# Efficiently Legal: Fine-tuning Justice

**Helena Pérez Puente**
University of the Basque Country (UPV)
MSc Language Anlysis and Processing
hperez028@ikasle.ehu.eus

**Laura Isabel Vargas García**
University of the Basque Country (UPV)
MSc Language Anlysis and Processing
lvargas010@ikasle.ehu.eus

## Abstract

Machine Translation has advanced significantly, but specific domains like law, with its complex terminology or "legalese," still represents challenges for general models. Training from scratch for such specialised domains is computationally expensive and data-hungry. We aim to create an efficient NMT system for the legal domain by fine-tuning models of the MarianMT framework using reduced parallel data from the JRC-Aquis corpus, offering a more resource-efficient alternative to training a model from scratch. We fine-tuned with both monolingual (EN-FR) and multilingual (EN-ROMANCE) data. Results show that even with small parallel datasets, fine-tuning improves translation quality significantly, boosting BLEU scores. The translation models further enhanced performance, proving that fine-tuning can deliver high-quality translations with less computational cost. This approach makes domain-specific NMT systems for legal texts both feasible and efficient.

## 1 Introduction

As is well known, data acquisition remains one of the primary challenges in any Deep Learning-based approach, due to the vast amounts of data required. Neural Machine Translation (NMT) is no exception. While standard translation tasks pose little difficulty for high-resource languages, significant challenges arise when dealing with specialized domains such as the legal domain, where high-quality data is still scarce.

Legal language, often referred to as legalese, is characterized by features that increase the complexity of translation. These texts frequently contain terminology whose meanings differ significantly from their general usage. For instance, the term action may refer to a lawsuit, and consideration denotes a concept related to contractual obligations rather than general reflection. Legal documents also make extensive use of Latin expressions such as inter alia, and borrowings from other languages, such as voir dire from French.

Fine-tuning remains the state-of-the-art technique for domain adaptation in NMT, particularly when handling highly specialized terminology. Luong and Manning (2015) demonstrated that NMT systems can greatly benefit from fine-tuning on in-domain corpora. This method is especially advantageous as it does not require substantial computational resources.

However, the challenge intensifies when attempting to obtain parallel data for a specific language pair within a specialized domain. The more specialized the corpus required, the harder it is to obtain. Our project addresses this data scarcity issue. We show that even a relatively small parallel dataset (EN-FR) can yield significant improvements in BLEU scores without causing catastrophic forgetting. Furthermore, we demonstrate that domain-specific corpora from related Romance languages (EN-ROMANCE) can be effectively used—even with non-multilingual models such as HelsinkiNLP—to improve translation quality. This is particularly valuable in low-resource scenarios where high-quality in-domain data is limited and computational capacity is constrained.

## 2 Related Work

It has been demonstrated that a Neural Machine Translation (NMT) model initially trained on large amounts of generic parallel data can benefit substantially from specialization in a specific domain. During the fine-tuning process, the model is trained for a few additional iterations using in-domain data, while keeping the original loss function unchanged. Servan et al. (2016) showed that although full retraining can yield better performance, fine-tuning achieves competitive results in only 1% of the time required for complete retraining.

Our study addresses two key research questions: First, is a relatively small in-domain parallel dataset

sufficient for effective fine-tuning? This question is relevant not only in low-resource scenarios but also in settings with more abundant data, where it is crucial to avoid catastrophic forgetting. Second, can domain-specific data from related languages be used when high-quality in-domain parallel data is lacking?

Several studies have demonstrated the efficiency of multilingual models for this task. One such model is mBART, a Transformer-based architecture that is multilingual, featuring a bidirectional encoder and an autoregressive decoder. Lee et al. (2022) showed that fine-tuning mBART with a small amount of domain-specific data yields better results than training a new model from scratch with a large in-domain dataset. This highlights the effectiveness of general pretraining on diverse translation data and shows that substantial performance gains can be achieved even with limited domain data.

Verma et al. (2022) further demonstrated that mBART pretraining benefits from exposure to domain-specific terminology in multiple related languages through transfer learning. In their study, they proposed two fine-tuning strategies for domain adaptation: the language-first approach, which first adapts the model to a specific language pair using general bitext and subsequently to the domain using in-domain bitext; and the domain-first approach, which first adapts to the domain using multilingual in-domain bitext and then to the specific language pair. Their results indicated that mBART-initialized models consistently outperform randomly initialized ones. Notably, the domain-first strategy achieved substantial domain adaptation while requiring 50 times less bitext, remaining competitive in performance.

Building on these findings, our study proposes using a multilingual in-domain corpus from Romance languages to fine-tune the model. Romance languages share lexical roots, which is especially evident at the subword level, a property that has proven highly beneficial for transfer learning. It has been shown that transfer learning across languages or domains is more effective when there is greater subword unit overlap across datasets. This is particularly relevant in the legal domain, which contains a high concentration of technical terms and learned borrowings, many of which derive directly from Latin. For example: constitución (es), costituzione (it), constituição (pt), constitution (fr).

However, in this project we also aimed to test our study on a less complex model—specifically, one without the language modeling component. Helsinki-NLP are models that, although not multilingual in architecture, supports both one-to-one and one-to-many translation. It achieves one-to-many capabilities by incorporating target language tags in the input, thereby directing the model to produce output in the specified language. As such, it is a Transformer that does not share an encoder or decoder across languages, but instead trains a model for each language pair. In our case, this language pair is defined broadly as EN-ROMANCE.

## 3 Experimental Setting

### 3.1 Models

To evaluate our hypothesis—that an efficient NMT system for the legal domain can be obtained through fine-tuning rather than training from scratch—we adopted pre-trained models from the Helsinki-NLP group, specifically from the MarianMTModel library available in Hugging Face Transformers.

We explored two fine-tuning strategies:

- **Single-language-pair approach:** fine-tuning the `Helsinki-NLP/opus-mt-en-fr` model.
- **Multilingual approach:** fine-tuning the `Helsinki-NLP/opus-mt-en-ROMANCE` model, which supports translation between English and multiple Romance languages.

The underlying framework for these models, MarianMT (Junczys-Dowmunt et al., 2018), is a neural machine translation toolkit designed with research efficiency and flexibility in mind. Unlike other toolkits built with larger deep learning libraries, Marian is self-contained and implemented in C++11, which contributes to both its high computational efficiency and its portability.

MarianMT models follow a standard encoder-decoder architecture with attention mechanisms. A key strength of the framework lies in its support for dynamic computation graphs, which allows for greater flexibility in implementing novel architectures and training algorithms. This makes it particularly well-suited for rapid experimentation, prototyping, and domain adaptation tasks such as this one.

The MarianMT framework also supports a wide array of language pairs—over a thousand pretrained models—making it highly versatile. Its

focus on speed, modularity, and reproducibility aligns well with our goal of building a resource-efficient and effective domain-specific translation system.

### 3.2 Corpora

For both fine-tuning and evaluation, we used the **JRC-Acquis** corpus[1] (Steinberger et al., 2006) , a large collection of parallel legal texts from the European Union. This corpus is based on the Acquis Communautaire—the body of EU law—and includes legislative and regulatory documents. It is aligned at the sentence level across up to 22 official EU languages, making it a valuable resource for training domain-specific machine translation systems.

In our study, we focused on four language pairs: English as the source language and four Romance languages as targets—French, Spanish, Italian, and Portuguese. These selections reflect both legal relevance in the EU context and linguistic diversity within the Romance language group.

To prepare the data for training and evaluation, we performed preprocessing using the `sacremoses` library. The resulting corpora were formatted to match the requirements of the Helsinki-NLP models.

We built two parallel datasets for fine-tuning:

- **EN-FR Dataset:** This single-language-pair dataset consists of 80,000 tokenized sentence pairs for training and 10,000 for validation. Tokenization was performed using the pre-trained tokenizer associated with `Helsinki-NLP/opus-mt-en-fr`. For evaluation, we used a test set of 10,000 raw (untokenized) English sentences, allowing us to assess model performance on naturally occurring input.

- **Multilingual EN-ROMANCE Dataset:** This dataset combines four language pairs (EN-FR, EN-ES, EN-IT, EN-PT), with 20,000 sentence pairs per pair in the training set (a total amount of 80,000), and 2,500 per pair in the validation set (a total amount of 10,000). We used the tokenizer from `Helsinki-NLP/opus-mt-en-ROMANCE` for all multilingual data. Two test sets were created: one for English–French and one for English–Italian, each containing 10,000 raw pair sentences for evaluation.

This careful dataset preparation ensures both compatibility with MarianMT's training pipeline and the ability to compare single-language and multilingual fine-tuning strategies under controlled conditions.

### 3.3 Experiment Steps

After preparing and organizing the corpora, we proceeded with the fine-tuning of the selected Helsinki-NLP models. Three different fine-tuning configurations were explored:

- **EN-FR fine-tuning:** `Helsinki-NLP/opus-mt-en-fr` was fine-tuned using the EN-FR dataset.

- **Multilingual fine-tuning:** `Helsinki-NLP/opus-mt-en-ROMANCE` was fine-tuned using the multilingual dataset.

- **EN-FR fine-tuning on multilingual model:** `Helsinki-NLP/opus-mt-en-ROMANCE` was also fine-tuned on the EN-FR dataset to compare its performance in a single-language setting versus multilingual training.

All fine-tuning was conducted using the Hugging Face Transformers library in a Google Colab environment with GPU acceleration. Each model was trained for three epochs, using a learning rate of $5 \times 10^{-5}$ and a weight decay of 0.01. These hyperparameters were selected based on our low-computational-resource setting.

Once fine-tuning was complete, we evaluated both the base and fine-tuned versions of each model on their respective test sets. Translation generation was performed using the standard Hugging Face inference pipeline, which involves:

- Tokenizing the raw English test input using the model-specific tokenizer (`tokenizer()`),

- Generating output tokens with `model.generate()`,

- Decoding the token IDs back into text using `tokenizer.decode()`.

The evaluation was structured as follows:

- `Helsinki-NLP/opus-mt-en-fr` (base and fine-tuned) was evaluated using the EN-FR test set.

[1] https://opus.nlpl.eu/JRC-Acquis/corpus/version/JRC-Acquis

- `Helsinki-NLP/opus-mt-en-ROMANCE` (base and multilingual fine-tuned) was evaluated on both the EN-FR and EN-IT test sets.
- The EN-FR fine-tuned version of `Helsinki-NLP/opus-mt-en-ROMANCE` was also evaluated using the EN-FR test set to assess whether adapting a multilingual model to a single language pair yields any performance advantages or degradation.

As evaluation metric, we used the BLEU score, computed with the `sacrebleu` library. BLEU is particularly suitable for legal translation tasks, where structural consistency is more important than creativity. Legal texts often rely on fixed phrasing and terminology, making BLEU a reliable indicator of translation quality in this context.

## 4 Results

This section presents the BLEU scores obtained for each model across different fine-tuning configurations and test sets. The goal is to observe how different fine-tuning strategies impact translation quality in the legal domain.

| Modelo | BLEU |
|---|---|
| EN-FR test | |
| Helsinki-NLP **EN-FR** (**base**) | **58,524** |
| Helsinki-NLP **EN-FR** (**fine-tuned** 10k-train) | **58,702** |
| Helsinki-NLP **EN-FR** (**fine-tuned** 80k-train) | **62,127** |
| EN-FR test | |
| Helsinki-NLP **EN-ROMANCE** (**base**) | **58,061** |
| Helsinki-NLP **EN-ROMANCE** (**fine-tuned multiling**) | **59,721** |
| Helsinki-NLP **EN-ROMANCE** (**fine-tuned EN-FR**) | **1,227** |
| EN-IT test | |
| Helsinki-NLP **EN-ROMANCE** (**base**) | **51,510** |
| Helsinki-NLP **EN-ROMANCE** (**fine-tuned multiling**) | **52,675** |

Figure 1: BLEU scores achieved by each base and fine-tuned model in the experimental setup

### 4.1 Helsinki-NLP EN-FR

The baseline Helsinki-NLP EN-FR model achieves a BLEU score of 58.524 on the EN-FR test set. When fine-tuned on 10,000 in-domain sentence pairs, the BLEU score slightly improves to 58.702, indicating a marginal benefit from even small-scale fine-tuning. A larger fine-tuning dataset of 80,000 sentences yields a more substantial gain, with the BLEU score increasing up to 62.127.

From a qualitative perspective, we analyzed 10 randomly selected translations from the test set to better understand the nature of the improvements achieved through fine-tuning. Compared to the baseline, the fine-tuned model demonstrates significantly closer alignment with the reference translations.

First, the fine-tuned model shows improved use of legal and institutional terminology, accurately employing domain-specific vocabulary such as *dénominations*, *nomenclature*, and *stock régulateur*. It also adheres more consistently to the stylistic norms of French legal language, for instance, using *no* instead of *nº* in legal references, and correctly phrasing citations like "*no 7 de l'Union européenne*", as opposed to the baseline model's "*nº 7/2004*".

Moreover, the fine-tuned model preserves standard legal phrasing more reliably, maintaining formal and formulaic structures typical of regulatory texts. For example, it correctly generates "*a arrêté la présente décision*", whereas the base model erroneously uses "*a adopté*", which alters the legal nuance. Grammatical accuracy also improves, particularly with respect to number agreement,for instance, correcting "*satisfaisant*" to "*satisfaisants*" in "*résultats satisfaisants*".

Despite these advances, the model still encounters some difficulties when translating structurally complex segments, such as lengthy regulatory clauses. This suggests that while fine-tuning significantly enhances translation quality, further improvements may require additional training or specialized architectural strategies to better handle structural divergence between English and French legal texts.

### 4.2 Helsinki-NLP EN-ROMANCE

For the multilingual Helsinki-NLP EN-ROMANCE model, tested on the EN-FR test set, the baseline version achieved a BLEU score of 58.061. Fine-tuning this model on the multilingual dataset (including 20,000 examples per Romance language) raised the score to 59.721, showing an improvement of 1.66 points in translation quality when adapting the model to legal language across multiple Romance targets. From a qualitative perspective, an analysis of 10 randomly selected translations from the test set reveals that the fine-tuned model delivers a marked improvement in translation quality. Notably, it demonstrates a better handling of legal terminology, for example, choosing *prévoir* over

the more generic *avoir*, as used by the base model. Additionally, the fine-tuned model more effectively resolves gender and syntactic ambiguities in complex legal structures, correctly translating *une tierce personne* instead of *un tiers*. It also exhibits improved mastery of domain-specific vocabulary, opting for terms like *dénominations* rather than *désignations*, and *dispositions* instead of *articles*, thereby reflecting a more formal and legally accurate register.

However, fine-tuning the EN-ROMANCE model exclusively on EN-FR data led to a severe drop in BLEU score (1.227). When inspecting the output, we find that the translations are completely incoherent. For instance:

- **Generated translation**: *"for For l Or fairly fonds de ourselves Nous»es« em quem que EUR 250 million."*
  **Reference**: *"Pour l'exercice budgétaire 2002, le montant de la réserve monétaire est ramené à 250 millions d'euros."*

- **Generated translation**: *"los wanted 2001 l sua l sua l sua l sua..."* (repetitive and nonsensical)
  **Reference**: *"modifiant la décision 2001/76/CE en ce qui concerne les crédits à l'exportation de navires."*

This sharp decline is a clear case of catastrophic forgetting, a phenomenon where the model loses previously learned capabilities—in this case, its multilingual translation abilities—due to overspecialization on a narrow subset of the data. Recent work by Saunders and DeNeefe (2024) explains that when a multilingual model is fine-tuned on just one language pair, it can overfit to that pair and damage its general translation ability, even for the same pair it was fine-tuned on. This explains why our fine-tuned (EN-FR) multilingual model performed so poorly: it lost the balance it originally had across languages.

On the EN-IT test set, the baseline multilingual model scored 51.510, and the version fine-tuned on the multilingual dataset scored 52.675, reflecting a consistent, although modest, improvement through fine-tuning.

## 5 Conclusions and future work

This study demonstrates that it is possible to improve a domain-specific machine translation system through fine-tuning with a relatively small parallel corpus (80,000 sentence pairs) for a specific language pair (English–French), without inducing catastrophic forgetting. This fine-tuning yields an improvement of 3.6 BLEU points, which, in qualitative terms, reflects a more accurate adoption of legal terminology, evidenced by translations such as *tampon* being replaced with *régulateur*, and *adopted* with *arrêté*. These results were achieved without resorting to a multilingual transformer architecture such as mBART. Instead, we show that using a simpler and more accessible model such as Helsinki-NLP's transformers, which are widely used in research, can be highly effective in enhancing translations within specialized domains such as the legal field.

Although the results are not as strong as those achieved with a dedicated one-to-one model, in scenarios where high-quality parallel domain data is scarce, leveraging data from closely related languages can be a viable alternative—especially when these languages share a significant number of subword units, as is the case with Romance languages. This strategy could be particularly beneficial for low-resource Romance languages such as Asturian, Galician, or Catalan.

In our evaluation, we observe a BLEU score improvement of 1.66 points using only 20,000 French legal domain sentences for fine-tuning the EN-ROMANCE model, an encouraging result, particularly when contrasted with a modest 0.17-point improvement obtained using only 10,000 sentences when fine-tuning the EN-FR model. This suggests that the model effectively leverages legal terminology seen in related languages and successfully transfers this knowledge to the target language (French). Similar gains are observed for Italian, where a 1.16-point improvement in BLEU is noted, reinforcing the notion that closely related languages can benefit mutually from shared fine-tuning.

However, in one-to-many translation models that rely on a target-tag approach to specify the output language, fine-tuning on a large amount of data from a single language (e.g., 80,000 sentences) may risk triggering catastrophic forgetting. Future research could investigate optimal data thresholds for fine-tuning in multilingual settings to mitigate this risk while still achieving significant domain adaptation.

## References


Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

E.-S. A. Lee, S. Thillainathan, S. Nayak, S. Ranathunga, D. I. Adelani, R. Su, and A. D. McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Danielle Saunders and Steve DeNeefe. 2024. Domain adapted machine translation: What does catastrophic forgetting forget and why? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12660–12671, Miami, Florida, USA. Association for Computational Linguistics.

Christophe Servan, Josep Maria Crego, and Jean Senellart. 2016. Domain specialization: A post-training domain adaptation for neural machine translation. volume abs/1612.06141.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

N. Verma, K. Murray, and K. Duh. 2022. Strategies for adapting multilingual pre-training for domain-specific machine translation. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 31–44, Orlando, USA. Association for Machine Translation in the Americas.