Comparative evaluation of the MT output of different MT systems on legal texts (EN-FR)

María Viana Rozas
Universidad del País Vasco/
Euskal Herriko Unibertsitatea
mviana009@ikasle.ehu.eus

María Sierro Fernández
Universidad del País Vasco /
Euskal Herriko Unibertsitatea
msierro001@ikasle.ehu.eus

Abstract

The present paper aims at evaluating the MT output of different MT systems trained and tested with texts from the legal domain, with emphasis on the language pair EN-FR. In particular, these MT systems were trained with a parallel dataset from Europarl, and tested on court cases from the European Court of Human Rights (ECHR). Regarding the selected MT systems, we focus on the MT output of supervised NMT with JoeyNMT, supervised NMT with (fine-tuned) Helsinki-NLP, SMT with Moses, and a zero-shot approach with BERT-Multilingual. The motivation behind this work lies in learning which MT systems perform better on this type of texts, considering that legal documents are usually a big challenge for MT.

1 Introduction

It has been argued that the translation of legal texts is important for guaranteeing the right of "individuals who do not speak or do not have sufficient proficiency in the language used by the court authorities" to understand the trial they are involved in (Vigier-Moreno and Pérez-Macías, 2022). However, the manual translation of court cases presents several difficulties, such as the fact that "native speakers of the language used by the court authorities may often have to translate into the other language" (Vigier-Moreno and Pérez-Macías, 2022).This is considered more complicated than translating into one's native tongue (Vigier-Moreno and Pérez-Macías, 2022), and the use of machine translation could be helpful for these professionals. On the other hand, court texts have several characteristics (such as legal terminology and phraseology, complex syntax or highly convoluted sentence structures) which present a challenge for MT systems (Vigier-Moreno and Pérez-Macías, 2022). In light of the usefulness and complications associated to the automatic translation of legal texts, our goal is to contribute to this field by evaluating the output of different MT systems trained and tested with texts from the legal domain. Regarding the selected language pair for our experiments, it will involve the translation from English to French. The reason for this is that the work carried out in this paper will be helpful for increasing the size of a corpus of legal texts in French language which will be used in a Master's thesis. This thesis focuses on another challenging task related to legal texts (anonymization) and the aforementioned corpus would be employed for training and testing different NER models which would recognize entities for anonymization.

Thus, this paper will be structured as follows: section 2 will introduce related work on MT systems; section 3 will present the selected datasets (the Europarl corpus and a corpus of cases from the ECHR); section 4 will comment on the MT methods and models used in our experiments; section 5 will show the results of our experiments; in section 6 we will analyze our results quantitatively and qualitatively; and we will finish with some concluding thoughts in section 7.

2 Related Work

Today, machine translation (MT) presents several challenges that need to be taken into consideration when building or working with MT models. These challenges include domain mismatch, limited availability of training data, handling rare words, processing long sentences, ensuring accurate word alignment, and optimizing beam search, among others (Koehn and Knowles, 2017). The application of neural networks has significantly improved the performance of MT models, prompting researchers to explore innovative approaches to adapt these models for enhanced translation quality. Even further, the introduction of neural machine translation (NMT) has provided opportunities to leverage different techniques, such as data combination and resource integration, to achieve further improvements (Zhixing Tan and Liu, 2020).

Within the field of legal text translation, there remains a strong reliance on manual translation due to the unique challenges it poses. Some legal concepts are inherently complex and not easily transferable to the target language (Salih, 2018). To facilitate the translation process, various resources are utilized, including computer-assisted legal research systems for comparative legal analysis and specialized dictionaries for contextual understanding (Muravev, 2020).

Although there are currently limited machine translation models specifically dedicated to legal texts, ongoing research is actively exploring this domain. For instance, (Fabrizio Gotti and Farzindar, 2008) propose an approach to machine translation of court judgments between English and French languages. Their methodology employs statistical translation models (SMT) and provides a comparative analysis of its performance against machine translation offered by Google.

In recent years, the development and sharing of language-specific datasets for legal texts in diverse languages have accelerated. These datasets have played a vital role in enabling MT models to learn and evolve as valuable tools for translators in the legal domain (Ralf Steinberger and Gilbro, 2014).

3 Selected datasets: Europarl and ECHR

To compare and evaluate the performance of different methodologies, the ’Europarl’ dataset was utilized as the primary data source (Koehn, 2005). It is important to note that there are alternative datasets available specifically for legal texts, such as the ’Digital Corpus of the European Parliament’ (DCEP) (Najeh Hajlaoui and Varga, 2014) and ’JRC-Acquis’(Ralf Steinberger and Varga, 2006), both of which are published by the European Commission.

In addition to the aforementioned dataset used for training and validation, a separate test subset was created using court cases from the European Court of Human Rights (ECHR) as part of the project. This subset consists of six texts for each language, serving as an evaluation set for systems trained on the other dataset. It is worth mentioning that none of the datasets used in the project contain any form of labeling or annotations.

Table 1 provides an overview of the dataset sizes in terms of the number of sentences contained within each one. Additionally, it compares the number of sentences present in the parallel (en-fr) Europarl corpus with its monolingual version, which lacks sentence alignment. In the case of the ECHR dataset, a comparison is made before and after the cleaning process, which will be explained in further detail in the subsequent section.

Dataset	**Type**	**English**	**French**
Europarl	Parallel	50,196,035	51,388,643
	Monolingual	2,218,201	2,190,579
ECHR	Raw	340	340
	Clean	330	330

Table 1: Comparison of the size of the datasets based on number of sentences.

3.1 Europarl

The Europarl dataset is a comprehensive collection of European Parliament proceedings that are accessible online. The dataset, introduced by (Koehn, 2005), is transformed into a parallel corpus encompassing 11 different languages: Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese, and Swedish. Each language pair consists of approximately 30 million words, and the texts date back to 1996, with a primary focus on statistical machine translation (SMT). In our work, we leverage English-French corpus pairs from this dataset, not limited to SMT, to evaluate various methodologies.

Table 1 illustrates that both datasets are comparable in size, with the English dataset being slightly larger. This characteristic facilitates minimizing learning loss on both language sides.

Regarding the corpus composition, the authors (Koehn, 2005) explain that each text corresponds to a distinct topic discussed in the Parliament. Consequently, they have identified the texts based on their topics and attempted to align them across languages, aiming to preserve as much information as possible. In doing so, meta-information associated with each day’s proceedings is included.

For our project, we obtained the data from the pre-divided parallel corpus, which provides separate documents for each language. Within each language division, the corpus consists of text files containing sentence-level language pairs. For instance, the following example showcases an English-French sentence pair: ’I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period.' - 'Je déclare reprise la session du Parlement européen qui avait été interrompue le vendredi 17 décembre dernier et je vous renouvelle tous mes vux en espérant que vous avez passé de bonnes vacances.'

The sentence alignment in the Europarl dataset, as performed by the authors (Koehn, 2005), relies on an implementation of the algorithm developed by Gale and Church. This alignment algorithm matches sentences based on length similarity in the sequence and merges sentences if necessary to maintain consistency in length. While this alignment algorithm may have some margin of error, it has yielded favorable results due to the limited number of sentences per paragraph.

3.2 ECHR

As it was mentioned in the introduction, the work carried out in this paper will also serve for increasing the size of a corpus of legal texts in French language, which will be used in a Master's thesis. Taking into account that such a Master's thesis was inspired by the research conducted by (Ildikó Pilán and Batet, 2022), we have selected texts for our test set which coincide with the work carried out by them. In particular, we collected six texts for each language from the database of the European Court of Human Rights (HUDOC). HUDOC is a comprehensive database that grants public access to an extensive collection of legal documents pertaining to cases brought before the Court. This database encompasses judgments, decisions, reports, and other materials that contribute to the comprehension and interpretation of human rights law in Europe. Furthermore, HUDOC is highly versatile, containing a diverse range of texts from approximately 35 languages [1].

For our project, we have utilized 'The Text Anonymization Benchmark' (Ildikó Pilán and Batet, 2022) as a basis, which comprises 1,268 court cases in the English language. From this collection, we randomly selected six cases and sought their official translations within HUDOC. By doing so, we ensured the ability to construct a parallel corpus based on this data without compromising quality.

To create the corpus subset, which serves as a test set for evaluating the performance of various models and methodologies, we employed a semi-automatic approach. The corpus was generated through a Python script that performed sentence segmentation automatically, followed by a manual alignment process to match sentences line by line.

4 MT Methods and Models

In this study, we have employed various methodologies utilizing the aforementioned datasets to facilitate a comprehensive performance comparison. The objective is to determine the optimal approach for employing Neural Machine Translation (NMT) with legal texts. Specifically, we have explored the following methods: supervised NMT with Joey NMT, supervised NMT with Helsinki-NLP, Statistical Machine Translation (SMT) with Moses, and Zero-shot translation using bert-base-multilingual-uncased. By evaluating and comparing the outcomes obtained from these different approaches, we aim to identify the most effective method for NMT implementation with legal texts.

4.1 Supervised NMT with JoeyNMT

The initial approach employed in this study is supervised NMT, which relies on paired parallel data comprising source language sentences and their corresponding translations in the target language. This method involves training a neural network model to learn the translation process by mapping representations of the source sentences to their respective target sentence representations. The core architecture utilized in supervised NMT is the encoder-decoder model (See Figure 1).

To implement this approach, we leveraged the experiments conducted during the course, utilizing JoeyNMT. Initially, the dataset was prepared and pre-processed through tokenization and cleaning procedures using the 'Byte Pair Encoding' (BPE) method. BPE tokenization aids in improving the translation of rare or out-of-vocabulary (OOV) words. Subsequently, we utilized JoeyNMT, developed by the University of Heidelberg, for training. JoeyNMT offers a flexible configuration system, enabling users to define various aspects of the translation model, including network architecture, training parameters, and data preprocessing options.

While supervised NMT requires a substantial amount of parallel training data for both the source and target languages to achieve optimal translation performance, it is advantageous compared to other methods such as unsupervised or semi-supervised NMT due to the availability of parallel datasets.

[1] For further information, refer to (Echr)

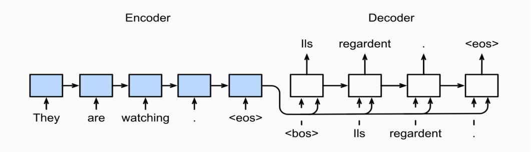

Figure 1: The encoder-decoder architecture

$$
\begin{aligned}
T &= \arg\max_{T} P(T|S) \\
&= \arg\max_{T} P(T)P(S|T)
\end{aligned}
$$

Figure 2: SMT equation (Garg and Agarwal, 2018)

Moreover, supervised NMT generally yields favorable results when trained with extensive data. However, for morphologically rich languages, this approach may encounter challenges in capturing morphological variations and generating accurate inflections in the target language if the dataset does not cover the entire range of morphological patterns (See (Marion Weller-Di Marco and Fraser, 2022)).

4.2 Supervised NMT with fine-tuning

In this approach, we opted for a more advanced deep learning technique by fine-tuning the Helsinki-NLP model specifically for the translation task using a Seq2seq architecture. Our methodology is based on the guidance provided by (Kumar, 2021), which outlines a step-by-step process for this procedure.

To accomplish this, we made use of the Europarl dataset to fine-tune the 'Helsinki-NLP/opus-en-fr' model available in the Hugging Face library. The data was preprocessed using the Transformers Tokenizer, which tokenizes the inputs and formats them according to the model's requirements. Subsequently, we employed the pre-trained model with Seq2seq training, which involves training an encoder-decoder model to convert sequences from one domain to another.

This approach offers several advantages. Firstly, it allows us to leverage the knowledge and insights acquired by pretrained models on large-scale datasets. Secondly, it reduces training time and resource requirements since we build upon an existing model. Additionally, fine-tuning enables the model to adapt its parameters to capture the specific patterns and structures related to translation tasks, thereby enhancing the accuracy and fluency of the translations. Moreover, it is a highly flexible method that allows the use of pretrained models to adapt to different domains.

However, it is important to note that the effectiveness of the fine-tuning process heavily relies on the quality and quantity of resources employed. If the task lacks sufficient or high-quality data, fine-tuning may not yield optimal results. Limited data can lead to issues such as overfitting or inadequate model adaptation, resulting in suboptimal translation performance. Additionally, the effectiveness of the approach depends on the quality of the pre-trained model used.

4.3 SMT with Moses

SMT, which stands for Statistical Machine Translation, is a method that employs statistical models to automatically translate text from one language to another. This approach relies on analyzing extensive amounts of bilingual or parallel text data, such as pairs of sentences, to discern patterns and statistical relationships between words and phrases across different languages.

The core principle of SMT is that each sentence S in a source language can potentially be translated into a target language as sentence T. To determine the most likely translation, SMT approaches assign a probability P(T|S) to each sentence pair (S, T), indicating the likelihood that T is the correct translation of S. This probability estimation forms the basis of the translation process (See Figure 2).

To address the challenge of Machine Translation, statistical approaches incorporate two essential components: the Language Model and the Translation Model. The Language Model, also known as the target language's language model, estimates the probability P(T) for a given sentence in the target language. On the other hand, the Translation Model estimates the probability P(S|T) of a source language sentence given a target language sentence. By combining the probabilities derived from the Language Model and the Translation Model, the joint probability of sentences S and T can be computed.

SMT encompasses different types, such as word-based and phrase-based approaches (refer to (Garg and Agarwal, 2018) for further details). In this case, we used a word-based approach, focusing on processing and aligning text at the word level. We also employed MOSES, an open-source toolkit that facilitates tokenization and language modeling,

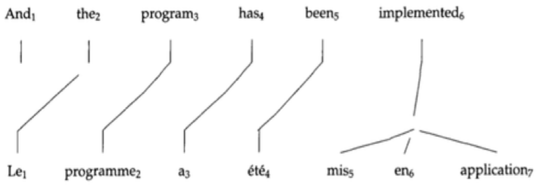

Figure 3: Sentences aligned using word alignment (Garg and Agarwal, 2018)

along with GIZA++ for word alignment. Proper alignment of text segments in parallel corpora is crucial to achieve accurate translations (See Figure 3).

Similar to the supervised approach, SMT benefits from the availability of parallel corpora, which tend to be more abundant in comparison to monolingual resources. Furthermore, SMT relies on interpretable statistical models, providing insights into the translation generation process. However, this approach requires precise alignments in parallel corpora and linguistic resources like language models and bilingual dictionaries. Obtaining and maintaining these resources can be costly and demanding in terms of additional effort.

4.4 Zero-shot with BERT-Multilingual

The term "zero-shot" in machine translation typically refers to the capability of a translation system to translate between language pairs for which it has not received explicit training. However, in our case, we have performed a zero-shot fine-tuning using the Bert-base-multilingual-uncased model.

Following the methodology discussed in class, we have implemented a zero-shot approach using our aligned parallel corpus dataset. To begin, we prepared the dataset by organizing it into individual columns: the source language (English sentences), the target language (French sentences), and labels. Random labels were assigned to the sentence pairs, which introduces a challenge during training since the training process is also randomized. Consequently, unlike the supervised method, increasing the number of epochs does not improve the results.

The Bert model, despite having been trained in different languages, becomes difficult for this task when trying to use its version of 'SequenceClassification' and adapt it to MT.

Zero-shot machine translation offers several advantages, including the potential for quick deployment of translation systems for new language pairs without requiring extensive training data. It also presents a cost-effective solution by reducing the need for training separate models for each language pair.

However, zero-shot machine translation may face difficulties in achieving the same level of translation quality as language-specific models. In our case, the absence of a corpus with predefined labels poses challenges in obtaining satisfactory results. To employ this method effectively, the language pairs would need to be labeled beforehand to facilitate training.

5 Results

In order to quantitatively analyze the MT output of the selected MT systems, we employed the 'bleu' score (with the exception of the zero-shot approach, were we used accuracy for evaluation). Table 2 summarizes our results regarding the 'bleu' score.

	Training	**Test**
JoeyNMT	28.05 (one epoch)	8.79
Helsinki-NLP	43.83 (one epoch)	16.29
Moses	-	9.56 - 10.27

Table 2: 'Bleu' scores for different MT systems, on training and testing subsets respectively.

It is important to note that the results of the Moses model indicate both the 'bleu' score resulting from testing the model after training (9.56) as well as the 'bleu' score resulting from testing the model after optimization with the development set (10.27).

During the training process of the Zero-Shot approach we have used 'Accuracy' as a evaluation method and not the 'bleu' score. In addition, we have trained it only with 6 epochs, since after this the system does not improve. The best result obtained during training was 0.368421, and using the subset test 0.4167.

On the other hand, the qualitative analysis that we will present in the next section will focus only on the translation of the test set produced by the fine-tuned Helsinki-NLP. We have taken this decision because that is the model which provided the best results when testing.

We have made available the notebooks with the code applied for obtaining the previous 'bleu' scores and accuracy; the translations of the MT systems; as well as the test set of court cases from the ECHR in section 8 "Supplemental Material". Nevertheless, the results of the SMT system were obtained by executing commands using SSH and we do not have a notebook with code for this system. In order to get these results, we followed the tutorial provided by the professor Gorka Labaka for the subject of Machine Translation of the Master in Language Analysis and Processing at UPV/EHU, adapting it to our datasets.

6 Analysis

In this section, we carry out a comparative evaluation of the MT output of the selected MT systems, both quantitatively and qualitatively.

6.1 Quantitative evaluation: ’bleu’ score and accuracy

The results indicated in the previous section show that the best performing system was the fine-tuned Helsinki-NLP, with a ’bleu’ score of 43.83 after training it for one epoch, and a ’bleu’ score of 16.29 on the test set.

According to the information available in the Google Cloud guide on ’Evaluating Models’ (Cloud, n.d.), ’bleu’ stands for ’BiLingual Evaluation Understudy’ and it "measures the similarity of the machine-translated text to a set of high quality reference translations". This Google Cloud guide (Cloud, n.d.) also provides a scale for interpreting ’bleu’ scores. In this scale, a ’bleu’ score between 40 and 50 (such as the one that we obtained when training the fine-tuned Helsinki-NLP) implies that the translation has high quality. However, a ’bleu’ score between 10 and 19 (such as the one that we obtained when testing the fine-tuned Helsinki-NLP) indicates that it is hard to get the gist of the translated text. Furthermore, a ’bleu’ score of less than 10 (such as the one that we obtained when testing JoeyNMT and Moses) points to almost useless translations according to the aforementioned scale.

In light of these interpretation of the ’bleu’ score, it could be argued that our results for the test set are not satisfactory, even for the best performing model. Nevertheless, it is important to take several aspects into consideration. Firstly, we only trained the selected NMT models for one epoch, due to time and resource constraints (as an example, this training took three hours and a half for the fine-tuned Helsinki-NLP). This means that there is a lot of room for improvement, and training these models for more epochs could boost performance. Secondly, even if both the training and test sets belong to the legal domain, their content differs in important ways. As it was indicated in section 3 "Selected datasets", the training set (Europarl corpus) contains topics discussed in the European Parliament, while the test set (court cases from the ECHR) focuses on judgements. This implies that the vocabulary and style of the training and test sets are not exactly equivalent, and this affects the results when testing the MT systems. Lastly, the size of our test set (340 sentences) was small if compared to the size of the training set (around two million sentences for the monolingual sets, as indicated in section 3 "Selected datasets"). Since we prepared the test set ourselves by downloading the texts, segmenting, and aligning them, the reasons of its small size are again our time and resource constraints. The small size of the test set could have an impact on the reliability of the testing performance and further research could carry out more experiments by increasing its size.

Regarding the accuracy obtained with the zero-shot approach as mentioned above, this is a random process, so the results obtained varied as the training was repeated. The absence of a dataset with previously adjudicated labels has hindered the possibility of obtaining a better performance. For this type of approach, it might have been more feasible to use datasets of different domains or topics so that the system could train on the basis of different labeled data according to the type of topic to which they belong. In this way it would be possible to obtain more deterministic results working with Zero-shot.

6.2 Qualitative evaluation

As we saw in the previous section, the translations of the test set produced by JoeyNMT and Moses are almost useless, according to the intepretation of the ’bleu’ score provided in the Google Cloud guide on ’Evaluating Models’ (Cloud, n.d.). Since the translation produced by the fine-tuned Helsinki-NLP has higher quality (even if the ’bleu’ score is still low) we will center our qualitative analysis on this last translation of the test set. More precisely, due to the processing limitations of Google Colab, we could only translate and therefore analyze one of the texts contained in the test set, but we consider that it will be enough to observe the most common errors.

For a qualitative evaluation of the translation obtained by the system, we will make use of the criteria proposed by Prof. Nora Aranberri: fluency, adequacy, and typology errors. In this regard, we believe that the translated text is surprisingly fluent and we would rank it with a 3 (good) in a rank from 1 to 4 because there are some minor erros but the texts flow smoothly. Relating the low ’bleu’ score with this surprising fluency, we consider the reason to be the use of synonyms and different verb tenses in the translation. This means that the translated text is fluent and comprehensible, but the vocabulary and grammar do not match exactly and this causes the ’bleu’ score to be low. We can see this happen in the following example (symbols omitted for clarity): ’Le demandeur a plaidé non coupable le 2 août 2002 et le procès a été reporté devant la cour de la Cour de justice Stoke-on-Trent le 7 octobre 2002.’ (MT) vs. ’Il plaida non coupable le 2 août 2002 et le procès fut renvoyé au 7 octobre 2002 devant la Crown Court de Stoke-on-Trent.’ (ref. translation).

In this vein, we consider that adequacy is also surprisingly high for such a low ’bleu’ score. We would also rank it with a 3 (most) in a rank from 1 to 4 because the translation contains almost all the meaning in the source, as in the previous example.

Regarding typology errors, we found that the most common errors related to terminology and accuracy (e.g. ’l’avocat de la couronne s’est adressé à la Cour’ [MT] vs. ’le ministère public (Counsel for the Crown) s’adressa au tribunal’ [ref. translation]).

In sum, in light of the selected translation, it seems that the low ’bleu’ score was due to the inability of this metric to entirely capture the nuances related to the quality of the text. After this qualitative analysis, we argue that a translated text with this quality could be helpful for translators working in their non-native language. Additionally, with some post-edition, the translation could also be useful for people involved in court cases when they do not speak the language of the court authorities because it conveys most of the meaning of the source. Moreover, such a translation could serve the previously indicated purpose of being used as synthetic data for increasing the size of a corpus of legal texts in French language to train and test NER models for anonymization.

7 Conclusions

The findings of this study suggest that the optimal system for translating legal texts would involve fine-tuning a model such as Helsinki-NLP. It is important to note that while the datasets used in this study were designed for SMT approaches, the results achieved through supervised training with JoeyNMT using only one epoch were comparable to those of SMT. We believe that with greater resources and time, this system could have surpassed SMT in terms of performance.

Furthermore, it would be beneficial to conduct further research on the use of multiple datasets in the legal domain, as presented in this study and by (Ralf Steinberger and Gilbro, 2014), to obtain a comprehensive understanding of the best approach for NMT in legal translation.

In conclusion, it is important to acknowledge that this study did not encompass all available NMT methods, like unsupervised learning, crosslingual embeddings, back-translation (a method known for its effectiveness), and LSTM with or without attention. Consequently, certain models may appear superior to others within this benchmarking analysis, but may not necessarily reflect their comparative performance outside of this study. Therefore, future research should expand the range of NMT methods used in order to ensure a broader and more objective perspective. This will enable the creation of more valuable resources for the translation of legal texts.

8 Supplemental Material

The Supplemental Material of the present paper includes the Colab notebooks containing the code that we used to obtain the results evaluated in this paper (with the exception of the SMT system, as previously indicated). Additionally, it contains the translations of the test set produced by Helsinki-NLP, JoeyNMT, and Moses; as well as the test set of court cases from the ECHR segmented and aligned by us:

8.1 Colab Notebooks

Supervised NMT with JoeyNMT: https://drive.google.com/file/d/1xd_XIlj6Ysug-i2NDFtQ5chFYYHhTKBO/view?usp=sharing

Supervised NMT with (fine-tuned) Helsinki-NLP: https://drive.google.com/file/d/1ciaIQIpIM_Ph9ryWVmGCOjD6TmR0sWXg/view?usp=sharing

Zero-shot approach: https://drive.google.com/file/d/1meV1nUMXBPsgrCJtWExyIaiPVvnNvSCq/

view?usp=sharing

8.2 Translations of the test set

Translation of the (partial) test set by Helsinki-NLP: https://drive.google.com/file/d/1oYjMVXdgsEHpHOE-6z0HzprZ1JkzfeWi/view?usp=sharing

Translation of the test set by JoeyNMT: https://drive.google.com/file/d/1oHmp1XS555i-eUVRCutq7xbI2nTU7Apu/view?usp=sharing

Translation of the test set by Moses after initial training: https://drive.google.com/file/d/1WsjILXwHM1Mzw6cOPbKaGjInuSg0ehFp/view?usp=sharing

Translation of the test set by Moses after model optimization: https://drive.google.com/file/d/15aj0XSzU6QK6hXjRN3FD5HUjBk5T7bPy/view?usp=sharing

8.3 Test set of court cases from the ECHR

English test set: https://drive.google.com/file/d/1BVdrTASF1lbTcfcIKHu3DUeVIIh3w5u0/view?usp=sharing

French test set: https://drive.google.com/file/d/1lxrreq3SIWFcYQJS_9qiMj9IADriXpZN/view?usp=sharing

References

Europarl parallel corpus.

Google Cloud. n.d. Evaluating models.

Echr. Hudoc - european court of human rights.

Elliott Macklovitch Fabrizio Gotti, Guy Lapalme and Atefeh Farzindar. 2008. Automatic translation of court judgments. *8th AMTA conference*, pages 370–379.

Ankush Garg and Mayank Agarwal. 2018. Machine translation: A literature review. pages 1–12.

Lilja Øvrelid Anthi Papadopoulou David Sánchez Ildikó Pilán, Pierre Lison and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, page 1–31.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. Proceedings of Machine Translation Summit X: Papers:79–86.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Sravan Kumar. 2021. How to fine-tune pre-trained translation model.

Matthias Huck Marion Weller-Di Marco and Alexander Fraser. 2022. Modeling target-side morphology in neural machine translation: A comparison of strategies. *Center for Information and Language Processing*, pages 1–10.

Yury Muravev. 2020. Machine translation and legal tech in legal translation training. *SPBPU DTMIS '20: Proceedings of Peter the Great St. Petersburg Polytechnic University International Scientific Conference "Digital Transformation on Manufacturing, Infrastructure and Service"*, pages 1–7.

Jaakko Väyrynen Ralf Steinberger Najeh Hajlaoui, David Kolovratnik and Daniel Varga. 2014. Dcep -digital corpus of the european parliament. page 3164–3170.

Alexandros Poulis Manuel Carrasco-Benitez Patrick Schlüter Marek Przybyszewski Ralf Steinberger, Mohamed Ebrahim and Signe Gilbro. 2014. An overview of the european union's highly multilingual parallel corpora. page 1–18.

Anna Widiger Camelia Ignat Tomaž Erjavec Dan Tufiş Ralf Steinberger, Bruno Pouliquen and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06):2142–2147.

Kawa Mirza Salih. 2018. Difficulties and challenges of translating legal texts from english into arabic. *The Scientific Journal of Cihan University – Sulaimanyia*, pages 9–20.

Francisco J. Vigier-Moreno and Lorena Pérez-Macías. 2022. Assessing neural machine translation of court documents: a case study on the translation of a spanish remand order into english. *Revista de Llengua i Dret, Journal of Language and Law*, 78:73–91.

Zonghan Yang Gang Chen Xuancheng Huang Maosong Sun Zhixing Tan, Shuo Wang and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools. *AI Open*, page 5–18.