et al., 2019) have significantly improved performance across NLP tasks. Multilingually pre-trained models (Conneau et al., 2020; Xue et al., 2021) have also been used for *zero-shot cross-lingual transfer* (Hu et al., 2020; Ruder et al., 2021), i.e., fine-tuning (further training) in one or more source languages and applying the model to other target languages at inference.

NLP for legal text has become popular (Zhong et al., 2020; Hendrycks et al., 2021; Chalkidis et al., 2021a,b; Xiao et al., 2021), but to our knowledge only Chalkidis et al. (2021a) have considered cross-lingual transfer of neural models in legal NLP. They introduced a multilingual dataset, Multi-EURLEX, for legal topic classification and explored zero-shot cross-lingual transfer using multilingually pre-trained models like XLM-R (Conneau et al., 2020) combined with adaptation (Houlsby et al., 2019; Zaken et al., 2021) to retain multilingual knowledge from pre-training. Multi-EURLEX, however, contains to a large extent *parallel* text (same content in multiple languages), which is unrealistic in real-world cross-lingual transfer. Also, Chalkidis et al. (2021a) did not consider *translation-based* methods (Lample and Conneau, 2019), which machine-translate the target language documents to a source language, or machine-translate the labeled source documents to the target languages and use the translations to train models for the target languages. *Teacher-student* approaches, which leverage multilingual teacher models to soft-label unlabeled documents of the target language(s) to train a student (Eisenschlos et al., 2019), were also not considered. We address these limitations in this work.

- We construct, use, and release a new, **more realistic version of Multi-EURLEX** that contains non-parallel training documents in four languages (English, French, German, Greek), along with the same (parallel) development and test documents for those languages as in the original dataset.
- To establish **'upper' performance bounds** for zero-shot transfer, we fine-tune XLM-R separately per language, as well as jointly in all four languages, simulating a scenario where there are equally many training documents in all languages, confirming that adapters improve cross-lingual transfer. Unlike Chalkidis et al. (2021a), we find that jointly fine-tuning for all languages leads to better performance, compared to monolingual fine-tuning. We partly attribute this difference to the fact that the original dataset contains parallel documents (same content), which reduces the benefit of jointly training in multiple languages.
- We show that **translation-based methods *vastly* outperform cross-lingual fine-tuning** with adapters, which was the best zero-shot cross-lingual transfer method of Chalkidis et al. (2021a). This suggests that exploiting modern Neural Machine Translation (NMT) systems is a much better zero-shot cross-lingual transfer strategy in real life, at least for legal topic classification.
- We develop a **bilingual teacher-student**. A multilingually pre-trained teacher is fine-tuned on

labeled documents of the source language and their machine-translations in the target language. The teacher then soft-labels all the documents it was trained on, and also soft-labels unlabeled documents of the target language. A student is then trained to predict all the soft labels. Its performance **exceeds the monolingual 'upper bound', i.e., fine-tuning directly in the target language**. Also, the student supports both the target and the source language, which allows a company to support **both languages with a single model**.

2 Related Work

Pre-trained Transformers have boosted performance across NLP, including cross-lingual transfer (Conneau and Lample, 2019; Conneau et al., 2020; Xue et al., 2021). Adapter modules (Houlsby et al., 2019) have been used to transfer pre-trained models to low-resource or even unseen languages (Pfeiffer et al., 2020, 2021). Also, Eisenschlos et al. (2019) proposed MultiFiT, a teacher-student framework that allows pre-training and fine-tuning monolingual students in a target language, using a multilingually pre-trained teacher to bootstrap the student with soft-labeled documents of the target language.

Gonalves and Quaresma (2010) performed legal topic classification in English, German, Italian, Portuguese using monolingual SVMs and their combination as a multilingual ensemble. Chalkidis et al. (2021a) studied zero-shot cross-lingual transfer in legal topic classification, introducing Multi-EURLEX. They found that fine-tuning a multilingually pretrained model in a single language leads to catastrophic forgetting of the multilingual knowledge from the pre-training and, thus, performs poorly in zero-shot transfer to other languages. To retain the multilingual knowledge, they used adaptation strategies (Houlsby et al., 2019; Pfeiffer et al., 2020). Their results also show that zero-shot cross-lingual transfer is more challenging in legal topic classification, compared to more generic classification tasks (Hu et al., 2020; Ruder et al., 2021).

3 The New Multi-EURLEX Version

We use Multi-EURLEX (Chalkidis et al., 2021a), a multilingual dataset for legal topic classification comprising 65k EU laws officially translated in 23 EU languages.[1] Each document (EU law) was originally annotated with relevant EUROVOC[2] concepts by the Publications Office of EU. EUROVOC is a taxonomy of concepts (a hierarchy of labels). We use the 127 'Level 2' labels, obtained by Chalkidis et al. (2021a) from the original EUROVOC annotations of the documents.

Limitations of Multi-EURLEX: One limitation of Multi-EURLEX is that the number of training documents is not the same across languages. For languages spoken in the older EU member states, there are 55k training documents per language, but for many others, there are much fewer training documents (e.g., 8k for Croatian, 15k for Bulgarian). This makes zero-shot cross-lingual transfer results difficult to compare, because the training set size varies across experiments, a factor not controlled for by Chalkidis et al. (2021a). More importantly, when training in several source languages, most of the source language documents are parallel (same content in multiple languages), which is unrealistic in most real-life applications and may produce misleading results. For example, in one of their baselines, Chalkidis et al. (2021a) jointly fine-tune a multilingually pre-trained model on the (parallel) training documents of all the 23 languages, and observe no performance benefit compared to fine-tuning a different instance of the model per language, possibly because of the fact that the training documents are parallel (same content). By contrast, we find that the multilingually fine-tuned model is substantially better than the monolingual ones, when the training documents are not parallel.

Updated Harder Version: We, therefore, construct, use, and release a new, more realistic version of Multi-EURLEX, where there are no parallel training documents across languages. For the new version, we randomly selected 12k (11k training, 1k development) documents per language, limiting the languages to four, namely English, German, French, Greek, and making sure there are no parallel documents. Using four languages allowed us to avoid parallel documents, but still have a reasonably large training set (11k) per language. The test sets are still parallel (5k training per language, as in the original Multi-EURLEX) to allow comparisons to be made when changing the target language. The four languages are from three different families (Germanic, Romance, Hellenic), which makes cross-lingual transfer harder.

[1] Multi-EURLEX is available at `https://huggingface.co/datasets/multi_eurlex`. Our modified version will be made publicly available when this work is published.

[2] `http://eurovoc.europa.eu/`

Model	#M	MT	BS+SL	Source en	Target Languages de	fr	el	Target Avg
'Upper' performance bounds (labeled training documents available in all 4 languages)								
Monolingual FT (Fine-Tuning on labeled documents of a particular language only)								
XLM-R (E2E)	4	✗	✗	68.2 ± 0.8	65.8 ± 0.7	67.0 ± 1.7	64.6 ± 0.4	65.8
XLM-R +Adapters	4	✗	✗	68.8 ± 0.1	65.0 ± 0.7	68.1 ± 0.4	64.9 ± 0.2	66.0
Multilingual FT (jointly Fine-Tuning on labeled documents of all 4 languages)								
XLM-R (E2E)	1	✗	✗	70.0 ± 1.0	68.9 ± 1.0	69.1 ± 1.5	67.4 ± 0.6	68.5
XLM-R +Adapters	1	✗	✗	70.4 ± 1.6	69.2 ± 1.1	69.9 ± 1.6	67.1 ± 0.5	68.7
Zero-shot Cross-lingual Methods (no labeled training documents available in the Target languages)								
Cross-lingual FT (FT on Source documents only, test in each Target language directly)								
XLM-R (E2E)	1	✗	✗	—	55.2 ± 5.2	58.1 ± 2.9	42.8 ± 6.5	52.0
XLM-R +Adapters	1	✗	✗	—	61.7 ± 1.9	60.6 ± 0.8	48.1 ± 1.8	56.8
Translate Test (FT on Source documents only, test on Target documents translated to Source)								
XLM-R (E2E)	1	✓	✗	—	63.3 ± 1.8	68.1 ± 0.8	66.5 ± 1.0	66.0
XLM-R +Adapters	1	✓	✗	—	62.8 ± 1.0	**68.7** ± 0.2	67.2 ± 1.2	66.2
Translate Train (translate the Source training documents to each Target, FT on the translations)								
XLM-R (E2E)	4	✓	✗	—	66.7 ± 1.5	67.2 ± 1.1	64.1 ± 1.4	66.0
XLM-R +Adapters	4	✓	✗	—	67.2 ± 1.0	67.0 ± 1.2	64.8 ± 1.7	66.4
Bilingual Teacher-Student (jointly FT on Source documents and their translations in a Target language)								
XLM-R (E2E) 🧙💭👶	4	✓	✓	**69.1** ± 1.3	**67.4** ± 0.1	66.1 ± 0.3	65.0 ± 0.4	66.1
XLM-R +Adapters 🧙💭👶	4	✓	✓	67.8 ± 1.3	66.9 ± 0.3	67.6 ± 1.2	**67.9** ± 0.1	**67.5**
Multilingual Teacher-Student (jointly FT on Source documents and their translations in all Target languages)								
XLM-R (E2E) 🧙💭👶	1	✓	✓	62.3 ± 1.6	60.9 ± 0.3	66.8 ± 0.2	48.4 ± 0.3	58.7
XLM-R +Adapters 🧙💭👶	1	✓	✓	65.0 ± 0.2	62.6 ± 0.2	**68.7** ± 0.8	50.5 ± 0.0	60.6

Table 1: Test R-Precision (RP, %) results ± std. deviation over 3 runs with different random seeds. E2E: End-to-End Fine-Tuning (FT). +Adapters: Updating only Adapter layers and classification head during FT. #M: number of models fine-tuned. MT: machine-translated documents used. BS+SL: Boot-Strapping with Soft Labels. 🧙💭👶: Teacher-Student approach. Best zero-shot scores per language shown in **bold**. Teacher scores in the Appendix.

4 Experimental Setup and Methods

We experiment with XLM-R (Conneau et al., 2020) in the two best-performing configurations of Chalkidis et al. (2021a): (a) *End-to-end* (E2E) fine-tuning, where all model parameters are updated, and (b) *Adapter-based* (Houlsby et al., 2019) fine-tuning, where we only update the parameters of additional bottleneck (adapter) layers between the pre-trained Transformer blocks. We compare both configurations across several training settings:

'Upper' Performance Bounds: Firstly, we examine the performance of XLM-R fine-tuned in a *monolingual* fashion, i.e., separately on the labeled documents of each language (source or target), or in a *multilingual* fashion, i.e., jointly on training documents of all four languages. In real life, labeled data in the target languages are rarely available. Typically a company has trained a system on English labeled documents and wishes to deploy it in other languages with very few (or no) labeled documents. However, these experiments show how high performance would be in an ideal case with labeled documents in each target language (as many as in the source language). We call them an 'upper' bound, because we would expect performance to be inferior in zero-shot cross-lingual transfer, where no labeled documents are available in the target languages. Nevertheless, our best zero-transfer method, actually surpasses some 'upper' bounds.

Cross-lingual Fine-Tuning (FT): Chalkidis et al. (2021a) showed that when fine-tuning a multilingually pre-trained model for a particular language, the model 'forgets' to a large extent its knowledge of the other languages and performs poorly in zero-shot cross-lingual transfer, unless adaptation mechanisms are used; but even then, zero-shot performance was much lower than the 'upper' bounds.

Translation-based Methods: Following Conneau et al. (2020) and Xue et al. (2021), we also consider methods that exploit machine-translated documents.[3] In *Translate Test*, we fine-tune XLM-R for the source language; given a target language document at inference time, we simply translate it to the source language and use the fine-tuned (for the

[3] We use the EasyNMT (Reimers, 2021) framework.

source language) XLM-R. In *Translate Train*, we machine-translate the labeled training documents of the source language to the target language, and use the translations (and the original labels) to fine-tune XLM-R for the target language; at test time, we evaluate on labeled test documents written in the target language (not machine-translated).

Teacher-Student: Inspired by Eisenschlos et al. (2019), we first fine-tune a *bilingual teacher* XLM-R using labeled documents in the source language and their machine translations (and original labels) in the target language. Then, we use the teacher to soft-label (assign a probability per label to) the source and machine-translated documents it was trained on, and to soft-label additional unlabeled documents of the target language; we use the 12k training documents of the target language without their labels. We then train a *student* XLM-R (on all the documents the teacher soft-labeled) to predict the soft labels. The student (and the teacher) is bilingual, i.e., it supports both the target and the source language. This allows a company to support both languages with a single model, which has cost benefits. We also experiment with a *multilingual teacher-student* approach, where a single multi-lingual teacher is jointly fine-tuned on labeled documents of the source language and their machine translations in all target languages. The teacher then soft-labels all the documents (and translations) it was trained on and additional unlabeled documents of the target languages. The student is again trained to predict the soft labels.[4] In this case, all four languages are supported.

5 Experimental Results

Table 1 reports test results. Following Chalkidis et al. (2021a), we report average R-Precision (RP) (Manning et al., 2009) alongside ($\pm$) standard deviation over 3 runs with different random seeds. Starting from the 'upper' bound results, we find that jointly fine-tuning on all four languages performs substantially better than fine-tuning monolingual models. By contrast, Chalkidis et al. (2021a) reported no benefit when jointly fine-tuning XLM-R for multiple languages. However, in their experiments there were many more training documents per language and the documents were parallel translations (same content), which reduced the benefit of jointly training in multiple languages (in our case, four times more documents with different content).

Cross-lingual FT with Adapters performs approx. 10 points lower in the target languages on average, compared to the corresponding monolingual 'upper' bound (56.8 vs. 66.0). Translate Test and Train, which were not considered by Chalkidis et al. (2021a), vastly outperform Cross-lingual FT with Adapters, which was the best zero-shot method of the same authors, and perform on par with the monolingual 'upper' bounds.[5] The bilingual student with Adapters improves the average performance on target languages slightly further (67.5), exceeding the monolingual 'upper' bound with Adapters (66.0). This improvement can be attributed to the additional (originally unlabeled) documents of the target languages and the soft labels that the student uses. Recall that the student has the further practical advantage of supporting two languages.

The multilingual student performs much worse on average, compared to the bilingual student, even with Adapters; with an exception for French where the student performs best (68.7) compared to all other methods. The results seem to be related to (affected by) the translation quality across target languages and the quality of the teacher's soft labels. We conduct an analysis for both aspects (translation and soft labels quality) in Appendix A.

6 Conclusions and Future Work

We considered zero-shot cross-lingual transfer in legal topic classification, introducing a more realistic version of Multi-EURLEX without parallel documents. We showed that translation-based methods vastly outperform cross-lingual fine-tuning of multilingually pre-trained models, the best previous zero-shot transfer method for Multi-EURLEX. We also developed a bilingual teacher-student zero-shot transfer approach, which exploits additional unlabeled documents of the target language and performs better than a model fine-tuned directly on labeled target language documents, while supporting both languages with a single model.

In future work, we aim to better understand the reasons of the poor performance of the *multilingual* teacher-student and hopefully to address them, in order to deploy a single zero-shot cross-lingual transfer model for multiple target languages.

[4]The student sees soft labels even in the manually labeled target documents and their translations, since soft labels have been found beneficial in manually labeled documents too (Fornaciari et al., 2021). Preliminary experiments confirmed this.

[5]The same conclusions can be drawn with other source languages (French, German, Greek); see Appendix B.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021a. 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.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael J. Bommarito II, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021b. Lexglue: A benchmark dataset for legal language understanding in english. *CoRR*, abs/2110.00976.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. 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)*, pages 5702–5707, Hong Kong, China. 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.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Teresa Gonalves and Paulo Quaresma. 2010. Multilingual text classification through combination of monolingual classifiers. In *CEUR Workshop*, volume 605.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *NeurIPS*.

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. *CoRR*, abs/1902.00751.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *Introduction to Information Retrieval*. Cambridge University Press.

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. 2021. 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.

Nils Reimers. 2021. Easy NMT - Easy to use, state-of-the-art Neural Machine Translation.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

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.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *CoRR*, abs/2106.10199.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. 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*, pages 5218–5230, Online. Association for Computational Linguistics.

A Quality Assessment

We observed that the multi-lingual teacher-student under-performs compared to the rest of the zero-shot cross-lingual settings, while also its bilingual counterparts show strong results. We hypothesis that these overall negative results (or positive for French) are correlated with the translation quality across target languages and the quality of the soft labels generated by (predicted) the teacher.

Translation Quality: In Table 2, we report the quality of machine-translations measured with the METEOR score (Banerjee and Lavie, 2005). We observe that the quality from English to French (0.73) is substantially better compared to the one from English to German or Greek (0.68). This quality disparity could potentially affect the the performance of all methods that use machine-translated documents, i.e. translate-train, translate-test, bilingual/multilingual teacher-student. Indeed, we observe in Table 1, that these methods are consistently better in French, while being comparable in German, and worse in Greek. This is quite expected as both French, and German use the Latin alphabet, and share a larger part of vocabulary compared to Greek, using the Greek one.

Soft Labels Quality: In Figure 1, we estimate the quality of soft labels via the absolute differences in between *gold* and *soft* labels predicted by the multilingual Teacher model across all document subsets (original in English, machine-translated in target languages, and additional unlabelled documents), and languages considered by the student. We compute differences, as the averaged Mean Absolute Error (MAE) across documents in documents subset:

$$\overline{\text{Diff}} = \frac{1}{N}\sum_{n=1}^{N}|G_n - S_n| \qquad (1)$$

where $N = 12,000$ is number of documents translated from English to a target language, and G_n, S_n are the *gold* and *soft* labels per document. We observe that the quality of the soft labels vastly varies both across documents subsets (considering the mean difference reported per violin with a thick blue horizontal line), and across documents per subset (distribution in each violin).

The average differences ($\overline{\text{Diff}}$) per language (source or target) fully correlate with the performance of the student model in the respective language, measured in RP, as reported in Table 1. Specifically, soft labels for French documents (machine-translated or unlabelled) are more accurate ($\overline{\text{Diff}} \simeq 0.25$) compared to the rest: $\overline{\text{Diff}} \simeq 0.45$ for German, and $\overline{\text{Diff}} \simeq 0.60$ for Greek. These results (soft label quality) seem to justify the performance improvement in French, compared to performance decrease in German and Greek. These results could also be affected by the quality of NMT (Table 2).

Based on these findings, we acknowledge that bootstrapping should be reconsidered in the future with respect to the quality of translations and soft labels. Such improvements could include filtering of documents with very uncertain soft labels

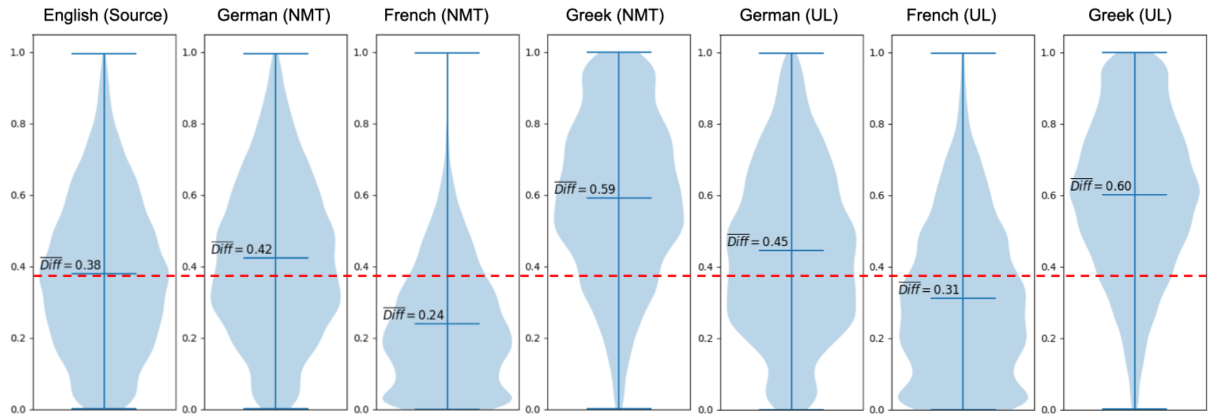

Figure 1: Difference (left blue parts) in between *gold* and *soft* labels predicted by the multilingual Teacher model, measured as Mean Absolute Error (MAE). Results reported per document subset (original in English (source), machine-translated (MT) in target languages, and additional unlabelled (UL)) and target language.

METEOR scores		
en-to-de	**en-to-fr**	**en-to-el**
0.680	0.733	0.680

Table 2: Quality of machine-translations, English (en) to targets (German (de), French (fr) and Greek (el)), provided by the NMT systems measured in METEOR.

(probabilities), e.g., very close to a threshold (e.g., $t = 0.5$), or weighting with respect to the labeling uncertainty. Similarly, one could possibly filter out, exceptionally low quality translations, measured via language modeling metrics (e.g., perplexity] with a language-specific pre-trained language model.

B Additional Results

In this section, we provide additional results of the same experiments described in Section 4, and presented in Section 5 across more language pairs, i.e., source-target combinations, such as German to the rest, French and Greek, respectively. Given the results, we can draw very similar conclusions.

C Responsible NLP - Details

C.1 Experimental Details

We follow the best hyper-parameters reported by Chalkidis et al. (2021a). For end-to-end (E2E) fine-tuning with XLM-R, we use a learning rate of 3e-5. When we use adapter modules, we use a learning rate of 1e-4, and the botteneck size is 256. For additional details consider Appendix A of Chalkidis et al. (2021a).

C.2 Licensing / Intended Use / Privacy

Both the dataset and code base of Chalkidis et al. (2021a) are available under CC-BY-4.0 license and we re-distribute the augmented dataset (incl. translations) and the updated code under the same license.

C.3 Computational Details

In all of our experiments we fine-tune the XLM-R model (Conneau et al., 2020) consists of 278M params with batch size (BS) equal to 8 and learning rate equal to 3e-05. When adapters modules were used we selected a Bottle-neck Size, the number of hidden units (K), to be equal to 256 as in the work of Chalkidis et al. (2021a) this number gave the best results. All experiments ran on an NVIDIA DGX-1 station with 8 NVIDIA V100 16GB GPU cards. In Table 6 we show the run-time (Hours:Minutes) of every experiment across the 3 runs performed with different random seed.

D Translation Details

We performed the translations using the EasyNMT[6] framework utilizing the *many-to-many* M2M_100_418M model of (Fan et al., 2020) for el-to-en and el-to-de pairs and the OPUS-MT (Tiedemann and Thottingal, 2020) models for the rest. A manual check of some translated samples showed sufficient translation quality.

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

Model	#M	NMT	SL + BS	Source de	Target Languages en	fr	el	Target Avg
Zero-shot Cross-lingual FT (No labeled data in target languages)								
Cross-lingual FT (German Only)								
XLM-R	1	✗	✗	65.84 ± 0.68	57.43 ± 1.61	53.95 ± 2.48	44.97 ± 1.09	52.1
XLM-R + Adapters	1	✗	✗	64.98 ± 0.72	61.30 ± 1.70	58.28 ± 0.60	49.02 ± 1.09	56.2
Translate Test documents to Target language								
XLM-R	1	✓	✗	65.84 ± 0.68	65.65 ± 0.72	65.66 ± 0.78	63.57 ± 0.74	65.0
XLM-R + Adapters	1	✓	✗	64.98 ± 0.72	65.66 ± 1.16	64.76 ± 0.50	64.70 ± 1.61	65.0
Translate Train documents to Target language								
XLM-R	N	✓	✗	65.84 ± 0.68	67.36 ± 1.62	65.64 ± 1.14	64.32 ± 1.21	65.8
XLM-R + Adapters	N	✓	✗	64.98 ± 0.72	66.03 ± 1.40	65.74 ± 1.53	63.85 ± 0.18	65.2

Table 3: Test R-Precision (RP, %) results ± std. deviation over 3 runs with different random seeds. E2E: End-to-End Fine-Tuning (FT). +Adapters: Updating only Adapter layers and classification head during FT. #M: number of models fine-tuned. MT shows if machine-translated documents are used. BS+SL shows if teacher-student Boot-Strapping with Soft Labels is used.

Model	#M	NMT	SL + BS	Source fr	Target Languages en	de	el	Target Avg
Zero-shot Cross-lingual FT (No labeled data in target languages)								
Cross-lingual FT (French Only)								
XLM-R	1	✗	✗	67.01 ± 1.69	65.26 ± 0.85	57.04 ± 2.74	49.27 ± 2.17	57.2
XLM-R + Adapters	1	✗	✗	68.05 ± 0.35	64.98 ± 1.66	61.44 ± 1.80	51.31 ± 1.86	59.2
Translate Test documents to Target language								
XLM-R	1	✓	✗	67.01 ± 1.69	66.73 ± 1.86	59.49 ± 2.26	46.16 ± 0.42	57.5
XLM-R + Adapters	1	✓	✗	68.05 ± 0.35	66.72 ± 1.11	59.59 ± 0.24	46.98 ± 2.56	57.8
Translate Train documents to Target language								
XLM-R	N	✓	✗	67.01 ± 1.69	69.01 ± 0.55	67.51 ± 1.59	67.62 ± 0.42	68.0
XLM-R + Adapters	N	✓	✗	68.05 ± 0.35	68.02 ± 1.11	66.99 ± 1.01	66.00 ± 0.95	67.0

Table 4: Test R-Precision (RP, %) results ± std. deviation over 3 runs with different random seeds. E2E: End-to-End Fine-Tuning (FT). +Adapters: Updating only Adapter layers and classification head during FT. #M: number of models fine-tuned. MT shows if machine-translated documents are used. BS+SL shows if teacher-student Boot-Strapping with Soft Labels is used.

Model	#M	NMT	SL + BS	Source el	Target Languages de	fr	en	Target Avg
Zero-shot Cross-lingual FT (No labeled data in target languages)								
Cross-lingual FT (Greek Only)								
XLM-R	1	✗	✗	64.57 ± 0.39	46.30 ± 3.23	43.09 ± 1.37	41.54 ± 2.02	43.6
XLM-R + Adapters	1	✗	✗	64.86 ± 0.19	49.89 ± 3.81	48.56 ± 4.28	47.98 ± 4.75	48.8
Translate Test documents to Target language								
XLM-R	1	✓	✗	64.57 ± 0.39	64.69 ± 0.49	64.59 ± 1.53	64.62 ± 0.48	64.6
XLM-R + Adapters	1	✓	✗	64.86 ± 0.19	65.41 ± 1.13	62.89 ± 0.95	64.88 ± 0.50	64.2
Translate Train documents to Target language								
XLM-R	N	✓	✗	64.57 ± 0.39	65.29 ± 1.51	64.31 ± 2.27	64.77 ± 1.30	64.8
XLM-R + Adapters	N	✓	✗	64.86 ± 0.19	66.22 ± 0.22	64.76 ± 1.24	65.80 ± 1.56	65.6

Table 5: Test R-Precision (RP, %) results ± std. deviation over 3 runs with different random seeds. E2E: End-to-End Fine-Tuning (FT). +Adapters: Updating only Adapter layers and classification head during FT. #M: number of models fine-tuned. MT shows if machine-translated documents are used. BS+SL shows if teacher-student Boot-Strapping with Soft Labels is used.

Setting	Adapters	Avg Run Time
Monolingual	✗	2h
Monolingual	✓	4h
Multilingual	✗	5h
Multilingual	✓	9h
Cross-lingual + MT	✗	2h
Cross-lingual + MT	✓	4h
Bilingual 🧙💭👶	✗	13h
Bilingual 🧙💭👶	✓	10h
Multilingual 🧙💭👶	✗	18h
Multilingual 🧙💭👶	✓	15h

Table 6: Run-time (Hours:Minutes) of every experiment in Tesla V100 GPU across the 3 runs performed with different random seed.