Data Augmentation to Low-Resource Languages Approach: Galician case

Author: Sofía García González

Advisors: Germán Rigau Claramunt and José Ramom Pichel Campos

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

February 2022

Departments: Computer Systems and Languages

Laburpena

En este TFM se han hecho dos propuestas de aumento de datos para el entrenamiento de dos modelos de traducción bilingües en Español-Gallego e Inglés-Gallego a partir de corpus paralelos del Portugués. Una de las propuestas para adaptar el portugués al gallego ha sido la transliteración y otra la traducción con RBMT Apertium. Los resultados obtenidos han sido prometedores en el par inglés-gallego y dejan la puerta abierta a futura investigación.

Palabras clave: Traducción Automática, Aumento de Datos, Transliteración

Abstract

In this Final Thesis, two proposals for data augmentation have been made for the training of two bilingual translation models in Spanish-Galician and English-Galician from parallel corpora of Portuguese. One of the proposals for adapting Portuguese to Galician has been transliteration and the other translation with Apertium RBMT. The results obtained have been promising in the English-Galician pair and leave the door open for future research.

Keywords: Machine Translation, Data Augmentation, Transliteration

Contents

List of Figures

List of Tables

1 Project definition

This final thesis consists on taking advantage of parallel corpora for Portuguese, in this case Spanish-Portuguese and English-Portuguese to convert the Portuguese part into Galician language and train a bilingual models to Galician. Following this strategy there are three main objectives. First of all, testing if data augmentation to Galician language taking advantage of Portuguese corpora is a valid strategy in machine translation. Secondly, this data augmentation is going to be done following two different strategies: transliteration and translation with a RBMT model. So, the second objective is to compare these two strategies and especially check whether a simple transliteration from Portuguese to Galician is valid to carry out data augmentation and train good bilingual models. Finally, the third objective is to test the quality of Galician language when it is translated or transliterated from Portuguese.

The reason why this third objective will be evaluated has linguistic reasons. Galician and Portuguese are romance languages with a common past, the *Galaico-Portugués*, a language spoken from the end of VIII to the middle of XIV centuries in *Gallaecia*, the medieval kingdom that occupied present-day Galicia, northern Portugal and part of Asturias and León. Since the Portugal independence in the s.XII and the establishment of the Portuguese standard, Galician and Portuguese began to evolve as distinct languages. The extension of the territory in which *Galego-Portugués* was spoken during X century can be seen in figure 1[1]. Whereas Portuguese evolved since s.XII until nowadays as a standardized language, Galician did not have a standard until the eighties of the past century. This phenomenon and the influence of *Castilian*, the official language in the territory, made that Galician gradually became castilianised until today. Nowadays, and especially since the *standard* creation, many syntactic structures and vocabulary shared between Galician and Portuguese are used less and less in Galician at risk of disappearing over time. This is why we think that the use of Portuguese corpora not only benefits the increase of data, but also the quality and linguistic perpetuation of Galician forms that are gradually falling into disuse even though they have been used naturally at a popular level until a few decades ago.

The structure of this final thesis will be the following. In section 2 antecedents and related work will be covered. First the evolution of machine translation and then the situation in this field of low-resource languages in general and Galician in particular. Then, in section 3 the methodology followed to carry out the different experiments and strategies will be itemised. Moreover, in the following section, 4 different approaches that this thesis shows to this field will be presented. And, finally, in sections 5 and 6 the results and the conclusions and future work will be exposed and developed respectively. And, finally, a last section about future work to do around this topic will be seen in section 7.

[1] Source: https://es.wikipedia.org/wiki/Diferencias_entre_el_gallego_y_el_portugues

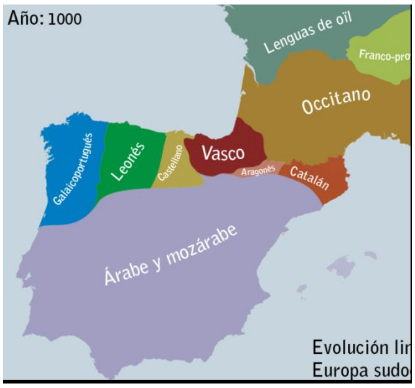


fig. 1. *Galego-Portugues* extension in s.X A.D.

2 Antecedents

In this section, previous work made in machine translation will be exposed. Firstly, machine translation evolution and different achievements in this specific field of Natural Language Processing (NLP) will be briefly covered 2.1. Furthermore, previous Machine translation investigation in low resource languages 2.2 and the specific situation in Galician language 2.2.2 will also be explained.

2.1 Machine Translation Evolution

As Forcada (2020) mentions:

> Machine translation can be defined as any use of computer systems to transform a computerised text written in a source language into a different computerised text written in a target language, thereby generating what is known as a raw translation.

In comparison to human translators, machine translation has two main advantages. Firstly, it can manage a huge amount of data in less time and, moreover, it can do that at a lower cost. Furthermore, machine translation devices can help their users to understand texts, videos... in languages that they do not know, making it easier the understanding between speakers of different languages. Moreover, many companies use these systems to translate their information or products to be able to open up to a wider public. But, in spite of that, machine translation has its own limitations too. The main ones are, firstly, that although in recently years new machine translation systems are getting much better results, raw translations are poorer than human ones. Secondly, they still have problems with figurative meaning making machine translation to specific domains or genres so poor, as for example, poetry. And, finally, recent systems as neural machine translation (NMT) need a huge amount of data to train their models, thereby, minority languages, such as Galician, are at a significant disadvantage in this field.

Regarding to machine translation system (MT) types, since the late forties to nowadays, two types of MT systems can be distinguished. Firstly, from the forties, in the MT beginning to the nineties the main approaches were based on rule-based machine translation (RBMT). This approach does not need previous corpora in source and target languages to be able to translate. The structural and grammar rules are the ones that make the translation possible. Later, from the eighties to nowadays, systems based on corpus-based machine translation (CBMT) have raised. These systems need parallel corpora in which source and target corpus have thousands or millions aligned sentences. The first main types of CBMT were: Example Based Machine Translation (EBMT) and Statistical Machine Translation (SMT). Since the end of the eighties until some years ago SMT, which uses probabilistic models to count the number of times certain occurrences appear in the bilingual training corpora, have been used by both companies and researchers. But since 2016, NMT has achieved the state-of-the-art in many pair of languages. This technology is based, in a very superficial way, on how the brain learns and process the information. (Forcada, 2020). In figure 2[2] a machine translation evolution scheme can be seen. In the following subsections each machine translation type will be explained in depth.

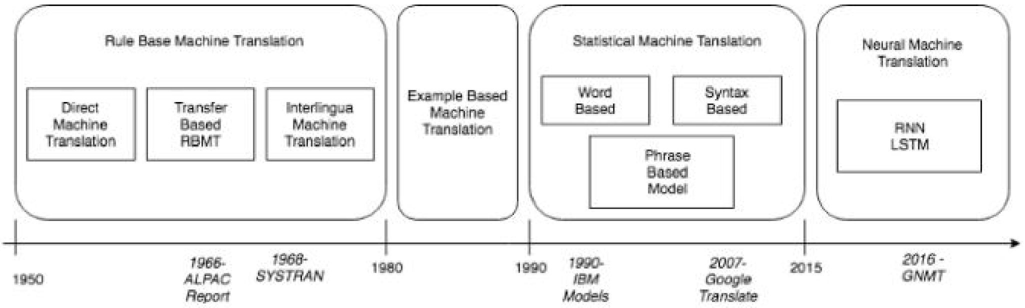

Rule Base Machine Translation
Direct Machine Translation
Transfer Based RBMT
Interlingua Machine Translation
Example Based Machine Translation
Statistical Machine Tanslation
Word Based
Syntax Based
Phrase Based Model
Neural Machine Translation
RNN LSTM
1950
1966- ALPAC Report
1968- SYSTRAN
1980
1990
1990- IBM Models
2007- Google Translate
2015
2016 - GNMT

fig. 2. Machine Translation Evolution

2.1.1 Rule-Based Machine Translation

As it has been previously said, RBMT are systems based on linguistic rules. Because of that, linguist professionals are necessary to maintain them. Furthermore, they can be very time consuming and, as languages are constantly changing and new terms or structures appear, they need a continuous maintenance. RBMT were the first type of machine translation systems and are still used as, in spite of the time necessary to maintain them, they are still very accurate to some language pairs, specially between similar ones (Bayón and Sánchez-Gijón, 2019).

RBMT systems are divided into three different types: (i) direct, (ii) transfer and (iii) *interlingua*. These different systems can be graphically seen in the famous Vauquois Triangle

[2] Source: https://towardsdatascience.com/evolution-of-machine-translation-5524f1c88b25

showed in figure 3[3].

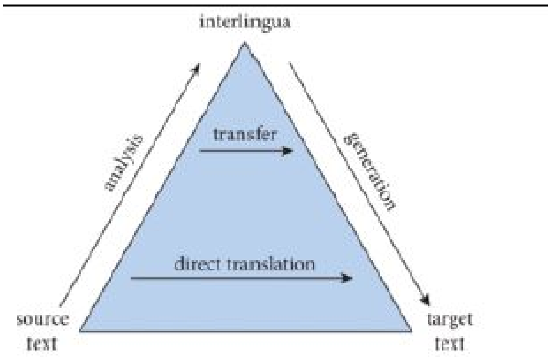


fig. 3. Vanquois Triangle

Direct translation consists mostly of word-by-word translation. The main tool of this strategy is a bilingual dictionary and some morphological information about the two pair of languages to translate. The sentence structure from one language to the other can be changed but, at the end, this system is not developed enough to achieve good translations in very distant languages, as the translation tend to be so literal and has a lot of limitations. As L.Specia and Y.Wilks (2021) mentions, it was a good start, but more advanced systems were necessary to RBMT development.

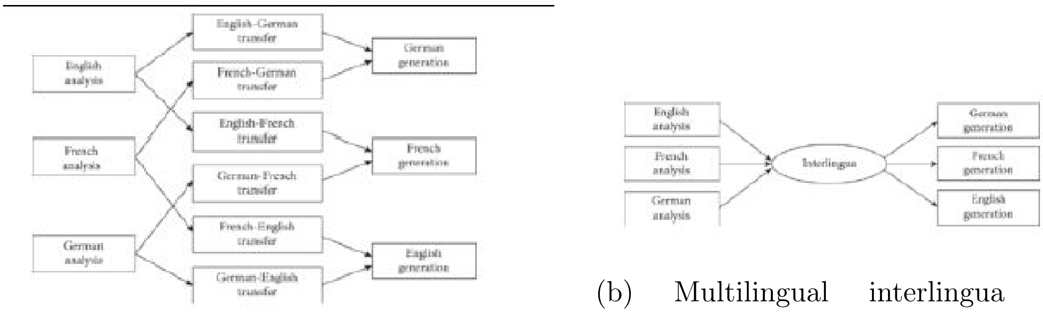
On the other hand, transfer systems are more sophisticated than direct translations. This system type makes rules to codify contrastive knowledge about differences between the pair of languages of interest. The three main components in transfer systems are: (i) Analysis and Rules to convert source text into syntactic or semantic representation; (ii) Transfer. Rules to convert source representation into target representation of the other language and (iii) Generation. Rules to convert the abstract representation of the target language into text (L.Specia and Y.Wilks, 2021). Transfer RBMT is the system used in *Apertium* (Armentano Oller et al., 2007)[4] that has many romance languages pairs, including Spanish-Galician and Portuguese-Galician.

Finally, *interlingua* system allows to produce multilingual machine translation in a simpler way than transfer. It is similar to the previous system but the analysis of source and target languages are independent between them as well as the generation. So, it allows to translate from a language to many others in an easier way. Because of that, as it can be seen in figure 4[5], multilingual *interlingua* systems are simpler than the transfer approach. This is because transfer multilingual systems need two modules for each language pair whereas *interlingua* only needs a module for each pair of languages.

[3] Source: L.Specia and Y.Wilks (2021)

[4] As this is one of the systems used in this project, it will be deeply explained in methodology section, 3.1.1

[5] Source: L.Specia and Y.Wilks (2021)



(b) Multilingual interlingua approach(three lan-guages, six modules)

(a) Multilingual transfer approach(three languages, twelve modules)

fig. 4. Transfer and Interlingua systems

2.1.2 Example-Based and Statistical Machine Translation

Example-based and statistical machine translation are similar systems. Both are corpus-based and need sentence alignment[6]. On the one hand, Example-Based Machine Translation(EBMT) is considered the first corpus-based approach and it is an intermediate stage between RBMT and SMT. The inputs of this system are not complete sentences, but small parts of them. Moreover, the run-time process can be divided into three steps:

1. **Matching:** Look for fragments of the sentences and their corresponding translation.
2. **Extraction:** The process extract the corresponding translations from the target side of the example corpus.
3. **Recombination:** The process to combine the previous matches to produce a complete text. The recombination process can be done at a character, word or pattern level.

This type of MT needs a lot of computation effort and the results were not specially good, so EBMT was never marketed. It does fine in very specific domains but SMT systems became much better some years later. Some examples of EBMT systems are CMU-EBMT (Brown, 2011), CUNEI or OpenMaTrEx (L.Specia and Y.Wilks, 2021).

On the other hand, SMT aims to create from the parallel corpus probabilistic models using statistical methods. It appeared from the first time in the nineties (Brown et al., 1993) and was the state-of-the-art at company and research levels until the neural system arrival. Since its beginning until the last decade its strategies were evolving. All of them try to find the most probable translation (f) given a source text (e), as it can be seen in equation 1. The strategy in these systems is the following. After applying the Bayes algorithm the sentences of the corpus are decomposed in sub-problems that generate translation models, $P(f|e)$ and language models $P(e)$. The first ones are generated from a parallel corpus

[6] All the information of this subsection was taken from L.Specia and Y.Wilks (2021) and Mandal (2019)

and calculates how probable is a translated sentence or a word based on the target corpus sentences. Additionally, the language models are learned from a monolingual corpus and estimates how good is a text in the target language based on the monolingual corpus. Once the translation and language model are initialized, the decoder searches, using the translation model, the optimal combination in the translation and estimates the translation fluency from the language model. One example of SMT architecture can be seen in figure 5[7].

$$e* = argmaxP(e|f) \tag{1}$$

At the beginning, (Brown et al., 1993) created what was later called the IBM alignment models. These models were word-by-word alignments, so the SMT model was based on word by word probabilities. Another two strategies that started at the beginning of the second millennium in SMT were phrase-based and sytanx-based models. The difference between them is that phrase-based systems are based on n-grams, which did not necessarily mean that they were based on grammar sentences, as n-grams could be based on any combination. Whereas syntax-based tries to make alignments taking into account syntax models as trees or context free grammars. The phrase-based models were the state of the art until 2016 and many companies as Google or Microsoft used them for their translation systems. One example of Phrase-Based Statistical model in Galician language is the Fernández et al. (2010) for English-Galician pair.

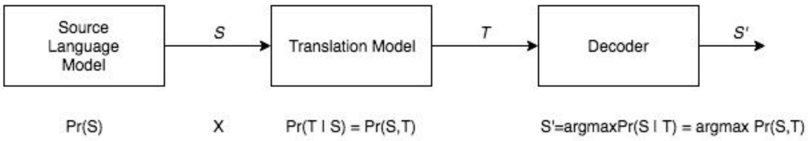


fig. 5. Word-by-Word SMT system

2.1.3 Neural Machine Translation

Although NMT has been massively used by researchers and companies since some years ago, the first attempts were made much earlier. Forcada and Ñeco (1997) and Castano et al. (1997) did the first approaches to Neural Machine Translation in the nineties although it had to evolve a lot until get nowadays performance. Forcada and Ñeco (1997) used two feed-forward neural networks[8] called the encoder and decoder[9]. This system took the strings symbol by symbol and the encoder did the internal representation. The same year,

[7] Source: (Mandal, 2019)

[8] In feed-forward neural networks there are an input layer, one or more possible hidden layers and an output layer. In this layer the vectorized information is always multiplied by the weights.In this first type of neural networks the information always moves in one direction.

[9] One example of feed-forward neural network can be seen in figure 6

Castano et al. (1997) used a recurrent network in which one word was taken by a time and produced the target word that was associated to the network with the maximum value. But the lack of data and resources at that time made these first approaches to fail (Mohamed et al., 2021). Some years later, Schwenk (2012) used a neural translation model in phrase-based SMT system. It learnt the probabilities of sentence pairs in this phrase-based system and achieved a development in English-French results. In spite of that, it was not until 2013 that, as Mohamed et al. (2021) mentions, the first pure NMT models were created by Kalchbrenner and Blunsom (2013). They used two recurrent continuous translation models. The first model used convolutional layers to produce the sentence representation whereas the second one estimated the length of the target sentence. This system generated firstly a four-gram representation of the source sentence and from it created a representation of the target sentence of which the second model had predicted the length (Mohamed et al., 2021). Two years later, in 2014, Sutskever et al. (2014) presented a Long Short Term Memory to translate sentences from English to French. This approach got close to the state of the art of the phrase-based machine translation at that moment. In the same year, Cho et al. (2014) proposed a recurrent neural network and a gated recursive convolutional neural network. It did well in short sentences with unknown words but the longer the sentence, the worse the results. Finally, in the IWSLT 2015, NMT was able to overcome the state of the art of SMT in English-German, a hard language pair because of their morphology and grammatical differences (Mohamed et al., 2021). Nowadays, Bidirectional Encoder Representation of Transformers (BERT) is the google's state of the art language modelling architecture based on transformers (Jagtap et al., 2020). Furthermore, multilingual models are giving very good results in many language pairs.

Regarding to the NMT architectures, they can be broadly divided into: Recurrent Neural Networks, Sequence to Sequence and, finally, Attention Mechanisms as, for example, Self-Attentional Transformers (Jagtap et al., 2020) and (Mohamed et al., 2021).

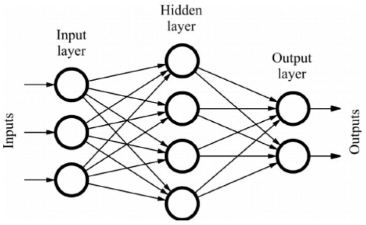


fig. 6. Feed-Forward Neural Networks Architecture

1. RECURRENT NEURAL NETWORKS

   In current neural networks the inputs are multiplied by respective weights to obtain hidden layers and then the output. The vanilla neural network maps the inputs and the outputs establishing a relation between them, but the outputs are the functions of current inputs. On the contrary, Recurrent Neural Networks introduce the concept of "memory" as they take into account previous hidden layers. Because of the Vanishing

Gradient[10] problems that these systems gave, the Long Short-Term Memory RRN systems were used. LSTM can "remember" the important previous information and "forget" the invalid one (Jagtap et al., 2020). An example of a Recurrent Neural Network cell can be seen in figure 7[11]. As Jagtap et al. (2020) mentions, RNN can be considered a word-to-word translation using Neural Networks.

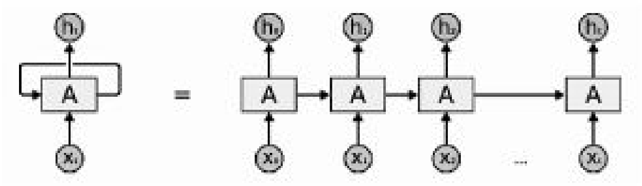


fig. 7. Recurrent Neural Network Cell

2. SEQUENCE TO SEQUENCE NEURAL NETWORKS

Sequence to Sequence models approach consists on architectures take the whole sentence as an input instead of doing it word by word. Due to its encoder-decoder mechanism, the encoder will convert the input sentence into a vector and then, taking into account this vector, the target sentence will be decoded.

In figure 8[12] a vanilla sequence to sequence architecture can be seen. In this figure, the encoder and the decoder are both LSTM neural networks. Each cell is updated taking at a time step one word. This is the architecture that has the state of the art in NMT models. In spite of that, this system has problems to remember long sequences. (Jagtap et al., 2020). As a solution to this type of problems, attention mechanisms emerged.

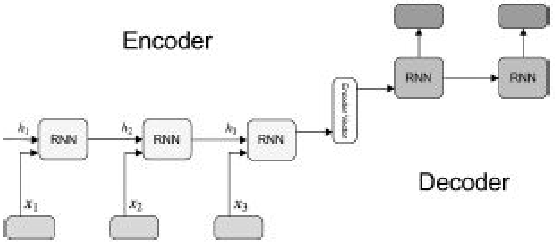


fig. 8. Sequence to Sequence Architecture

[10] The vanishing gradient refers to the fact that in a RNN the error signal decreases exponentially in the final layer. If the vanishing is too small, it prevents the weight to change its value, so it can give problems in the training process

[11] Source: Jagtap et al. (2020)

[12] Source: Jagtap et al. (2020)

3. ATTENTION MECHANISMS

   Attention mechanism is based on what to pay attention to and not how much information remember. It means that all the intermediate context vectors are important to the final context vector and not only the last one. (Jagtap et al., 2020). *Bahadannau* Attention was the first attention method used. It consisted on taking a weighted sum of all intermediate context vectors from all the time-steps (Jagtap et al., 2020).But,as this method needs computational complexity and power, which makes it useless, self-attention mechanism emerged. Self Attention, instead of using hidden states as inputs, uses the inputs directly. So it allows to do the process in parallel instead of step by step (Jagtap et al., 2020). Transformers are an example of self-attention mechanism models.

   Self-attentional Transformers use a mechanism in which a representation of a sequence is generated by relating different positions of it. Self-attentional transformers also learns intra-input and intra-output dependencies. Both encoder and decoder consist of six identical layers. Encoder layer contains two main components: a multi-head self-attention mechanism and a position-wise feed forward network. On the other hand, each decoder layer contains three components: a masked multi-head self-attention mechanism, a multi-head self-attention mechanism and a position-wise feed forward network (Jagtap et al., 2020). In figure 9[13] the transformer architecture can be seen.

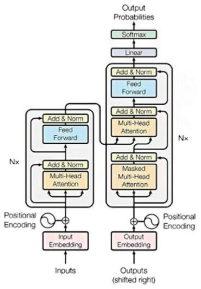


fig. 9. Self-Attentional Transformer Architecture

2.1.4 Evaluation Metrics in Machine Translation

One of the main difficulties in machine translation field is the evaluation. The main trouble in this task is that in language, and especially in the translation field, there is more than one possible answer. Many times more than one translation can be correct, so even differences with respect to the gold standard[14] not necessarily have to be errors.

[13] Source: Mohamed et al. (2021)

[14] Gold standard are translations made by linguists or translators that can be considered as the reference to the machine or translator to compare with the translation produced by the machine. They also can be real texts.

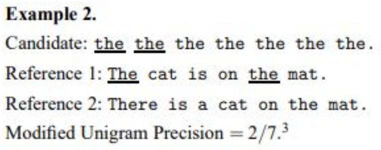
Machine translation evaluation metrics can be divided into human and automatic ones. As can be imagined, human evaluations are more reliable. This is because the human knowledge of language is more diverse than that of a machine, allowing it to decide whether a translation is correct beyond the gold standard. But, on the other hand, the human correction of huge amount of data would need a lot of effort, time and specialised people making it uncompensated. Regarding to automatic evaluation, it is much more faster making it more useful, but, in spite of that, it usually need human post-verification as it is not as reliable as human one (L.Specia and Y.Wilks, 2021).

The most important human evaluation metrics are: error analysis, attribute evaluation, pair-wise comparison or ranking and productivity tests. Error analysis consists on looking for errors that are rated according to the type, frequency and severity. Furthermore, attribute evaluation is usually divided into adequacy and fluency and it is rated in a scale from the worst to the best adequacy and fluency translation. Finally, comparison is the action of decide between more than one possible translation to choose the best one (Aranberri et al., 2017).

On the other hand, the most common automatic evaluation metrics are: BLEU, TER, chrF, ROUGE and COMET.

1. **BLEU**

   BLEU (Bilingual Evaluation Understudy) released in 2002 is the most used evaluation metric because as it is easy to use, fast and very correlated to human evaluations (Papineni et al., 2002). It solves the current problem that precision metric gave. Precision was not a good metric to evaluate machine translation as it calculated if the different words of the generated sentence were present in the reference one and divided it by the number of words that were in the reference sentence. But sometimes this metric could give good scores to sentences that made no sense. BLEU approaches a modified precision in which the system counts how many times this word appears in the reference sentence and calculates how many times this word appears in the translated one in comparison to the times it appears in the reference. This score is more reliable than precision as it can be seen in figure 10. Using current precision the translated sentence would have $7/7 = 1$ an hypothetical perfect translation. But using the modified unigram would have only a $2/7 = 0.28$. So, the BLEU score is more accurate than precision. BLEU can be calculated depending on different n-grams size. In spite of being one of the most used evaluation metrics,



Example 2.
Candidate: the the the the the the the.
Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.
Modified Unigram Precision = 2/7.[3]

fig. 10. Precision and BLEU score differences

   it has some inconveniences. For example, it does not take into account morphology differences and only the exact forms of words are considered as correct. Furthermore, BLEU scores are very conditioned to the tokenization and preprocess. The use of different tokenizers can vary a lot the BLEU scores, so many times the comparisons between different projects are not very reliable. Because of that in 2018 was released sacreBLEU, a python script which expects detokenized outputs to apply its own preprocessing and tokenization process. Doing so, the results are more comparable (Post, 2018). Nowadays is considered the reference value to WMT. The scores of BLEU goes from 0 to 100 and they can be interpreted as[15]:

- <10: unuseful
- 10-19: Difficult to understand the meaning of the message.
- 20-29: The meaning is understandable but there are significant errors.
- 30-40: They can be considered good translations.
- 40-50: High quality translations.
- 50-60: Very high quality translations. .
- > 60: Quality better than human.

2. **METEOR**

   METEOR (Metric for Evaluation of Translation with Explicit Ordering) is an automatic metric proposed in 2004 to solve some problems that BLEU gave (Banerjee and Lavie, 2005). It includes fragmentation score and accounts to word order, token matching considering stemming (the root of the word), synonymy, paraphrase and is optimize with weight-scoring to correlate it with human evaluation. It takes more into account the Recall in the contrary to BLEU score. If it has more than one reference, it calculates the correlation between all the references available (L.Specia and Y.Wilks, 2021). It works well at sentence-level, having a 0.964 correlation with human evaluation but it performs worse at corpus level giving a 0.4 correlation with human evaluation.

3. **chrF**

   Character *n-gram* score is an automatic evaluation metric proposed in 2015 (Popović, 2015). Some advantages of this metric is that it is language and tokenization independent. Moreover, it has a high correlation with human-evaluation. The chrF formula can be seen in figure 11[16]. In this picture, chrp represents the percentage of *n-grams* in the hypothesis that have a counterpart in the reference. On the other hand, chrP is the percentage of character *n-gram* in the reference which also take part in the

[15] Source of score meaning: `https://cloud.google.com/translate/automl/docs/evaluate`. In spite of that, it is important to take always into account that metrics are only for guidance, so even when the results in BLEU are higher than 60 it is known that it does not always mean a translation better than a human one.

[16] Source: Popović (2015)

hypothesis. Finally, *beta* is a parameter that assigns *beta* times more importance to recall than to precision. Recent experiments have shown that F-score at *n-gram* level is more accurate than the metrics at word level.

$$\text{CHRF}\beta = (1+\beta^2)\frac{\text{CHRP}\cdot\text{CHRR}}{\beta^2\cdot\text{CHRP}+\text{CHRR}} \qquad (1)$$

fig. 11. chrF formula

4. **TER/HTER**

   TER (Translation Edit Rate) metric is based on the original WER (Word Error Rate) for automatic speech recognition. TER calculates the number of deletions, insertions and substitutions to be made to adequate the generated translation to its reference. And, moreover, its derived, HTER (Human Translation Edit Rate) consists on taking as reference human-corrected versions of the machine translation. So, HTER is a semi-automatic evaluation process as it needs a post-editing phase made by humans. TER and HTER were first proposed in 2005 as another approach to machine translation evaluation (Snover et al., 2006). TER metric, as all the previous ones except to BLEU, goes from 0 to 1 but in this case, the lowest the result, the better the translation, because a low result means that the changes that would be necessary to do are not so much.

5. **ROUGE**

   ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used both to machine translation and summary text generation evaluation (Lin and Och, 2004). ROUGE calculates the unigram precision, recall and F1-score to have an overall evaluation combining different metrics.

6. **COMET**

   COMET (Crosslingual Optimized Metric for Evaluation of Translation) (Rei et al., 2020) is the actual state-of-the-art metric evaluation in machine translation since the WMT 2019 Metrics. It is a Py-torch neural framework that allows to train multilingual machine translation evaluation models. The goal of this framework is to automate the evaluation to improve the efficiency and speed.

In this subsection the most important metrics have been explained but not all of them are going to be used to evaluate the performance of our models, only three of them have been chosen: BLEU, chrF and TER. This is because they are the most common and they measure different aspects of the translation so the evaluation is more reliable.

2.2 Machine Translation in Low Resource Languages

Low resource languages, as Galician, have always been one of the main *handicaps* of machine translation, especially since the SMT and NMT arrival. The main problem with this type of systems is that a huge amount of parallel data is necessary to train their models, so, languages with low resources (LRL) give poorer results than high-resource ones (HRL) . Even methods as unsupervised machine translation (UMT) (Lample et al., 2017) and Back-Translation (Edunov et al., 2018) that has demonstrated to be very efficient for data augmentation of big languages as English-French and English-German pairs respectively, also have had poor results in LRL because of data scarcity (Xia et al., 2019), (Artetxe et al., 2018). UMT consists on using monolingual data from the source and the target languages and train a model encoding sentences of both languages in the same feature space(Lample et al., 2017) to be able to translate from source to target language. Regarding Back-Translation, it consists on translating monolingual data of the desired target language using a base-line model. Doing it, parallel data is created being the source language synthetic. Once again, data scarcity, even in monolingual data, make these methods not useful to LRL. Because of that some researchers as Bayón and Sánchez-Gijón (2019) still highlight ideas as:

> The main conclusion is that although NMT seems promising in frequent language combinations, especially if English is involved, it is not obtaining the desired results in low-resource languages such as the pair Spanish-Galician. NMT has not yet unseated RBMT and PBMT, performing, in fact, worse than these systems.

NLP has maintained in different fields, as MT, the biases established in our society , in which minor languages and, as a consequence, their speakers, suffer from under-representation. In spite of that, different techniques to avoid this issue have been carried out during this time, above all to increase data through data augmentation approaches or multilingual neural systems.

2.2.1 Transfer Learning in Multilingual Models and Data Augmentation

The first approaches in NMT to develop LRL training models consisted on transfer learning methods (Zoph et al., 2016), (Gu et al., 2018) in multilingual models. The idea of this operation is training a first model in a High Resource Language (HRL) and then use these parameters to initialize the training process of a related Low Resource Language (LRL) through the parameters information sharing. The LRL training part works as fine-tuning of the previous trained model and have demonstrated to be useful to increase the accuracy. In spite of that, these operations require time and computational effort, so Neubig and Hu (2018) proposed training joined parallel data of HRL and LRL models taking advantage of the similar vocabulary and syntax to simplify the process. Neubig and Hu (2018) demonstrated that this methodology was effective at that moment, improving the previous results in this field. They mixed the transfer learning approach with a joint training between HRL and LRL parallel data in a multilingual context. They achieved good results in a 58 languages to English model, in which the pair HRL-LRL Portuguese-Galician was included. In spite of that, these methods are not ideal, as, although the languages can be very similar, they still have differences in syntax and vocabulary (Xia et al., 2019). Another approach to bilingual models that also takes advantage of similar language HRL-LRL pairs is to adapt the HRL to the LRL to be able to train a model in which the LRL is synthetic data. This is the approach that, indeed, has served as the basis and motivation of this thesis (Xia et al., 2019). It can be done training translation models (Xia et al., 2019) or transliterating the HRL into the LRL[17](Karakanta et al., 2018), (Ramnath et al., 2021). All of them increased, in 2019, the BLEU results compared to previous state-of-the-art achievements using some HRL-LRL language pairs as, for example, Portuguese-Galician, Bielorussian-Russian or Hindi-Gujarati respectively.

Finally, it is important to underline that in the past two years UMT and multilingual MT have developed their performances in low resource scenarios (Goenaga and Labaka, 2021). One example is the proposal of Fan et al. (2021) that have created a many-to-many translation model that can perform translation between each pair of these 100 languages not being English one of them. Indeed, Galician is one of the languages in this model. Their approach have shown a competitive performance respect to bilingual models in comparison to previous multilingual approaches.

2.2.2 Galician in Machine Translation

As it has been mentioned in previous sections, from the first types of machine translation until the neural one, there have been different projects involving Galician language, especially for Portuguese-Galician pair (Armentano Oller et al., 2007), Port2gal, Spanish-Galician (Armentano Oller et al., 2007), (Bayón and Sánchez-Gijón, 2019) and English-Galician (Fernández et al., 2010), (Xia et al., 2019). Moreover, Galician is also included in different multilingual systems (Wang et al., 2019), (Aharoni et al., 2019), (Fan et al., 2021). Even there have been cases of translation into other languages, such as (Leng et al., 2019) that proposed an unsupervised pivot method to translate between Danish- Galician and other distant languages.

But the intention in this subsection is to focus on the two previous projects that have been a model to the present thesis, as they also follow the strategy of taking advantage of Portuguese data to train English-Galician and Galician-English models.[18] These are the *Carvalho*(Fernández et al., 2010) and the method proposed by Xia et al. (2019) including another HRL-LRL pairs. In *Carvalho* the English-Portuguese pair from Europarl.v3 was used as the basis corpus and the Portuguese language was translated to Galician using *Apertium*. After doing that, some post-editing methods had to be done. For example, transliterating with Port2gal toolkit the vocabulary that *Apertium* was not able to translate. After that, they trained a PBSMT and achieved similar scores, although a bit lower,

[17] As transliteration and translation between HRL and LRL is one the main proposals of methodology in this thesis, the difference between these techniques will be better explained in section 3.1.

[18] In this case, this project expands this methodology to the Spanish-Galician pair.

than Google English-Galician translation at that year, in BLEU metric[19]. On the other hand, Xia et al. (2019) proposed a data augmentation method to train a model from several minor languages to English taking advantage of the HRL related to the specific LRL. These HRL-LRL languages were: Portuguese-Galician, Turkish-Azerbaijani, Russian-Belorussian and Czech-Slovak. Regarding the data, both monolingual in all the languages, and parallel data in HRL-LRL, HRL-EN and LRL-EN were taken from TED corpus and Wikipedia. Regarding the methodology, they proposed different approaches:

1. Standard Supervised NMT. They train a base model with the concatenation of LRL-English and HRL-English corpus. Then, they fine-tune the model on the concatenation of the base and augmented datasets (Xia et al., 2019).

2. Standard Unsupervised MT model from the the LRL data, which, as was previously mentioned, gave extremely poor results.

3. Standard Supervised Back-Translation using on the one hand EN-LRL dataset and, on the other hand, EN-HRL dataset. EN-HRL gave, as might be expected, better results in standard Back-Translation.

4. HRL to LRL adaptation from an EN-HRL dataset through vanilla supervised and unsupervised machine translation models. This approach does not give significant better results.

5. HRL to LRL adaptation through word to word substitution. This is one of the approaches that gives better results. They obtain a bilingual dictionary by finding the optimal mapping between the source and target word embedding spaces. They exploit identical words between both languages and compute the distance between mapped source and target words with the CSLS similarity measure.

6. Unsupervised initialization modification between HRL-LRL. They use the previous induction dictionary method, learn joint word segmentation model in original monolingual data in LRL and synthetic monolingual data LRL created from the word substitution and train an UMT fashion model between the original monolingual LRL data and the synthetic one. This method gives in Galician language slightly better BLEU results than word to word substitution.

7. Modified UMT and word substitution mixed corpora to train a LRL-EN model. This method slightly outperforms the two previous approaches.

8. They tried again the word to word substitution on the one hand and the modified UMT on the other hand but now the HRL data comes from a back-translation from English data. This method gives slightly worse results but indeed outperforms the state-of-the-art of previous projects.

[19] In 2010 Google translator used PBMT

9. Finally, the best results came from word substitution mixing original HRL data and HRL data created from back-translation from English.

All the approaches in this paper using HRL-EN data to adapt it to the LRL related language outperformed the previous state-of-the-art in Galician-English bilingual models. Indeed, Portuguese-Galician pair gave the best results in all the tests together with Czech-Slovak pair (Xia et al., 2019).

Regarding Spanish-Galician machine translation, as they are related languages, RBMT and SMT have given, in general, better results than NMT and not many approaches in NMT have been done. Only the one by Bayón and Sánchez-Gijón (2019) that confirm, in 2019, her opinion that for related language pairs NMT has to develop a bit more to be really useful compared to SMT and RBMT. Nowadays, multilingual models in which Galician language is involved, are giving similar results to RBMT systems in some cases. In fact, in section 5 our models will be compared to some baselines that are multilingual models.

3 Methodology

In this section different resources and tool-kits used in this thesis will be explained. First of all, the difference between translation and transliteration will be covered 3.1, as well as the tool-kits used to do this process, *Apertium* 3.1.1 and Port2gal 3.1.2 respectively. Secondly, the corpus extraction and preparation, 3.2. Then, the tool-kits used to train the models 3.3, which are Fairseq 3.3.1 and OpenNMT-py 3.3.2 will be exposed and finally, the equipment used to train the models will be exposed 3.4.

3.1 Translation and Transliteration

One of the main purposes of this final thesis is to verify whether two specific methods are valid to carry out data augmentation in Galician from Portuguese language: Transliteration with Port2gal and translation with the RBMT system , *Apertium*. But, firstly, it is very important to clarify the difference between translation and transliteration, which, although similar, are not the same. Following the meaning that the Spanish Real Academy (RAE) gives[20], translation is to express in one language what is written or expressed in the other. Whereas transliteration is, following the RAE and the Oxford dictionaries, writing words or symbols of one language using the symbols or writing system of another. What this different definitions mean is that the aim of translation is to express exactly the same information from one language to the other. Many times the structures must be different, some terms have not a possible translation to reproduce exactly the same meaning in the other language with only word-to-word, so, some explanation is necessary, what, indeed, is the reason why machine translation is as difficult as it is. But, on the other hand,

[20] The reference is taken from the RAE as it was the best defined term in comparison to other English dictionaries

transliteration does not intend to be true to the source language meaning as this process consists on changing only orthography. When two languages are very similar, as Portuguese and Galician, a simple transliteration could be equivalent to a translation method. Some basic examples of symbol change would be: the *nh* of Portuguese to *ñ* in Galician or *lh* in Portuguese to *ll* in Galician. What usually happens is that Portuguese structures and vocabulary exists in Galician but it tends to be more formal and less used. So, producing texts in Galician from Portuguese could benefit to not to lose vocabulary and structures that are disappearing in Galician language, as it has been previously mentioned in section 1.

In table 1 there are two examples taken from the corpus English-Portuguese. Both sentences are written in Portuguese, Galician translated from *Apertium* and Galician transliterated. To both sentences be understood the English version is also written.

In the table, the vocabulary differences between Portuguese, translated and transliterated Galician are highlighted in bold. As can be seen in the first example, the only difference between the transliterated and the translated sentence is the word *montante* and *importe*. The transliterated text does not translate the vocabulary, so the word referent to "reference amounts" is the same as in Portuguese. But in the translated sentence, the word is translated to the correct form in Galician, *importe*. On the other hand, if the rest of the sentence is checked, there are no differences between the translated and the transliterated one. So, for words in Portuguese like *referência*, *comitologia* and *duraçâo* transliteration and translation methods are equally efficient. In the second example something similar happens. The word "Members", *colegas* in Portuguese does not change in the transliterated corpus. But in the translated one, the word is translated to the correct word in Galician, *compañeiros*. But, again, the rest of the sentence in the transliterated and translated corpus is the same. So, as it can be seen, transliterated and translated corpus are very similar and the differences can only be noticed in some specific words. A simple change of characters is very similar to a complete translation.

3.1.1 *Apertium*

As it was previously mentioned in section 2.1.1, *Opentrad Apertium* is a transfer RBMT open-source platform, although some modules are statistic or hybrids (Armentano Oller et al., 2007). It was created as an investigation project in which different universities and companies collaborated. At the beginning, it was created to translate between similar languages as, for example, Spanish-Catalan, Spanish-Galician, Czech-Slovak, etc. But since 2006 it also includes language pairs of different languages, as, for example, Spanish-English. Some of the similar-languages has 5 or 10 per cent of error rate. All the documentation and the tools necessary to execute them are available in their web page[21]. In spite of that, in this project the *Apertium* tools used were the ones available in Imaxin|Software company, as they are more developed. In figure 12[22], the *apertium* motor is shown. So the *Apertium* motor can be divided into: the de-formatter module, the morphology analyzer,

[21] `https://apertium.org/index.spa.html?dir=spa-eng&q=`

[22] Source: Armentano Oller et al. (2007)

Corpus	Sentences
Portuguese	As principais áreas de diferendo situaram se ao nível dos **montantes** financeiros de referência , da comitologia e da duração do regulamento .
Apertium	As principais áreas de diferendo situaron se ao nivel dos **importes** financeiros de referencia , da comitoloxia e da duración do regulamento
Port2gal	As principais áreas de diferendo situaron se ao nivel dos **montantes** financeiros de referencia , da comitoloxía e da duración do regulamento
English	The main areas of difference have been on the financial reference amounts , on comitology and on the duration of the regulation
Portuguese	Senhor Presidente , Senhor Comissário , muitos **colegas** já recordaram que o alargamento é um objectivo importante e ambicioso
Apertium	Señor Presidente , Señor Comisario , moitos **compañeiros** xa recordaron que o alargamento é un obxectivo importante e ambicioso
Port2gal	Señor Presidente , Señor Comisario , moitos **colegas** xa recordaron que o alargamento é un obxectivo importante e ambicioso
English	Mr President , Commissioner , many of the Members have already pointed out that enlargement is a major , ambitious goal

Table 1: Translation and Transliteration differences

the categorical lexical disambiguator, the structural and lexical transfer, the morphology generator, the postgenerator and the re-formatter. In addition to these modules, it includes compilers to transform linguistic data into a binary form used by each module. In fact, there is only one compiler to the lexical transfer module, the morphology generator and the post-generator (Armentano Oller et al., 2007).

- De-formatter (*desformateador*): It takes the text from the document format (XML or HTML).
- Morphology Analyzer (*analizador morfológico*): It takes the text and adds to each part of the text a *lemma*[23]. In the cases where there is a contraction, it splits it and

[23] A lemma is the root of the word. Equivalent to a dictionary entry

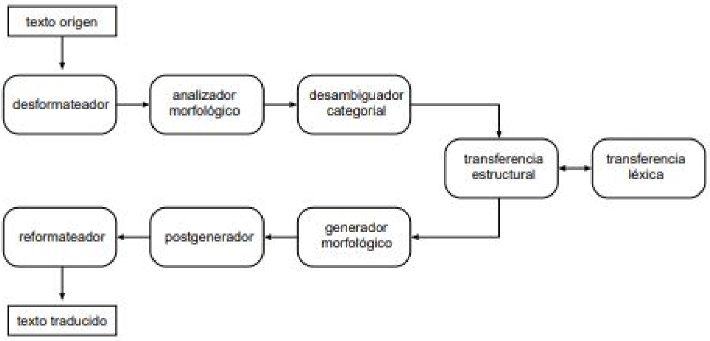


fig. 12. *Apertium* motor

extract the words that form the contracted word.

- The categorical lexical disambiguator (*desambiguador categorial*): In these common cases in which the same word can have different meanings or can represent different syntax categories, the disambiguator uses the statistical Markov model to choose the most probable one.

- The lexical transfer module (*transferencia léxica*: It uses a bilingual dictionary to take the most adequate equivalent in the target language taking into account the source one.

- The structural transfer module (*transferencia estructural*): It is the module in which the special phrases that involve changes from the source language, as the genre, the number, reordering, etc. are treated. For example, in Spanish the determinant has to coincide with the genre and the number of the noun and sometimes it changes in comparison to other languages as for example English, in which the determinant is invariable. So, this module carries out this type of analysis.

- The morphological generator (*generador morfológico*: This module takes the superficial lexic form in the target language and flexes the word in the correct form.

- The postgenerator (*postgenerador*): it deals with contracted and apostrophication forms.

- The reformatter (*reformateador*): It converts the target word again into the correct format (XML or HTML).

Although RBMT systems are less and less used, *Apertium* is still a platform in use and development. In recent years, some new modules have been created. For example, the *apertium-recursive* has been created in 2019 by Daniel Swanson as part of Google Summer of Code 2019 to deal with the structure differences in different languages avoiding the creation of many rules. On the other hand, the module *apertium-separable* deals with multi word expressions as can be the phrasal verbs in English. Multi-word expressions are complicated to a finite-state system as *Apertium*, especially when both parts are not together, for example, **take** this **away**. So, this module provides a framework to solve this kind of problems. Moreover, *apertium-separable* module was developed by Irene Tang as part of the Google Summer Code and handle contiguous and dis-contiguous multi word expressions. It receives a XML dictionary compiled into a finite state transducer. It maps the input text to the correct translation. Finally, the *apertium anaphora* module was developed by Tanmai Khanna as part of Google Summer of Code 2019 to deal with anaphora resolution. Which means, solving references to previous items or concepts in the same text. For example, if the text is referring to a man and later a personal masculine pronoun appears referring to this person it should be masculine and singular, such as he. This module uses Mitkov's algorithm to calculate probabilities in the text (Khanna et al., 2021). One example of this new architecture can be seen in figure 13.

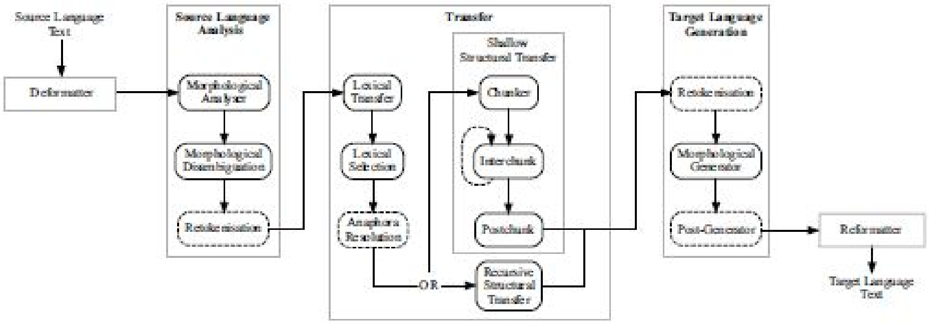


fig. 13. *Apertium* new modules

One of the main problems that *Apertium* gave in the Portuguese to Galician translation was that it misaligned some sentences in the process. So, the output file had four or two sentences less than the original Portuguese and English or Spanish files. As it would give problems in the training process, this issue had to be solved. Finally, tokenizing the Portuguese file and then adding a full stop at the end of each sentence before giving it to *Apertium* solved the problem. Furthermore, some post-edition after the translation was necessary. Although in previous projects Fernández et al. (2010) sent the words that *Apertium* did not recognize or did not translate to Port2gal to be then transliterated, in this thesis it was decided to do the post-editing process more manually. This is because even after sending the words to Port2gal some human post-editon was necessary. So, in this thesis a manual transliteration was done, especially to eliminate the letters that do not exist in Galician, such as ç or j[24]. With the *Notepad++* text editor help, some common terminations and letters in Portuguese that do not exist in Galician were replaced by the correct forms. Some of the main examples of this post-edition can be seen in table 2.

[24] In any was carried out word translations. The corrections were only in spelling or genre concordance. So, in this transliteration process some words could be still incorrect in Galician language

Moreover, when *Apertium* does not recognise a word in its monolingual dictionaries, a * symbol appears in the output file, when this word is not recognised in the bilingual dictionary it is marked with a @ and, finally, if there is a problem of concordance between both dictionaries, a # symbol highlights the word affected. So, all these symbols had also to be removed, as well as the final stop added in the previous step before the translation process. We know that in this corpus there are more mistakes to solve, especially because the Portuguese-Galician pair is one of the less developed in this *Apertium* version. This issue could be handle in future work.

Portuguese forms	Galician forms
ções	cións
çõe	ción
ço	zo
ça	za
ss	s
ô	o
ê	e
â	a
çóis	ós
lh	ll
j	x

Table 2: Galician translation post-edition

3.1.2 Port2gal

Port2gal is an open-source transliterator/translator tool-kit that also can be used online[25]. In this thesis the 2006 version was used, which is available to download in Pablo Gamallo webpage[26]. This tool-kit was originally created by Alberto García and developed by Pablo Gamallo. It consists on:

- A transliterator: It is a revised version from the original of Alberto García
- A set of verbal rules
- A small verb dictionary

The code is written in Perl and it can transliterate from Portuguese to both Galician

[25] https://gramatica.usc.es/ gamallo/php/translit/tradutorRAG.php

[26] `https://gramatica.usc.es/~gamallo/port2gal.htm`

language regulations: RAG and AGAL[27], in this case we chose the RAG regulation[28]. Port2gal did not gave the same unalingment issue that *Apertium* did, but it was necessary a small post-edition similar to the previous one, but not as much because in transliteration vocabulary is not really taken into account, as it was explained in transliteration section 3.1. An example of other NLP projects that have used Port2gal strategy is the one carried out by Garcia et al. (2016) that created a treebank of universal dependencies for Galician from Portuguese data achieving good and promising results.

3.2 Corpus

In this thesis two pair of languages models were trained. Spanish-Galician 3.2.1 and English-Galician 3.2.2, so parallel and monolingual corpus in both pair of languages were collected. It is important to highlight that finding appropriate parallel corpus was one of the most complex tasks in this project. This is because, as it was explained in section 1, although Portuguese and Galician are very similar languages, the Portuguese variety is not a superfluous issue, indeed it is very important. The main adversity was that most part of the bilingual parallel corpora available in the Internet both in Spanish-Portuguese and English-Portuguese mix European and Brazilian Portuguese varieties in the same corpus. One example is the *ParaCrawl* or Tatoeba corpus[29] and other big ones available in *Opus* web-page[30]. The main *handicap* of this situation is that these two varieties do not only have differences in vocabulary, but also in syntax. Although there are more, in table 3 an example of the two most common syntax differences is shown. In the table 3 the same sentence[31] is written in both Portuguese varieties, Galician and the translation in English. Syntax varieties in Portuguese from Brazil (Brazilian) and the correct form for Galician and Portuguese from Portugal (Portuguese) are highlighted in bold. As can be seen, in Portuguese and Galician there is an article before the possessive pronoun (o meu) whereas in Brazilian the possessive is alone (meu). The article before the possessive pronoun is mandatory in Galician except in very few cases, the absence of this article sounds very strange to a Galician speaker. On the other hand, the personal pronoun collocation after the verb, although has also some exceptions, is the correct form in most cases in Galician language. As it can be seen in the table, Brazilian tends to collocate it before the verb (nos oferece) whereas in Portuguese is also after the verb (oferece-nos). The differences in Brazilian are considered serious mistakes in Galician that can make the *Apertium*

[27] Although there is only an official regulation, RAG (*Real Academia Galega*), in Galicia exists another alternative and unofficial regulation (AGAL) which proposes grammar and vocabulary being more similar to Portuguese than the official regulation. For example, maintaining in Galician the 'nh' symbol instead of 'ñ' or 'lh' instead of 'll'. Although it is unofficial, there are some editorials that publish in this language regulation and it is used and defended by some academics.

[28] All the information about Port2gal is taken from the Pablo Gamallo webpage, `https://gramatica.usc.es/~gamallo/port2gal.htm`, as there is no paper published about this tool-kit

[29] Source: https://paracrawl.eu/

[30] Source: `https://opus.nlpl.eu/`

[31] This example is part of a sentence taken from English-Portuguese corpus. It was adapted to the Brazilian Portuguese and manually translated to Galician

translation and especially the *Port2gal* transliteration from Portuguese to Galician sound unnatural. Because of that, we decided to take corpus from the European Parliaments to be sure that the Portuguese was from Portugal as it would be very difficult to make sure that another type of corpus has or has not mixed varieties[32].

Languages	Sentences
Portuguese variety	oferece-**nos** a rara oportunidade de dirigir **o nosso** olhar
Brazilian variety	**nos** oferece a rara oportunidade de dirigir **nosso** olhar
Galician	ofréce**nos** a rara oportunidade de dirixir **o noso** ollar
English	gives us a rare opportunity to cast our eyes

Table 3: Brazilian and Portuguese varieties differences.

3.2.1 Spanish-Portuguese-Galician

The original Spanish-Portuguese corpus was taken from the UE Direction of General Translation (DGT) in 2007 for the *Acquis Communataire*. It is available in 24 languages of the European Union[33]. The DGT allows the use of a tool, TMXtract, available for Windows and Linux to extract the desirable language pair from the different zips available, in this case Spanish-Portuguese. To this corpus the DGT-TM-Release 2011 Spanish-Portuguese was extracted in tmx format. This tmx file was converted into two parallel txt files with a python script using the tmx library. These resulting files had 1.84M sentences and over 35 million of tokens. Some examples of the corpus in Spanish, Portuguese, Galician translated and Galician transliterated can be seen in table 4. Also, some examples of the test corpus can be seen in table 5. The split into train and valid sets was done in the data preprocess step explained in section 3.3.1.

[32] Indeed, the reason why we started to focus on this issue was that one of the first English-Galician models trained with Open-NMT-py and *ParaCrawl* corpus gave most of these mistakes in its results. As a result it was eventually discarded

[33] It is publicly available in `https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory`

Spanish	Portuguese	Galician translation	Galician transliteration
Fórmula	Fórmula	Fórmula	Fórmula
Modo de utilización	Via de utilização	Vía de utilización	Vina de utilización
Con trazador	Com marcadores	Con marcadores	Con marcadores
Sin trazador	Sem marcadores	Sen marcadores	Sen marcadores
Precio mínimo de venta	Preço mínimo de venda	Prezo mínimo de venda	Prezo mínimo de venda
Mantequilla > 82	Manteiga > 82	Manteiga > 82	Manteiga > 82
Sin transformar	Em natureza	En natureza	En natureza
Concentrada	Concentrada	Concentrada	Concentrada
Garantía de transformación	Garantia de transformação	Garantía de transformación	Garantía de transformación

Table 4: Spanish-Portuguese-Galician corpus

Spanish	Galician
ÍNDICE	ÍNDICE
Título I . Disposiciones generales	Título I. Disposicións xerais
Artículo 1 . Ámbito de aplicación	Artigo 1 . Ámbito de aplicación
Artículo 2 . Contenido de la solicitud o del escrito de iniciación	Artigo 2 . Contido da solicitude ou do escrito de iniciación
Artículo 3 . Representación	Artigo 3 . Representación
Título II . Procedimientos especiales de revisión	Título II . Procedementos especiais de revisión
Capítulo I . Procedimiento de revisión de actos nulos de pleno derecho	Capítulo I. Procedemento de revisión de actos nulos de pleno dereito
Artículo 4 . Iniciación	Artigo 4 . Iniciación
Artículo 5 . Tramitación	Artigo 5 . Tramitación
Artículo 6 . Resolución	Artigo 6 . Resolución
Capítulo II . Procedimiento para la declaración de lesividad de actos anulables	Capítulo II . Procedemento para a declaración de lesividade de actos anulables
Artículo 7 . Iniciación	Artigo 7 . Iniciación

Table 5: Test Spanish-Galcian from corpus LEGA

On the other hand, as test set, the LEGA corpus available in the Linguistic Corpus of University of Vigo (CLUVI) was used[34]. It is a Spanish-Galician legal-administrative translation corpus of 1012 sentences and almost six million tokens. The original file is in XML format and was also converted to two parallel txt files using a python script and xml library. After removing the duplicated lines, the final test set had 945 sentences. Finally, the Spanish monolingual corpus used to the Back Translation process in gl-es direction was taken from a corpus of the *Boletín Oficial del Estado(BOE)*[35] that contained over 15M sentences from 2016 to 2020. Only the first 2M were used. Some examples of the monolingual data can be seen in table 6. The reason why only this number of sentences was taken is that to BackTranslation as Soto García (2018) mentions in his thesis, backtranslation is is better if the backtranslated corpus is not as twice big or more than the original one. As ours is a small corpus, we prefer not to take many sentences to back translation. The same happens in English-Galician case.

[34] This corpus is publicly available to download in https://repositori.upf.edu/handle/10230/20051

[35] Source: https://elrc-share.eu/repository/browse/spanish-monolingual-corpus-from-contents-of-spanis
4bb236c2ce2711e9913100155d026706ae95b47f242d4c3c91a4a33821351c77/

Spanish Monolingual data
A1, A2. . . An=Número de asociados de cada una de las asociaciones solicitantes de una o varias de las modalidades de subvención
A14A ANABOLIZANTES HORMONALES.
A14A2 Anabolizantes hormonales asociados con vitaminas.
A14A2 ANABOLIZANTES HORMONALES CON VITAMINAS.
A14A1 ANABOLIZANTES HORMONALES SOLOS.
A-a: Anglo-árabe.
A A ANOTARSE EN LA RUBRICA QUE LE PERSONAL SANITARIO
A02A6 ANTIACIDOS CON OTRAS SUSTANCIAS.
A.: A) Anticipos de una mensualidad:
A02A7 Antiflatulentos asociados con otras sustancias.
A02A7 ANTIFLATULENTOS CON OTRAS SUSTANCIAS.

Table 6: Monolingual Spanish Data

3.2.2 English-Portuguese-Galician

As English-Portuguese original corpus, the Europarl-v7 was used, as Fernández et al. (2010) used in their project but in this thesis a more recent version was used. Although Europarl corpus is available in two txt files, the tmx format was taken to later transform it into two parallel files. This is because the two txt file available to download gave problems of alignment and empty lines. So, the same python script used to convert the Spanish-Portuguese file was used. Each file had 1.95M sentences. In the absence of English-Galician open-source legal corpus available on the Internet[36], we created our own legal test file. From the English-Portuguese corpus we took the first 1012 sentences to test. Then, the Portuguese file was firstly translated with *Apertium* and then completely post-edited by hand to check and correct all the mistakes to adapt it to be as close as a real Galician corpus as possible. To do that correctly, it was used, firstly, the *RAG* online dictionary[37] and a manual of Galician administrative and legal language (Cobas Medín, 2016)[38]. All the vocabulary and syntax forms correct in both languages, Galician and Portuguese, were maintained as the Portuguese form which is a common practice in formal registers as the legal one. After removing the duplicated lines, the test file had 1004 lines.

On the other hand, the monolingual English data to carry out the BackTranslation in the gl-en direction was taken from the Corp-HouseOfCommons-V2.rds document of ParlSpeech V2[39]. This document contains parliamentary text of United Kingdom from the eighties to the 2019. In this case sentences from the 2000 to the 2019 were taken, being 1.3M. The rds file was converted to csv and then the sentences were extracted with a python script. As the sentences were so big, almost paragraphs, a sentence splitter available at CITIUS to English, Spanish and Galician among other languages, *Linguakit* was used[40]. After sentence splitting, the monolingual corpus ended up having 14M sentences from which we took 2M. Some examples of the monolingual English data can be seen in table 7.

English Monolingual data
It is pretty vacant all round .
He's got rid of that Liberal penny again .
Answer the question .
The answer is that of course pay matters .
What are you going to do about it ?
If he will make a statement on the funding of schools .

Table 7: Monolingual English Data

[36] Actually an English-Galician legal corpus exists in CLUVI from the UNESCO but it is not available to download, only to consult.

[37] `https://academia.gal/dicionario`

[38] This book is also available on the Internet to download in the *Portal da Lingua* webpage `https://www.lingua.gal/c/document_library/get_file?file_path=/portal-lingua/recursos/02-manual_linguaxe_administrativa_SUPERIOR_v2.pdf`

[39] Source: `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/L4OAKN`

[40] It is freely available at CITIUS webpage, `https://citius.usc.es/transferencia/software/linguakit`

3.2.3 Tatoeba and Flores test sets

To check the performance of our models in out-of-domain scenarios two different test sets available in both Spanish-Galician and English-Galician pairs were chosen: Tatoeba (Tiedemann, 2020) and Flores (Goyal et al., 2021). In Tatoeba test sets, the Spanish-Galician and English-Galician sentences are different. But, on the contrary, in Flores corpus the sentences are the same in English, Spanish bilingual corpus.

Tatoeba offers benchmarks and train corpus in 2961 language pairs,covering 555 languages. Two of these language pairs are Spanish-Galician and English-Galician languages. The English-Galician pair benchmark has 1018 sentence pairs and the Spanish-Galician has originally 3121 sentences but only 1012 were taken to do the test (Tiedemann, 2020).

On the other hand, Flores-101 evaluation benchmark takes 3001 English sentences from Wikipedia domain that have later been translated by professional translators into 101 different languages, one of them, Galician. From the Spanish-Galician and English-Galician corpus only 1012 pair sentences were taken for each language pair (Goyal et al., 2021).

So, finally, all the test sets had 1012 sentences except to English-Galician tatoeba test that had 1018. With these benchmarks we were able to check our models in out-of-domain corpus and compare them to available English-Galician and Spanish-Galician baseline-models. The first ten sentences of Tatoeba corpus can be seen in table 8 to English-Galician pair and in table 9 to Spanish-Galician. On the other hand, part of the first three sentences of Flores corpus in the three languages can be seen in table 10.

English	**Galician**
That's the stupidest thing I've ever said.	É a cousa máis estúpida que dixen nunca.
If you don't eat, you die.	Se non comes, morres.
Do you need me to give you some money?	Precisas que che dea algo de cartos?
I was trying to kill time.	Estaba a tratar de matar o tempo.
I'm thirsty.	Teño sede.
Who painted this painting?	Quen pintou ese cadro?
I knew he would accept.	Sabía que el aceptaría.
I'm taking a walk in a park.	Estou camiñando por un parque.
An eye for an eye, a tooth for a tooth.	Ollo por ollo, dente por dente.
You should have accepted his advice.	Debiches acepta-lo seu consello.

Table 8: Tatoeba English-Galician test

´

Spanish	**Galician**
¿Qué estás haciendo?	Que andas a facer?
¡Feliz cumpleaños, Muiriel!	Feliz aniversario, Muiriel!
La contraseña es "Muiriel".	O contrasinal é "Muiriel".
La educación en este mundo me decepciona.	A educación neste mundo decepcióname.
Entonces tenemos un problema...	Daquela temos un problema...
¡Date prisa!	Bule!
¡Date prisa!	Apura!
¿Entonces qué?	Entón que?
Algunas veces no puedo evitar mostrar mis sentimientos.	Algunhas veces non podo evitar amosa-los meus sentimentos.
¡Puedes hacerlo!	Podes facelo!

Table 9: Tatoeba Spanish-Galician test

English	**Spanish**	**Galician**
"We now have 4-month-old mice [...]	« Actualmente, tenemos ratones de cuatro meses de edad [...]	"Agora temos ratos de 4 meses [...]
Dr. Ehud Ur, professor of medicine at Dalhousie University[...]	el Dr. Ehud Ur, docente en la carrera de medicina de la Universidad de Dalhousie [...]	O Dr. Ehud Ur, profesor de medicina na Universidade Dalhousie [...]
Like some other experts, he is skeptical about whether diabetes can be cured [...]	Al igual que otros especialistas, es escéptico acerca de si la diabetes tiene cura [...]	Do mesmo xeito que outros expertos, amósase escéptico de que a diabetes teña cura [...]

Table 10: Flores English-Spanish-Galician examples

´

3.3 Tool-kits

In this thesis two sequence modeling available tools: Fairseq, 3.3.1 and Open-NMT-py 3.3.2 were used. The reason why two different tools were employed was that, on the one hand Open-NMT-py is available to use in the Colab-Notebook and is easy to employ to get a first contact with machine translation[41] (Klein et al., 2017). On the other hand, Fairseq is a sequence-modelling toolkit written in Pytorch. It allows more tasks than machine translation, as language modeling or summarization. It achieved the state-of-the-art to English-French and English-German pair of languages and allows the use of many sequence-to-sequence models as: convolutional models, LSTM and Transformers[42] and also part of the code used in this project is available at `https://fairseq.readthedocs.io/en/latest/getting_started.html#training-a-new-model` (Ott et al., 2019).

3.3.1 Fairseq

In this subsection the data pre-processing and training parameters are going to be shown.

1. **Data Preprocess**

   To preprocess the bitext data, the *prepare-wmt18en2de.sh*[43] code available in fairseq was used and adapted to our case[44]. With this method the data is tokenized with Moses tokenizer, splitted into train and valid sets[45] and applies the byte pair encoding (BPE) from subword-nmt with a 32k vocabulary size. BPE is a subword segmentation algorithm that works as follows: Firstly, a dictionary from the training data vocabulary is created, in which each word is represented as a sequence of characters plus an end-of-word-symbol. Then, the most frequent symbol pair is identified and all its occurrences are merged. As much the merged operations, as bigger the vocabulary size. Then, the most infrequent words in both corpus are encoded as sequences of subword units (Sennrich et al., 2016) with the '@' symbol. Finally, the clean-corpus perl script is applied. With this script empty lines and sentences of more than 250 tokens or source/target length ratio exceeding 1.5 are removed to get better training performance. After this preprocessing step the final corpus sizes were

[41] The code used in this tool and the tutorial to reproduce the training is available at `https://forum.opennmt.net/t/tutorial-for-opennmt-py-using-colab/2895`

[42] The code used to this training process can be found at Github, `https://github.com/pytorch/fairseq`

[43] `https://github.com/pytorch/fairseq/blob/main/examples/backtranslation/prepare-wmt18en2de.sh`

[44] This preprocess method was also used to train the OpenNMT models as the one proposed in Colab gave very bad results and was very simple, as it this a basic tutorial.

[45] This code automatically splits the data into train and valid. But, as will be explained in the results section 5, we had to modify this first splitting to get better results.

obtained, which can be seen in table 11 where the number of sentences of each document: training and valid sets are shown. The corpus of *Apertium* translated Galician are represented as Galician-ap and the transliterated ones as Galician-tr. As was previously mentioned, the first size of the valid set had to be modified. Finally, 1012 sentences, the same number as to test set, were taken to the valid document, although this number of sentences was modified after applying the clean-corpus. So, the first number of sentences in the training and valid cells are referring to the ones splitted by the fairseq code and the second number to the modification made by us, which, indeed, gave better results.

On the other hand, the monolingual data preprocessing before the back-translation (using the en-gl or es-gl trained model to translate the monolingual target language into Galician), was done with the *prepare-de-monolingual.sh*[46] script. It had to be adapted to our case as this script is written to pre-process 25 different files of 1M sentences each one. As in our case the copus sizes are of 2M sentences the code had to be a little bit modified. At the end, the preprocess is the same as the bitext data but to one language only. The final sizes to Spanish and English monolingual corpus were 1,9M and 1,8M sentences respectively.

English-Galician-ap	English-Galician-tr	Spanish-Galician-ap	Spanish-Galician-tr
Training 1.85M/1.89M	Training 1.88M/1.89M	Training 1.77M	Training 1.78M
Valid 18.7k/957	Valid 19k/960	Valid 983	Valid 985

Table 11: Bilingual corpus sizes

2. **Training Parameters**

   The architecture chosen to train these models in fairseq was the transformer, specifically the transformer-iwsmlt-de-en. Firstly we used the architecture proposed in the code shown in the back-translation code example of transformer-wmt-en-de-big, but as our dataset is not so big it gave very poor results, so finally the transformer-iwsmlt-de-en was chosen. The parameters used in bilingual models are shown in table 12. All the parameters were taken from the example in Back-Translation code. We had only to adapt the update-freq from 16 to 2 as it was recommended that if there were less than 8 GPU, as this case is, this value should be reduced. Also, the parameters used to Back-Translation are shown in table 13. Back-Translation is a process used to develop bilingual models with monolingual data. It operates as a semi-supervised setup in which an intermediate system on the parallel data is trained and used to translate monolingual data of the target language in the source language. Then a

[46] Source: `https://github.com/pytorch/fairseq/blob/main/examples/backtranslation/prepare-de-monolingual.sh`

parallel corpus where the source language is *synthetic* data[47] and the target language is original one. This data is added to the original parallel corpus to train a final system with augmented data (Edunov et al., 2018). Although Edunov et al. (2018) mentions that beam is more effective than sampling in poor resource settings, it is referring to settings of 80k bitext sentences, so in this case we decided to choose the sampling algorithm to do the Back Translation as the setup where it performs the best is ranged between 640k and 5.2M sentences in the bitext (Edunov et al., 2018).

Parameters	**Values**
Arquitecture	Transformer
dropout	0.3
weight decay	0.0
label-smoothing	0.1
optimization	adam-betas (0.9, 0.98)
clip-norm	0.0
learning rate	0.001
warmup-updates	4000
max-tokens	3584
update-freq	2

Table 12: Bilingual models fairseq parameters

Parameters	**Values**
Arquitecture	Transformer
upsampling-primary	2
dropout	0.3
weight decay	0.0
label-smoothing	0.1
optimization	adam-betas (0.9, 0.98)
clip-norm	0.0
learning rate	0.001
warmup-updates	4000
max-tokens	3584
update-freq	2

Table 13: Backtranslation fairseq parameters

[47] *synthetic* data is the one that is not original of this language. For example, the Galician data generated by Portuguese-Galician adaptation by translation and transliteration can be considered as *synthetic* data.

3.3.2 OpenNMT-py

The Open-NMT-py tool was only used to do the first trainings and tests in the English-Galician datasets. These tests were made in Colab and the only GPU available to use was the one that Colab offers, so the trainings done in this tool were only the first approximations. As it was mentioned in the section below, the data preprocessing was the same as the one made to fairseq because the one proposed in Colab tutorial was very basic. The architecture used in Open-NMT was also the transformer, as it was the one proposed in the Colab code. All the parameters were maintained except the batch-size[48]. The parameters can be seen in table 14.

Parameters	**Values**
Arquitecture	Transformer
layers	6
rnn-size	512
word-vec-size	0.0
transformer-ff	2048
heads	8
max-generator-batches	2
dropout	0.1
batch-size	2000
accum-count	2
adam-beta2	0.998
warmup-steps	8000
learning-rate	2
max-grad-norm	0
param-init	0
label-smoothing	0.1
valid steps	1000
save-checkpoint-steps	1000
world-size	1
gpu-rank	0

Table 14: Open-NMT transformer parameters

3.4 Equipment

Neural Machine Translation needs big computation resources, because of that Graphics Processing Units (GPU) are necessary to train these type of models. To this thesis three

[48] As the Colab GPU has limitations of time, the batch-size value had to be changed to be able to train the model without an Out of Memory error (OOM) in Colab. It gave worse results but as it was only an approximation and first test to the later experiments in Fairseq, it was maintained as it.

different GPUs were used. Firstly, the one available in Google Colab that has time limitations and has not a lot of memory freely available. Secondly, a NVIDIA corporation TU116M [GeForce GTX 1660 Ti Mobile] (rev a1) GPU with 256 MB memory that is the one available in Imaxin was used to train the es-gl models and some of the gl-en ones. Finally, a more powerful GPU, Intel(R) Xeon(R) CPU E5-2640 v4 processor with 64GB of RAM and a GeForce GTX TITAN X with 12GB of RAM available at the UPV/EHU university was used to carry out the last experiments of fairseq, especially to train the backtranslated models. The difference in training time between second and the last GPUs is significant. What the second GPU trained in two days, the third GPU trained in twelve hours.

4 Our Approach

As it was mentioned in section 1 in this project a strategy to Galician data augmentation in machine translation is proposed. Although it is not the first time that data from the Portuguese have been exploited to data augmentation to Galician, as it was mentioned in section 2, it is the first time that transliteration and translation with a RBMT system are proposed as strategies to convert Portuguese into Galician to the make synthetic bilingual corpus.

Although there are some publicly available bilingual data in Galician-English and Galician-Spanish in webpages as Opus-mt or Tatoeba project, this data is usually taken from Internet pages as, mostly, Wikipedia, so the quality of the language is not always the best. On the other hand, although multilingual models are giving very good results, as it will be seen in section 5, training this type of systems need an important computational effort and amount of data. Furthermore, as it will be also seen in the Results section, they can easily mix the languages at the time of translating. So, given the scarcity of official data in Galician beyond the web resources, this strategy allows the use of resources of a majority and institutionalised language to train models in Galician with a good linguistic quality, especially at a syntactic and vocabulary level. Furthermore, it allows to adapt data of different domains with fewer resources than would be needed to train, for example, multilingual models and apply it to real scenarios as, for example, a market level. Finally, although in this case it was necessary to limit the corpus to the legal domain because of the reasons explained in subsection 3.1, RBMT *Apertium* could be also used to translate Brazilian variant, that it is usually the most available.

Furthermore, to be able to test the English-Galician models it was necessary to create our own English-Galician test set as there is no one available. So, this resource can be put publicly available to future use. And, although it was not possible to do the backtranslation in es-gl and en-gl directions because of the lack of time and resources, a part of monolingual legal Galician corpus has been created, although it only has around 400k sentences, which is not enough to carry out backtranslation, but it could be also useful to other NLP tasks.

5 Results and Analysis

In this section the results of the different experiments will be shown. They are going to be divided into the Open-NMT-py models in section 5.1 and the Fairseq models in section 5.2. In Open-NMT-py only English models have been trained as a first attempt. It is in fairseq where models in both pair of languages have been trained. So, Fairseq section will be also divided into Spanish-Galician 5.2.1 and English-Galician 5.2.2.

5.1 OpenNMT-py

As it has been previously said, OpenNMT-py was a first interaction and test to train translation models in an easier and faster way. As the GPU used was the one that Colab-Notebook offers, the training results are not high and not useful to compare with the ones of Fairseq because of the lack of capacity it had. So, the main purpose of these models were, firstly, start to see the different results of Apertium and transliterated corpus and, secondly, the difference between the fairseq splitted corpus and our own division by hand with less valid sentences. And, finally, only the English-Galician direction was trained in this tool, as it was only a starting point to this thesis. In table 15 the results of the fairseq splitted corpus (18.9k valid sentences) are shown. As it can be noticed there is a difference of almost five points in BLEU between translated and transliterated corpus. It is normal, as the vocabulary of the transliterated corpus is completely Portuguese, except the words that with the transliteration strategy are the same both in Galician and Portuguese. Because of that, most of the nouns, adjectives, adverbs are going to be considered incorrect by BLEU metric, whereas the syntax and sentence structure are most correct in Galician language too. On the other hand, in table 16 the results after adapting the valid corpus with less sentences can be seen. Although both the Apertium and transliterated corpus have increased their BLEU results, it is especially noticeable the transliterated corpus increase and the fact that the difference between translated and transliterated corpus is less. Whereas the Apertium corppus had not a significant increase, the transliterated corpus has a punctuation of three points more in BLEU. So, the valid sentences reduction helps to a better training and learning process. For this reason, this valid sentence reduction was maintained in the fairseq models.

Original fairseq splitting	
Apertium	Transliterated
19.11	14.34

Table 15: English-Galician BLEU with original train and valid division

Less valid sentences	
Apertium	Transliterated
19.54	17.09

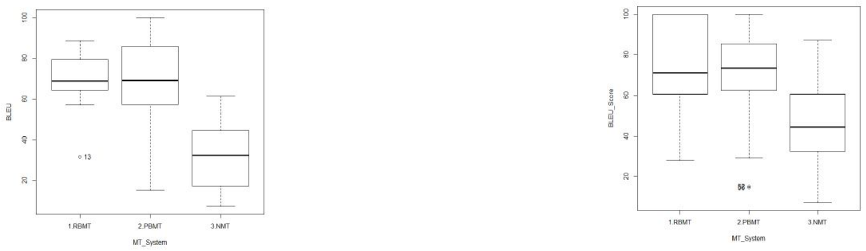
Table 16: English-Galician BLEU results with less valid sentences

5.2 Fairseq

In this section both Spanish-Galician and Galician-Spanish results can be seen in subsection 5.2.1. On the other hand, both English-Galician and Galician-English models results can be seen in subsection 5.2.2. Although the *fairseq-generate* command automatically evaluates with BLEU4 metric, the metrics chosen to evaluate all the models and compare them to the actual baselines are, as it was previously mentioned: Sacrebleu, chrF and TER. This is because, as it was explained in section 2.1.4, BLEU scores depend a lot on tokenization, so metrics as sacrebleu, chrF or TER are more accurate and reliable to compare different models. On the other hand, from now on, as it was called in table 11 the models achieved by translating the Portuguese part to Galician with Apertium will be called Galician-ap, whereas the transliterated ones will be called Galician-tr. Finally, in gl-es and gl-en direction, the backtranslated model will be called Galician-bk.

5.2.1 Spanish-Galician and Galician-Spanish models

Regarding to Spanish-Galician bilingual pair in NMT research, there is not much investigation. Only, as it was mentioned in the Antecedents section, Bayón and Sánchez-Gijón (2019) compared RBMT, PBMT and NMT machine translation to Spanish-Galician direction to check what machine translation system was the best in this case. The results of its investigation can be seen in figure 14. The results are based on around 32 segments of sentences. From this data, two BLEU scores have been measured. One taking into account the overall sentences, result that can be seen in picture a) of figure 14 and the other BLEU score is measured taking ito account the sentences longer than 30 words. These results are shown in picture b) of the same figure. In this paper, Bayón and Sánchez-Gijón (2019) concludes that RBMT and PBMT systems are still better than NMT to high related languages as is the case of Spanish-Galician. But she also argues that more investigation in Spanish-Galician NMT is needed. So in this section augmented bilingual, multilingual and RBMT systems are going to be better compared.



(a) Overall Results

(b) More than 30 words sentence

fig. 14. RBMT, PBMT, NMT Spanish-Galician-BLEU-scores

In table 17 results in both directions es-gl and gl-es on LEGA test are shown. As it can be seen, backtranslation in gl-es direction is the best result achieving a 67.10 sacrebleu score. Moreover chrF and TER metrics are correlated, as they are the best ones compared to the other models too. Although the difference between models taking into account these metrics is not as high as sacrebleu score except to TER results in gl-es direction between Galician-ap and Galician-bk which have the most significant difference between them in TER evaluation metric. On the other hand, it can be also perceived that in both directions the worse results are in the transliterated corpus but, as it was mentioned, the difference between them is smaller in TER and chrF metrics than in sacrebleu, especially in es-gl direction. Even so, transliteration results are still good taking into account that the vocabulary from Portuguese to Galician was not changed at all. So, although the Galician generated by these models seems not to be so accurate compared to the LEGA reference, the model can translate very well from Galician to Spanish.

Es-Gl			
Models	**Sacrebleu**	**chrF**	**TER**
Galician-tr	30.49	0.64	0.48
Galician-ap	33.10	0.65	0.47
Gl-Es			
Galician-tr	53.80	0.75	0.30
Galician-ap	56.50	0.77	0.28
Galician-bk	67.10	0.82	0.22

Table 17: Es-Gl and Gl-Es evaluation in LEGA test

After these results it was decided to compare these models with some of the baselines available to Spanish-Galician pair in the same test set: *Apertium* RBMT[49]. After the translation with RBMT system the only post-edition done was to remove the special characters

[49] There is an open-source tool available of *Apertium* but in this case the version used was *Gaio*, the one used in Imaxin to official legal translations and journals

generated when dictionaries do not recognise specific words: '*', '#' and '@'. On the other hand, NMT multilingual models available at hugging face were used as baselines too, such as: MBART-large-50-to-many models[50] and m2m100418M[51]. The results in the same test LEGA can be seen in table 18. In this table it can be seen that Apertium has the best es-gl results compared to the rest of the models, although M2M multilingual model also had surprisingly good scores, especially in gl-es direction in which even outperforms a RBMT system that is prepared to legal domains. Another interesting point is the poor performance of MBART in es-gl direction, this model mix completely both languages as it is shown in table 20. In fact, it is quite surprising the difference between the translation in es-gl with the translation in gl-es direction that can be also seen in table 19. In both tables the words that are translated in another languages are highlighted in bold. As it can be seen, even in gl-es direction it mixes the languages, but Galician translation in LEGA corpus is almost all in Spanish, maybe because of the lack of legal vocabulary that this model has.

Es-Gl			
Models	**Sacrebleu**	**chrF**	**TER**
MBART	8.60	0.37	1.30
M2M	79.90	0.92	0.09
Apertium RBMT	**84.10**	**0.94**	**0.07**
Galician-tr	30.49	0.64	0.48
Galician-ap	33.10	0.65	0.47
Gl-Es			
MBART	45.30	0.67	0.43
RBMT Apertium	80.90	0.93	0.10
M2M	**82.60**	**0.93**	**0.09**
Galician-tr	53.80	0.75	0.30
Galician-ap	56.50	0.77	0.28
Galician-bk	67.10	0.82	0.22

Table 18: Es-Gl and Gl-ES evaluation in LEGA test with all the models

[50] Source: `https://huggingface.co/facebook/mbart-large-50-one-to-many-mmt`. This model was trained with original corpora of 50 different languages in which are Spanish and Galician, so it allows the translation between this language pair.

[51] Source: `https://huggingface.co/facebook/m2m100_418M`. This model is a seq-to-seq that can translate between the 9900 directions of 100 languages

REAL DECRETO 520 / 2005, del 13 de mayo, por el que se **aproba** el Reglamento General de Desarrollo de la Ley 58 / 2003, del 17 de diciembre
En particular, **the legislación** contiene una nueva serie de normas reguladoras de revisión administrativa[...]
Title I, "" General Provisions, "" contiene el ámbito de aplicación

Table 19: gl-es-MBART-translation

REAL DECRETO 520 / 2005, **de 13 de mayo**, por que se adopte o **Reglamento General de Desarrollo de la Ley 58 / 2003, de 17 de diciembre**
A aprobación da **Ley 58 / 2003, de 17 de diciembre, General Tributaria**, significou unha reforma importante na revisión de actos administrativos,[...]
En particular, o Código de Procedimiento Penal[...]
Title I, "Disposiciones generales", contains the scope de aplicación

Table 20: es-gl-MBART-translation

Finally, it was decided to test our models and the baselines in out-of-domain corpus, so Flores and Tatoeba, already mentioned in section 3.2.3, were used. The results achieved in these test sets are shown in table 21. In this table there are some aspects to highlight. First of all, as it was expected, our models perform worse in Tatoeba and Flores as there is a lot of vocabulary that they have not learnt. Despite that, the backtranslated model in gl-es direction outperforms MBART in all the test sets. Furthermore, all the models give poorer and not good results in Flores test, maybe because the sentences are much longer than in Tatoeba and LEGA tests, so the longer the sentences, the poorer the performance of machine translation models. Regarding to the performance of the transliterated model in es-gl direction in Tatoeba and Flores it is extremely poor, as MBART in LEGA test, this model also mixes a lot both languages. Once again, the lack of vocabulary to these tests and the fact that most of it is Portuguese and not Galician makes the translation to Galician much worse than in the opposite direction. Another interesting fact is that in all the tests in es-gl direction, Apertium outperforms the rest of the models, even though they are completely out of domain corpus. But in gl-es direction M2M model slightly outperforms Apertium in corpus LEGA and Flores, which is quite impressive. In this case, a neural model is outperforming a RBMT specially prepared to legal domain.

Es-Gl	**Tatoeba**			**Flores**			**LEGA**		
	Sacrebleu	**chrF**	**TER**	**Sacrebleu**	**chrF**	**TER**	**Sacrebleu**	**chrF**	**TER**
MBART	27.60	0.51	0.56	12.20	0.42	0.77	8.60	0.37	1.30
M2M	55.20	0.71	0.32	21.70	0.52	0.67	79.90	0.92	0.09
RBMT Apertium	**71.70**	**0.83**	**0.17**	**23.20**	**0.53**	**0.63**	**84.10**	**0.94**	**0.07**
Galician-tr	5.30	0.31	1.06	3.60	0.33	1.05	30.49	0.64	0.48
Galician-ap	24.70	0.51	0.65	14.60	0.47	0.76	33.10	0.65	0.47
Gl-Es	**Sacrebleu**	**chrF**	**TER**	**Sacrebleu**	**chrF**	**TER**	**Sacrebleu**	**chrF**	**TER**
MBART	33.60	0.56	0.58	16.00	0.45	0.70	45.30	0.67	0.43
M2M	61.70	0.76	0.26	**23.40**	**0.53**	**0.62**	**82.60**	**0.93**	**0.09**
RBMT Apertium	**70.70**	**0.83**	**0.19**	23.20	0.53	0.63	80.90	0.93	0.10
Galician-tr	28.5	0.55	0.53	16.8	0.46	0.69	53.80	0.75	0.30
Galician-ap	32.9	0.55	0.53	17.9	0.48	0.68	56.50	0.77	0.28
Galician-bk	34	0.55	0.53	18.6	0.48	0.68	67.10	0.82	0.22

Table 21: Models Es-Gl and Gl-Es evaluation in Tatoeba, Flores and LEGA test sets

To end up, some translations of LEGA test achieved by our best models, galician-ap in es-gl direction and galician-bk in gl-es direction will be shown in tables 22 and 23 respectively. Again, the language mistakes are highlighted in bold.

Artigo único
Execución
É suficiente unha tentativa única cando o destinatário se encontrar como descoñecido
Aprobación do Estatuto Xeral do **Abogacía** Española
Artigo 42. o : Efectos da concesión ou **rexeita** da suspensión
Artigo 22. o : Efectos da interposiciónen **relación** ao exercicio doutros recursos
Artigo 36. o : **Quantía necesario** para o recurso normal

Table 22: es-gl-galician-ap LEGA translation

Reducción proporcional de las garantías **achegadas** para la suspensión
El órgano unipersonal podrá aplazar la conclusión de la vista a otro día que se determine [...]
Tribunal Económico-Administrativo Regional de Aragón , con sede en Zaragoza
En este caso , la garantía conservará sus efectos en el procedimiento económico-administrativo [...]
Artículo 66 Aplicación de resoluciones administrativas
Consecuencias de la simultaneidad
Artículo 50 Domicilio para notificaciones

Table 23: gl-es-galician-bk LEGA translation

5.2.2 English-Galician and Galician-English models

In this subsection our English-Galician models will be tested and later compared to three different baselines: en-ROMANCE[52] model, MBART and M2M[53]. As in Spanish-Galician models, the test used were: first of all, the test that has been created from Europarl translating the Portuguese language to Galician, firstly with Apertium and then manually, and then the Tatoeba and Flores test sets in English-Galician pair. In table 24 our models are tested with the test created from Europarl. As it can be noticed, in both directions the translated model, Galician-ap, has achieved again better results than the transliterated one. Furthermore, English-Galician direction gives bad results taking into account that it is not achieving in Sacrebleu score more or equal than 30 and chrF and TER results also not get the minimum to be considered a good or valid translation. Regarding to Galician-English direction, the results get better in both models, giving the galician-bk the best results, although the difference between the translated and backtranslated models are not as high as in Spanish-Galician language pair. Indeed, the backtranslated model has slightly grown in sacrebleu but it has not changed in chrF and TER tests, so in this two distant languages the backtranslated process may need more iterations and different strategies to get better.

En-Gl			
Models	**Sacrebleu**	**chrF**	**TER**
Galician-tr	24.50	0.55	0.65
Galician-ap	**28.20**	**0.58**	**0.62**
Gl-En			
Galician-tr	29.10	0.56	0.61
Galician-ap	31.50	0.58	0.57
Galician-bk	**32.20**	**0.58**	**0.57**

Table 24: En-Gl and Gl-En evaluation in my own test

On the other hand, in table 25 the results of our models compared to the baselines in our test are shown. In this case, in contrast to Spanish-Galician models, our models achieve the best results in comparison to baselines, even the transliterated one. It is something expected as the test is taken from the same corpus used to train them. Moreover, it is a very specific domain and the first translation of the Portuguese test part is done with the same RBMT system used to convert the training corpora. The results of the baseline models also increase in Galician-English direction, in which the differences between baselines and our models are smaller, especially the ones achieved by opusmt ROMANCE-en, but again

[52] Source: `https://huggingface.co/Helsinki-NLP/opus-mt-en-ROMANCE` this model allows the machine translation of English to many romance languages and vice versa.

[53] In this case the use of RBMT Apertium system en-gl, although it exists, was discarded as baseline model because of the poor performance and the fact that it is not even used.

the transliterated model outperforms the baselines. As the Galician of this test is synthetic as the one of our training corpora, it is normal that baseline models perform worse in this test set.

En-Gl			
Models	**Sacrebleu**	**chrF**	**TER**
Opusmt-en-ROMANCE	14.00	0.47	0.73
MBART	17.70	0.49	0.70
M2M	19.50	0.51	0.68
Galician-tr	24.50	0.55	0.65
Galician-ap	**28.20**	**0.58**	**0.62**
Gl-En			
Opusmt-ROMANCE-en	28.00	0.55	0.64
MBART	26.10	0.54	0.65
M2M	24.30	0.54	0.65
Galician-tr	29.10	0.56	0.61
Galician-ap	31.50	0.58	0.57
Galician-bk	**32.20**	**0.58**	**0.57**

Table 25: En-Gl and Gl-En evaluation in my own test with all models

Finally, in table 26 the results in Tatoeba and Flores test sets both with our models and the base-lines are shown. Again, as in Spanish-Galician models, it was expected that in a non-specific domain our models perform much worse. In this case M2M is again the best performer although MBART has also achieved very good results in comparison to Spanish-Galician models, On the other hand, in gl-en direction, again, all the models perform better. It is also interesting to mention that, although in Tatoeba and Flores the best results are achieved by M2M model, in the gl-en direction of Tatoeba test our back-translated model outperforms the ROMANCE-en one in spite of being an out-of-domain test, so it would be expected that it could achieve better results if it was trained with different corpora.

En-Gl	**Tatoeba**			**Flores**			**Europarl-Test**		
	Sacrebleu	**chrF**	**TER**	**Sacrebleu**	**chrF**	**TER**	**Sacrebleu**	**chrF**	**TER**
en-ROMANCE	25.30	0.50	0.59	20.00	0.51	0.62	14.00	0.47	0.73
MBART	37.30	0.59	0.47	25.4	0.54	0.57	17.70	0.49	0.70
M2M	**37.40**	**0.59**	**0.48**	**29.30**	**0.57**	**0.53**	19.50	0.51	0.68
Galician-tr	17.80	0.45	0.75	11.60	0.38	0.76	24.50	0.55	0.65
Galician-ap	22.40	0.47	0.69	18.70	0.49	0.67	**28.20**	**0.58**	**0.62**
Gl-En	**Sacrebleu**	**chrF**	**TER**	**Sacrebleu**	**chrF**	**TER**	**Sacrebleu**	**chrF**	**TER**
ROMANCE-en	26.50	0.52	0.56	31.10	0.59	0.55	28.00	0.55	0.64
MBART	41.30	0.60	0.45	27.60	0.58	0.59	26.10	0.54	0.65
M2M	**46.30**	**0.64**	**0.40**	**32.50**	**0.61**	**0.53**	24.30	0.54	0.65
Galician-tr	24.40	0.46	0.63	18.70	0.49	0.70	29.10	0.56	0.61
Galician-ap	28.1	0.49	0.61	21	0.52	0.67	31.50	0.58	0.57
Galician-bk	29.00	0.5	0.60	22.20	0.51	0.65	**32.20**	**0.58**	**0.57**

Table 26: Models En-Gl and Gl-En evaluation in Tatoeba, Flores and LEGA test sets

Finally, in tables 27 and 28 some examples of our models translation in our test both in en-gl and gl-en translation can be respectively seen.

Debe proseguir por este camiño
Aplausos do Grupo PSE
Comprendo o que está a dicir
Procedemos seguidamente á votación
Sometemo - lo á votación
Solicito a súa votación favorábel
Esta directiva é unha contribución nese sentido

Table 27: en-gl-galician-ap our test translation

I agree with your analysis
Thank you very much , Commissioner
I understand what the honourable Member means
There are also decisions against such a tax
I therefore have to take a decision
Personally , I fully agree with this position
That is why we will vote in favour of it

Table 28: gl-en-galician-bk our test translation

5.2.3 Analysis

In this subsection a small analysis of general results in both language pairs will be done. First of all, it can be concluded that, although Portuguese and Galician are very close and similar languages, a simple transliteration process gives, in general, worse results than the corpus achieved by translating the Portuguese part to Galician with *Apertium*. On the other hand, regarding models in Spanish-Galician language pairs, it could be seen that even nowadays RBMT systems are giving very good results in similar language pairs. So, they are systems that should be considered even nowadays in spite of the fact that they are time consuming and need a lot of maintenance. But, at the same time, M2M multilingual model has given quite surprising good results too even outperforming RBMT in some test sets, so neural systems are getting closer and closer to the sate of the art in similar language pairs on the contrary what was mentioned in Bayón and Sánchez-Gijón (2019) article. Regarding our models, although the gl-es direction in LEGA test has given quite good results it was not enough to outperform or be close enough to a RBMT or M2M multilingual performance. But the fact that in gl-es direction it could outperform a baseline as MBART in an out-of-domain test, suggests that with a little more investigation the results in this type of bilingual model could be much better.

With respect to English-Galician pair, this strategy performs better in relation to base-line models than in Spanish-Galician pair. In our legal test set it achieves in both direction the best results and even outperforms one of the base-line models, ROMANCE-en in Tatoeba test set in gl-en direction with the back-translated model, even though the model is trained in a legal domain. So, the results achieved by the English-Galician and Galician-English models even in out of domain corpus seem promising and also need more investigation to develop this performance.

6 Conclusions

To finish this thesis, some conclusions drawn after doing it will be mentioned. First of all, one of the main objectives of this thesis was to test if data augmentation from Portuguese was a valid strategy to train bilingual models to Galician language and we think that it was seen that it is. Secondly, the other objective was to compare transliteration and RBMT translation to data adaptation from Portuguese to Galician language in machine translation. In both bilingual models Spanish-Galician and English-Galician, it could be seen that transliteration is not enough to train good bilingual models in Galician, indeed RBMT translation has given much better results. Finally, the quality of the language could not be tested because of lack of time but it would be very interesting as future work.

On the other hand, the results achieved by the Spanish-Galician models were as good as a first time was expected, giving RBMT system the best results in most of the test sets and even being outperformed by one of the multilingual models. Although our models worked fine in LEGA corpus it was not enough to be close to RBMT translation or M2M. Finally, the English-Galician models give quite promising results especially the one translated with RBMT in en-gl direction and the backtranslated in gl-en direction. Although, as it was expected, it did much worse in out-of-domain tests than the baseline models, it seems that training it with another type of corpus could develop the performance.

To end up, some final thoughts after doing this project will be exposed. Firstly, we think that it is very important to define the purpose of any translation model, which means that, for example, if the purpose is to offer an official translation as a company, as it can be the case of *Imaxin*, beyond the results it is important to be able to solve the possible errors that the machine translation could give, which it is quite difficult in NMT systems. So, as it was mentioned in the previous section, it is important to reflect on whether the RBMT systems are disposable in similar languages or whether they still offer, despite their many drawbacks, important advantages such as control over their functioning and the real possibility of correction.For example, in a research if the machine translation only makes a mistake in a word it would be almost perfect, but in an official use at can be a journal or a webpage, it could be a fatal error that almost changes the complete meaning of the sentence. But, on the other hand, it is true that it is very different in the case of distant languages as English-Galician in which NMT has really developed the performance of machine translation. So, in cases as that, NMT is the only real alternative.

7 Future Work

To finish this thesis, some final ides to future work around this topic will be displayed. Because of the lack of time there are many experiments that could not be done in this thesis and could to tried in a future, such as:

- It was not possible to do the BackTranslation in en-gl and es-gl direction so with enough Galician monolingual data to carry out the backtranslation in this direction, this experiment could be done.

- Also related to the Backtranslation, in this final thesis only a first approximation was done, so it would be interesting to develop it with more iterations, different parameters, etc. to compare the performances.

- One of the main disadvantages of this project was the fact that it was necessary to limit the data to a very specific domain, so it would be good to use another type of corpus.

- Although in this case the data augmentation was done from Portuguese, to the English-Galician model it could be possible to take advantage of English-Spanish parallel data and translate this Spanish to Galician using the RBMT system *Apertium* to do the same experiment.

- We could not train as base line a model trained from original Spanish-Galician and English-Galician bilingual models from available data to compare it to the performance of the data augmented systems, so it would be important to compare them in future research.

- In this final thesis, Fairseq transformers were used to train the final models but also platforms as OpenNMT or different architectures mentioned in previous sections, 2 as LSTM could be used to compare the performance of them in respect to Fairseq or transformers.

- Finally, as at the beginning of this thesis one of the main objectives was to test the quality of the language that taking advantage of Portuguese give, a manual testing should be done, as there are some quality characteristics that automatic evaluations cannot detect.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*, 2019.

Nora Aranberri, Gorka Labaka, Arantza Díaz de Ilarraza, and Kepa Sarasola. Ebaluatoia: crowd evaluation for english–basque machine translation. *Language Resources and Evaluation*, 51(4):1053–1084, 2017.

Carme Armentano Oller, Antonio Miguel Corbí Bellot, Mikel L Forcada, Mireia Ginestí Rosell, Marco A Montava Belda, Sergio Ortiz Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez Sánchez, Felipe Sánchez-Martínez, et al. Apertium, una plataforma de código abierto para el desarrollo de sistemas de traducción automática. 2007.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2018.

Satanjeev Banerjee and Alon Lavie. 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, 2005.

María Do Campo Bayón and Pilar Sánchez-Gijón. Evaluating machine translation in a low-resource language combination: Spanish-galician. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 30–35, 2019.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

Ralf D Brown. The cmu-ebmt machine translation system. *Machine translation*, 25(2): 179–195, 2011.

M Asunción Castano, Francisco Casacuberta, and Enrique Vidal. Machine translation using neural networks and finite-state models. *Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 160–167, 1997.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

Domínguez Conde N. et al. Cobas Medín, D. Curso de linguaxe administrativa. nivel superior. 2016.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021.

Paulo Malvar Fernández, José Ramom Pichel Campos, Oscar Senra Gómez, Pablo Gamallo, and Alberto García. Vencendo a escassez de recursos computacionais. carvalho: Tradutor automático estatístico inglês-galego a partir do corpus paralelo europarl inglês-português. *Linguamática*, 2(2):31–38, 2010.

Mikel L. Forcada. Building machine translation systems for minor languages: Challenges and effects. *Revista de Llengua i Dret // Journal of Language and Law*, (73), 2020.

Mikel L Forcada and Ramón P Ñeco. Recursive hetero-associative memories for translation. In *International Work-Conference on Artificial Neural Networks*, pages 453–462. Springer, 1997.

Marcos Garcia, Carlos Gómez-Rodríguez, and Miguel A Alonso. Creación de un treebank de dependencias universales mediante recursos existentes para lenguas próximas: el caso del gallego. *Procesamiento del Lenguaje Natural*, (57):33–40, 2016.

Aranberri N. Goenaga, I. and G. Labaka. Machine translation. *Project European Language Equality (ELE) Grant agreement no. LC-01641480–101018166 ELE Coordinator Prof. Dr. Andy Way (DCU) Co-coordinator Prof. Dr. Georg Rehm (DFKI) Start date, duration 01-01-2021, 18 months*, 2021.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*, 2021.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*, 2018.

Rohan Jagtap, Dr Dhage, and N Sudhir. An in-depth walkthrough on evolution of neural machine translation. *arXiv preprint arXiv:2004.04902*, 2020.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709, 2013.

Alina Karakanta, Jon Dehdari, and Josef van Genabith. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1):167–189, 2018.

Tanmai Khanna, Jonathan N Washington, Francis M Tyers, Sevilay Bayatlı, Daniel G Swanson, Tommi A Pirinen, Irene Tang, and Hèctor Alòs i Font. Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, pages 1–28, 2021.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*, 2017.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Yichong Leng, Xu Tan, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. Unsupervised pivot translation for distant languages. *arXiv preprint arXiv:1906.02461*, 2019.

Chin-Yew Lin and FJ Och. Looking for a few good metrics: Rouge and its evaluation. In *Ntcir Workshop*, 2004.

L.Specia and Y.Wilks. Machine translation. In *The Oxford Handbook of Computational Linguistics 2nd edition*. 2021.

Some Aditya Mandal. Evolution of machine translation, 2019. URL `https://towardsdatascience.com/evolution-of-machine-translation-5524f1c88b25`.

Shereen A Mohamed, Ashraf A Elsayed, YF Hassan, and Mohamed A Abdou. Neural machine translation: past, present, and future. *Neural Computing and Applications*, 33 (23):15919–15931, 2021.

Graham Neubig and Junjie Hu. Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*, 2018.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, 2015.

Matt Post. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*, 2018.

Sahana Ramnath, Melvin Johnson, Abhirut Gupta, and Aravindan Raghuveer. Hint-edbt: Augmenting back-translation with quality and transliteration hints. *arXiv preprint arXiv:2109.04443*, 2021.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*, 2020.

Holger Schwenk. Continuous space translation models for phrase-based statistical machine translation. In *Proceedings of COLING 2012: Posters*, pages 1071–1080, 2012.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*, 2016.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas. URL `https://aclanthology.org/2006.amta-papers.25`.

Xabier Soto García. Basque-to-spanish and spanish-to-basque. machine translation for the health domain. 2018.

I. Sutskever, O. Vinyals, and Q.V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

Jörg Tiedemann. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*, 2020.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. Multilingual neural machine translation with soft decoupled encoding. *arXiv preprint arXiv:1902.03499*, 2019.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. Generalized data augmentation for low-resource translation. *arXiv preprint arXiv:1906.03785*, 2019.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*, 2016.