Cross-Lingual Transfer for Legal Text Classification: Revisiting MULTI-EURLEX and Beyond

Welti Nicolas
ENSAE
nicolas.welti@ensae.fr

Abstract

This paper explores the performance of multilingual Transformer models on legal text classification using the MULTI-EURLEX dataset. We investigate zero-shot transfer across EU languages and adaptation strategies that mitigate catastrophic forgetting. Our experiments replicate baseline results and test new approaches such as translation-based transfer and generalization to national laws, revealing both the strengths and current limitations of multilingual legal NLP.

1 State of the art

Multilingual learning has become a central focus in NLP, initially propelled by advancements in neural machine translation and further extended through the development of multilingual word embeddings and pretrained Transformer-based models such as XLM-R and mT5 [3]. These models enable zero-shot cross-lingual transfer, a capability increasingly leveraged in various NLP tasks.

Despite the emergence of legal NLP [1] as a distinct research area—with studies addressing tasks like legal judgment prediction, topic classification, and contract analysis—cross-lingual approaches remain underexplored. To address this, the MULTI-EURLEX [2] dataset was introduced, comprising 65,000 EU legal documents annotated with EUROVOC labels in 23 official languages.

2 Multi-Eurlex overview

This dataset consists of legislative documents from the European Union, each annotated with one or more labels from the hierarchical EUROVOC thesaurus—a multilingual, multi-domain classification system maintained by the EU. These documents have been officially translated into up to 23 EU languages. However, not all documents are available in every language; the degree of language coverage varies, with some documents present in all 23 languages and others in only a subset, depending on the official translation practices of the EU. Each dataset entry is anchored to a unique document identifier and includes the full text of the law in the available languages alongside the associated EUROVOC categories. This structure supports multilingual and cross-lingual experiments, particularly zero-shot transfer, where a model trained in one language is evaluated on others without additional language-specific training data. The dataset is thus designed to enable robust evaluation of multilingual legal text classification methods across a diverse set of languages, legal domains, and classification granularities.

As said above, the main experiment was the finetuning of multilingual models on documents in one language to enable zero-shot classification in other languages. Indeed, given that some EU languages don't have a very wide corpus of documents compared to others, the fine-tuning is only done in one language, and the goal is to see if the knowledge transfers to documents in other languages. So, for example we will take a multilingual model and finetune it on classifying only English documents. It will then be tested on documents in all languages. Let's take French as an example: the pre-

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

finetuning model "knows" both English and French, it then learns using its knowledge of English, the task of classifying laws. Theoretically, it could then, since it "knows" French, classify French texts in the same way it classifies the English versions. The issue is that, when learning the new task on English documents, the model risks "forgetting" its knowledge of French, since it is longer being optimized on it.

Indeed, naive fine-tuning leads to performance drops in non-English languages. The MULTI-EURLEX paper explored several adaptation strategies:

- **Frozen Layers:** Freeze the first N transformer blocks. Values of $N = 3, 6, 9, 12$ were tested.
- **Adapter Modules:** Trainable modules inserted in each encoder block, only these are fine-tuned. The modules are added after each feed-forward model in the encoder and consists of two dense layers.
- **BitFit:** Only the bias parameters are updated (0.04% of parameters).
- **LNFit:** Only the layer normalization parameters are updated (0.01% of parameters).

All these strategies aim to reduce the catastrophic forgetting by focusing the update on a subset of parameters. All the adaptations strategies show a significant improvement in model accuracy with adapter modules being the best, but BitFfit and LNFit do obtain good results despite only updating a very small portion of the parameters.

	GERMANIC					ROMANCE					SLAVIC			URALIC			
	en	**da**	**de**	**nl**	**sv**	**ro**	**es**	**fr**	**it**	**pt**	**pl**	**bg**	**cs**	**hu**	**fi**	**el**	**All**
One-to-one (Fine-tune XLM-ROBERTA or monolingually pretrained BERTs in one language, test in the *same* language.)																	
NATIVE-BERT	**67.7**	65.5	**68.4**	66.7	**68.5**	**68.5**	67.6	**67.4**	**67.9**	**67.4**	**67.2**	-	**66.7**	**67.7**	**67.8**	**67.8**	**67.4**
XLM-ROBERTA	67.4	**66.7**	67.5	**67.3**	66.5	66.4	**67.8**	67.2	67.4	67.0	65.0	66.1	**66.7**	65.5	66.5	65.8	66.6
Diff.	-0.3	+1.2	-0.9	+0.6	-2.0	-2.1	+0.2	-0.2	-0.5	-0.4	-2.2	-	0.0	-2.2	-1.3	-2.0	-0.7
One-to-many (Fine-tune XLM-ROBERTA *only* in English, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	**67.4**	56.5	52.4	49.0	55.7	55.2	54.0	55.0	52.0	50.5	46.9	51.2	49.6	48.8	46.4	33.3	49.3
First 3 blocks frozen	66.3	59.1	56.8	55.3	57.5	57.9	58.1	57.7	56.2	54.9	53.7	56.1	54.3	51.0	52.1	42.4	53.0
First 6 blocks frozen	66.3	59.1	57.4	55.7	57.9	57.2	56.9	57.9	53.9	55.4	51.9	55.8	52.6	47.3	48.7	39.6	51.7
First 9 blocks frozen	65.8	59.4	57.9	56.9	58.6	58.2	58.7	59.4	55.7	57.5	53.4	56.7	54.2	48.8	50.4	44.5	53.0
All 12 blocks frozen	27.2	21.4	24.6	24.6	23.0	21.6	23.4	21.9	20.1	25.1	22.8	23.1	24.3	22.8	21.9	19.0	22.2
Adapter modules	67.3	**61.5**	**59.3**	**57.8**	**59.5**	**60.3**	**61.0**	**60.4**	**58.8**	**58.5**	**57.5**	**59.2**	**56.8**	**55.3**	**55.6**	**46.1**	**56.1**
BITFIT (bias terms only)	63.9	59.3	57.0	54.0	58.2	57.8	57.4	56.9	56.4	55.5	54.0	55.6	54.8	51.2	54.8	42.1	53.7
LNFIT (layer-norm only)	63.1	58.9	55.7	54.1	56.6	59.1	59.1	58.0	56.6	57.2	55.7	55.4	52.8	51.4	50.7	39.9	53.3
Many-to-many (Jointly fine-tune XLM-ROBERTA in *all* languages, test in all languages, with alternative adaptation strategies.)																	
End-to-end fine-tuning	66.4	66.2	66.2	66.1	66.1	66.3	66.3	66.2	66.3	65.9	65.6	65.7	65.7	65.2	65.8	65.1	65.7
Adapter modules	**67.2**	**67.1**	**66.3**	**67.1**	**67.0**	**67.4**	**67.2**	**67.1**	**67.4**	**67.0**	**66.2**	**66.6**	**67.0**	**65.5**	**66.6**	**65.7**	**66.4**

Figure 1: Training results for Multi-EURLEX with level 3 labels (567 labels)

3 Beyond MULTI-EURLEX

Since the release of the multi-EURLEX dataset in 2021, subsequent research has built upon its foundation to enhance cross-lingual legal topic classification. A notable advancement is presented in the paper "Realistic Zero-Shot Cross-Lingual Transfer in Legal Topic Classification" [4]. This study addresses the limitations of the original multi-EURLEX dataset, which contains parallel documents—identical content across multiple languages—a scenario that doesn't reflect real-world applications. To create a more realistic benchmark, the authors developed a version of the dataset without parallel documents. Their experiments demonstrated that translation-based methods significantly outperform cross-lingual fine-tuning of multilingually pre-trained models, which was previously considered the best approach for zero-shot transfer in this context. Furthermore, they introduced a bilingual teacher-student zero-shot transfer method that leverages additional unlabeled documents in the target language, achieving better performance than models fine-tuned directly on labeled target language documents.

4 Dataset Details

The dataset is made up of 65000 documents put into a 55000 documents train set, 5000 for validation and 5000 for test. Not all documents have translations available in all 23 EU languages. But there are enough to have a full validation and test set in each language, a table with the details is made available in the appendix.

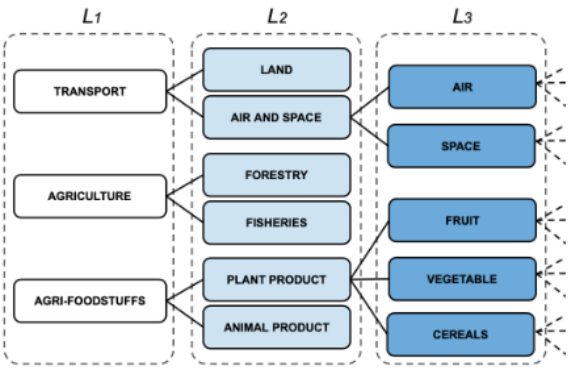

Figure 2: Examples from levels 1 to 3 from the EUROVOC hierarchy. More general concepts become more specific as we move from higher to lower levels.

EUROVOC provides hierarchical concepts. The dataset includes level 1, 2 and 3, if a document is labeled with a level 2 or 3 concept, it won't be labeled with its ancestor. A level 1 category analysis shows strong representation in domains like trade and agriculture.

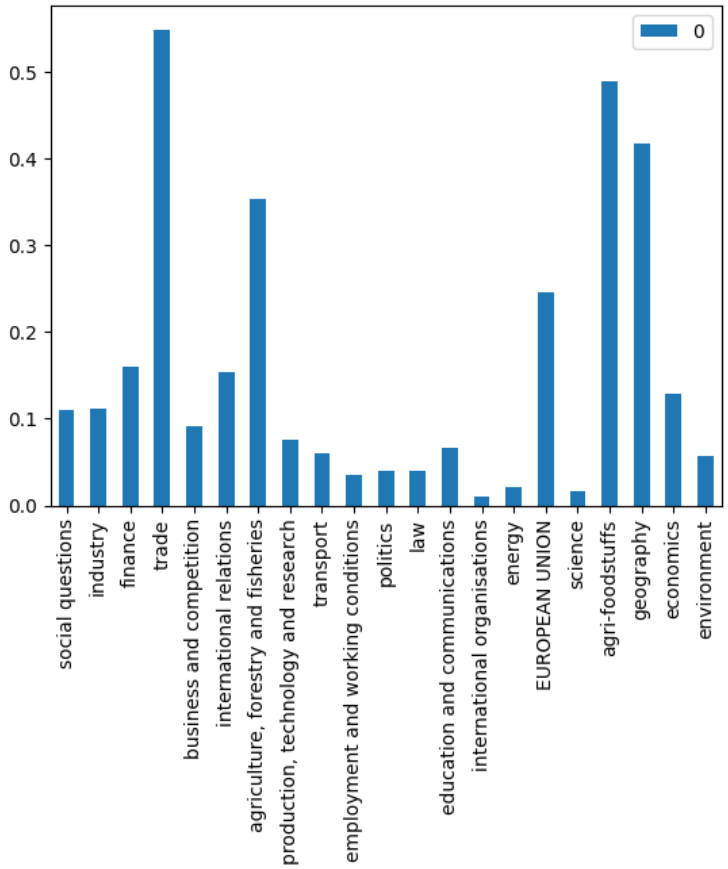

Figure 3: Distribution of Level 1 EUROVOC Categories

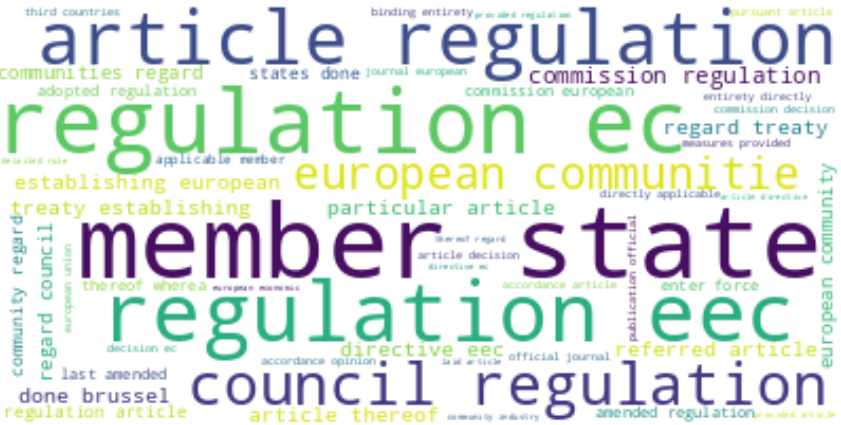

Figure 4: Global Wordcloud of Frequent Legal Terms

We see some keywords such as: member state, council, regulation... which we would expect, given the type of documents we are dealing with. Some of these were added as stop words in the next wordclouds.

Below, we have wordclouds made with texts from specific categories. These wordclouds match expectations from their respective categories. Indeed, it is expected that words like "price" and "market" would appear with high frequency in documents relating to energy, while transport doc-

uments would contain words like “flying” or “vessel” and business regulations documents would have terms such as “dumping”, “company” or “market share”.

Figure 5: Energy Category Wordcloud

Figure 6: Transport Category Wordcloud

Figure 7: Business and Competition Categories Wordcloud

5 Experiments

5.1 Reproducing MULTI-EURLEX

The first step was to (partially) reproduce the paper results, since no model weights were provided, the training had to be redone with our resources. The paper did provide a link to its code (written in tensorflow), and while it could not directly be used, due to broken dependencies and a complex structure, it did provide inspiration for a local pytorch implementation (available on the GitHub). This implementation reruns the training of fine-tuning, with and without adaptation strategies (frozen layers were chosen as an adaptation strategy due to their simplicity). While the results could not be compared one-to-one with the papers, since our training used level 2 labels (to make the next experiments simpler), the results matched expectations from the paper. Training was done in 10 epochs, and the model with best validation loss was saved. More details and graph are available directly in the notebook (retraining-model.ipynb).

We use the r-precision metric. It measures how well the model ranks relevant items by comparing the number of relevant items in the top-k predictions to the total number of relevant items in the dataset. Specifically, R-Precision is defined as the ratio of relevant items found in the top-k positions of the predicted ranking, where k is the number of relevant items in the ground truth. This metric provides insight into the precision of the model's ranking for each query or instance.

Model	English R-score	French R-score	Avg. Params Updated
Fully Fine-tuned	0.82	0.63	100%
Frozen Layers (N=6)	0.77	0.66	50%

Table 1: Performance of adaptation strategies (R score) on English and French test sets.

5.2 Translation-based Zero-Shot Classification

Inspired by the work of [4], we wanted to test using translation models instead of transfer learning to see if it could outperform the original multi-Eurlex paper's approach. The second experiment was to test whether the zero-shot classification approach would yield interesting results on national laws. For simplicity's sake, in both these experiments, the model is fine-tuned in English, and the tests are done on French documents. The adaptation strategy used was frozen layers (with $N = 6$, meaning that half of the layers were frozen).

The first experiment was to compare the performance of the frozen layers' models against the fully finetuned model ran on an English version of the documents. As said before, the fully finetuned model only manages good performance on English documents, due to forgetting of its multilingual capacity. On the other side, the model with frozen layers retains much better performance in French but is outperformed in English (it is still significantly better in English than in French). The fact that we have parallel translations of the documents is perfect for fine-tuning a translation model, and we want to see if it is good enough to create a document that will be accurately classified. The model "opus-mt-en-fr" was therefore downloaded from Hugging Face and finetuned for three epochs. The results:

The r-score drops to 0.1, meaning that we have a colossal drop in model capabilities. It seems the translation destroys key structure in the text, and blocks the classifier from many of its classification abilities. Some documents are still correctly classified, but the translation approach is not competitive in our case. There needs to be more work done on creating a suitable translation model.

5.3 Generalization to National Law

The second experiment is to check how general the legal knowledge of our model is. If it can transfer its legal knowledge from English to French, it might also be able to offer insights on French National Laws. Since we did not find French laws labelled with the EUROVOC classification, we just ran the classifier on 20 French legal documents selected from [dataset] before manually checking if the results made sense.

Across the 20 French national laws tested, the model showed a partial but uneven ability to align them with appropriate EU legal categories. In several cases, it correctly identified relevant themes such as taxation, agriculture, or health, particularly when EU competencies are well-established. However, the model frequently over-predicts broad EU-level categories like European Union law or EU institutions and European civil service, even when the laws concern purely national matters. Redundancy between closely related categories (e.g., employment, labour market, social protection) also appears, suggesting a need for finer semantic distinction or post-processing. Overall, the model generalizes well in EU-relevant domains but lacks precision on other domains that would only be mentioned in national laws (like the military for example).

For example, in one case, the model correctly assigned the labels taxation, European Union law, and Europe to a law approving a tax convention between France and Italy — a case clearly involving international and EU-level coordination. In contrast, for another document (Index 3), a 1951 military budget law specifying personnel leave quotas, received tags such as EU institutions and European civil service and transport policy, despite its purely national administrative scope and absence of any EU-relevant transport regulation. This misclassification points to the model's tendency to over-assign EU-related categories in contexts where they do not apply, the transport mislabelling seems to be coming from the fact that the word air is contained in the document. Military-related might be especially hard for this model as this is not a topic that would often come up in national laws.

The full 20 examples can be found in a csv file in the github repository (or in the relevant notebook directly). Below are the first 5 lines (they include the two examples cited above).

index	titre	contexte	predictions
0	LOI n° 2005-270 du 24 mars 2005 portant statut général des militaires (1)	DEUXIÈME PARTIE : DISPOSITIONS DIVERSES.	social protection, labour market, EU institutions and European civil service, personnel management and staff remuneration, employment
1	LOI n° 2004-1484 du 30 décembre 2004 de finances pour 2005 (1)	DEUXIÈME PARTIE : MOYENS DES SERVICES ET DISPOSITIONS SPÉCIALES TITRE Ier : DISPOSITIONS APPLICABLES À L'ANNÉE 2005 I. - OPÉRATIONS À CARACTÈRE DÉFINITIF A. - Budget général.	financial institutions and credit, civil law, economic policy, accounting, EU institutions and European civil service
2	Loi n°68-5 du 3 janvier 1968 PORTANT REFORME DU DROIT DES INCAPABLES MAJEURS		health, EU institutions and European civil service, European Union law, executive power and public service, labour market
3	Loi n°51-651 du 24 mai 1951 BUDGET DE L'EXERCICE 1951 : DEVELOPPEMENT DES CREDITS AFFECTES AUX DEPENSES MILITAIRES DE FONCTIONNEMENT ET D'EQUIPEMENT	Titre III : Dispositions spéciales Paragraphe 2 : Dispositions relatives au personnel.	EU institutions and European civil service, organisation of transport, transport policy, personnel management and staff remuneration, air and space transport
4	LOI no 90-456 du 1er juin 1990 autorisant l'approbation d'une convention entre le Gouvernement de la République française et le Gouvernement de la République italienne en vue d'éviter les doubles impositions en matière d'impôts sur le revenu et sur la fortune et de prévenir l'évasion et la fraude fiscales (ensemble un protocole et un échange de lettres) (1)		political geography, economic geography, Europe, taxation, European Union law
5	LOI no 95-95 du 1er février 1995 de modernisation de l'agriculture (1)	Titre II : Dispositions relatives à l'exploitation agricole Section 3 : De l'installation en agriculture.	agricultural policy, farming systems, European Union law, accounting, taxation

Figure 8: Legal texts with contexts and predictions

6 Conclusion

We were not able to implement an effective translation model, and our translator causes a significant drop in results qualities. One of the solutions could be to train a translator end to end, with a loss for the translation and for the classification task. Anyway, while translation-based classification might show promise (maybe not in our work, but in the works mentioned earlier), it is not scalable across all languages due to the lack of high-quality translation models, and our unability to fine-tune in document-lacking languages, this lack of document being one of the original reasons for the Multi-Eurlex experiments. Thus, parameter-efficient adaptation strategies such as frozen layers and adapter modules remain essential.

The experiment with French national law suggests that multilingual models trained on EU legislation can partially generalize to national contexts, but only when the law's topics align with topics treated in EU legislations. Further work is still needed to confirm this across languages. Future work should develop automated evaluation protocols for non-parallel multilingual legal corpora using translation and expert validation.

References

[1] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on eu legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy, 2019.

[2] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. Multi-eurlex: A multilingual dataset for eu legal text classification. *arXiv preprint*, 2021.

[3] Felix Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:1–46, 2020.

[4] Stratos Xenouleas, Giannis Panagiotakis Alexia Tsoukara, Ilias Chalkidis, and Ion Androutsopoulos. Realistic zero-shot cross-lingual transfer in legal topic classification. *arXiv preprint*, 2021.