Improving Neural Machine Translation of Spanish to Quechua with Transfer Learning

Stanford CS224N {Custom} Project

Kaiyu Ren
Department of Computer Science
Stanford University
kyren@stanford.edu

Abstract

Because it is impossible to gather sufficient training data for low resource languages, there is a growing interest in cross-lingual transfer learning. In this paper, we propose a new transfer learning method for the low-resource language pair Spanish-Quechua using parallel data from related language pairs.Our approach leverages the agglutinative nature of both Quechua and Finnish, where words are formed by stringing together morphemes, to improve translation quality. We also implements gradual unfreezing to enhance the performance of the model. Overall, we demonstrate the effectiveness of our method through experiments and discuss the potential for further improvement with additional parallel data and low-resource language pairs.

1 Key Information to include

- Mentor: Isabel Papadimitriou

2 Introduction

Quechua is a widely spoken **low-resource language** in South America with nearly 9 million speakers. However, there are few successful adaptations of machine translation systems for low-resource languages like Quechua. One of the challenges is finding or creating data-efficient models to translate Quechua-Spanish. The key characteristic of the Quechua language is its **agglutinative** nature, which involves a complex morphology where words are formed by adding affixes to a root word. The affixes can be added to the beginning or end of the root word, resulting in the formation of many different words using a limited number of roots. The agglutinative nature of Quechua also enables the formation of long words with multiple affixes, making the language difficult for non-native speakers to comprehend.

To address the challenges associated with Quechua, we propose using transfer learning by pretraining a Spanish to Finnish translation model, as Finnish is another agglutinative language like Quechua. By leveraging the experience gained in translating Spanish to Finnish, the neural machine translation (NMT) model can learn to handle the morphology of agglutinative languages, which could improve its translation accuracy for Quechua.

In this paper, we aim to answer the following research questions (RQs): (1) Can transfer learning through pretraining be effective for low-resource languages like Quechua? (2) How does the translation accuracy improve with techniques such as progressive unfreezing. Regarding RQ1, we use the opus model pretrained on Spanish and Finnish using a transformer model. For RQ2, our approach involves fine-tuning the pretrained Spanish to Finnish model on the Quechua-Spanish translation task using the progressive unfreezing technique.

In general, our findings indicate that transfer learning outperforms our baseline model for the translation task, and we observe some improvement in the model's scores by unfreezing the encoder.

3 Related Work

There have been several papers on the effectiveness of transfer learning in neural machine learning. Applying transfer learning has also shown to have been effective in enhancing the BLEU score in low resource settings [1] (Zoph, 2016). This paper uses French as the parent source language and uses Hausa, Turkish, Uzbek, and Urdu as the child source language. They were able to improve baseline NMT models by an average of 5.6 BLEU across all four language pairs. Ensembling and unknown word replacement added another 2 BLEU, bringing the NMT performance on low-resource machine translation close to a strong syntax-based machine translation (SBMT) system, exceeding its performance on one language pair.

As demonstrated by Ortega's work in 2020, incorporating segmentation and similar language constructs can enhance neural machine translation (NMT) systems for translating low-resource languages, such as Quechua and Spanish [2] (Ortega, 2020). While our study also considers the Quechua-Spanish language pair and leverages the use of Finnish as a related language for improving the NMT model, there are notable differences in our approaches. Firstly, our study focuses on translating Spanish to Quechua instead of the reverse. Additionally, we use a much larger dataset, consisting of 102747 training examples, 12843 testing examples, and 12844 validation examples, compared to Ortega's NMT model that was trained on a significantly smaller dataset of 17,500 parallel sentences and validated with 2500 parallel sentences.

Moreno's submission to the AmericasNLP machine translation shared task demonstrated the effectiveness of transfer learning for low-resource languages using a related language, as shown in their paper [3]. Their main contributions were two-fold: (1) collecting additional datasets from online sources, and (2) pre-training the system on a large Spanish-English dataset before fine-tuning it for the Spanish-Quechua task. These adaptations resulted in a significant improvement in BLEU score (0.56) and a small improvement in chrF (0.007). Additionally, they compared the performance of training on the JW300 dataset (which mainly consists of biblical parallel data) versus training on all available datasets.

4 Approach

4.1 Baseline Model

Since there is no prior work done on using Spanish to Finnish model for transfer learning, we created a model with the same same size and shape as the Spanish-Finnish model but reinitialized and trained from scratch. We run it on 10 epochs and we get the following result:

Model	BLEU Score	chrF Score	Epochs
Baseline	1.2937	0.1932	10

Table 1: Baseline Model.

4.2 Network Architecture

We utilized a transformer for NMT. The core of the transformer consists of a series of encoder and decoder layers, where each layer contains multi-head self-attention mechanisms and feedforward neural networks. The self-attention mechanism allows the model to focus on different parts of the input sequence when generating the output, while the feedforward network applies non-linear transformations to the intermediate representations.

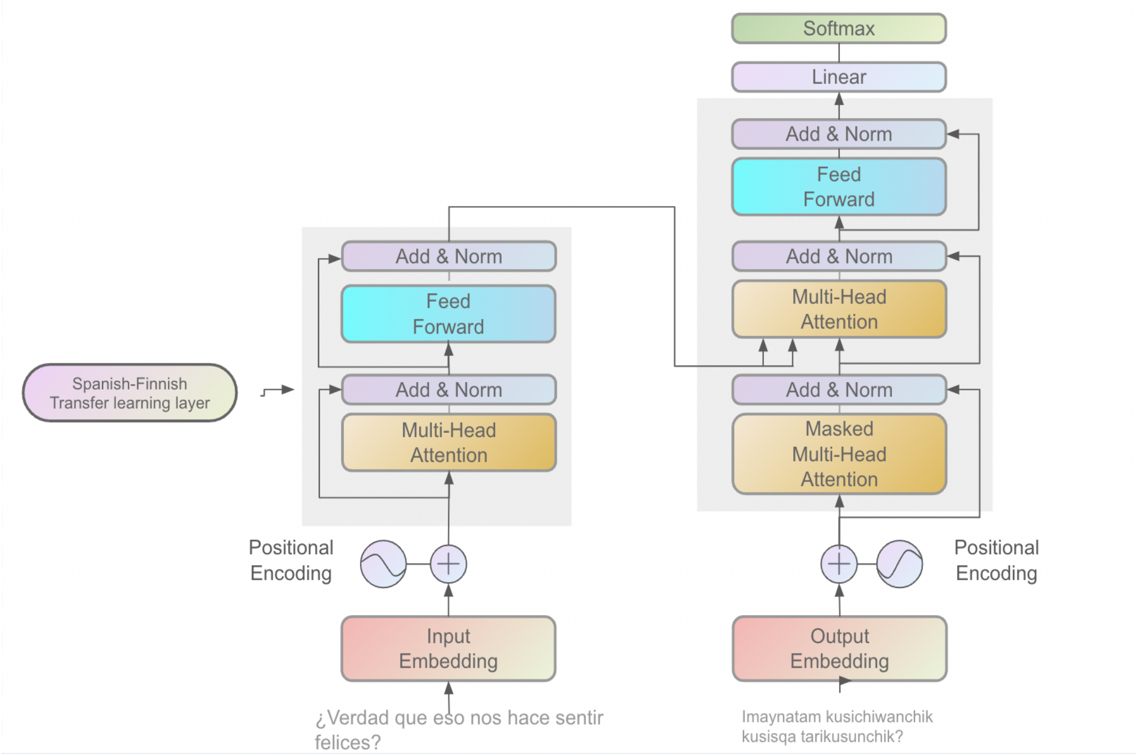

Figure 1: Transformer sequence to sequence model

5 Experiments

5.1 Data

The dataset we used is a recopilation of webs and other datasets and accessed via HuggingFace [4]. It includes "Mundo Quechua" by "Ivan Acuña", "Kuyakuykim (Te quiero): Apps con las que podrías aprender quechua" by "El comercio", "Piropos y frases de amor en quechua" by "Soy Quechua", "Corazón en quechua" by "Soy Quechua", "Oraciones en Español traducidas a Quechua" by "Tatoeba", "AmericasNLP 2021 Shared Task on Open Machine Translation" by "americasnlp2021".

The "AmericasNLP 2021 Shared Task on Open Machine Translation" contains a large part of the dataset we used since it was recently compiled by AmericasNLP and it contains three subdatasets. JW300 (quy, quz, en): texts from the religious domain available in OPUS. JW300 has 121k sentences. MINEDU (quy): Sentences extracted from the official dictionary of the Ministry of Education in Peru (MINEDU). This dataset contains open-domain short sentences. Finally, there is Dict_misc (quy): Dictionary entries and samples collected and reviewed by Huarcaya Taquiri (2020). We can see that the dataset is skewed towards religious texts and we are planning to address it in the final submission of the paper. The dataset was manually cleaned during compilation, as some words of one language were related to several words of the other language.

As a whole, the dataset contains 102747 training examples, 12843 testing examples, and 12844 validation examples. Each example contains an Spanish sentence and its Quechua translation.

5.2 Evaluation method

For the evaluation method, we are first going to use the BLEU score as a standard. The Bilingual evaluation study(**BLEU**) is a standard accuracy metric for machine translation. It calculates the precision of the machine-translated text by comparing it to one or more reference translations. The resulting score ranges from 0 to 1, with 1 indicating a perfect match between the machine translation and the reference translations.

$$BLEU = BP \cdot (\sum_{n=1}^{N} \cdot \frac{\log p_n}{N})$$

where BP is the brevity penalty, which is calculated as follows: BP = 1 if $c > r$ and BP = $e^{(1-r/c)}$ if $c \leq r$.

Here, c is the length of the candidate sentence, r is the length of the reference sentence closest to the candidate in length, p_n is the n-gram precision, which is the ratio of the number of n-grams in the candidate that appear in the reference to the total number of n-grams in the candidate, and w_n is a weight assigned to the n-gram precision to give more importance to higher-order n-grams. The weights are typically set to $1/N$.

However, we found that BLEU score is not comprehensive. chrF (character F-score) is a metric for evaluating the quality of machine translation output, which is based on the F1-score of character n-grams (substrings of length n) in the reference and the translation. Here is the equation for computing the chrF score.

$$\text{chrF} = \frac{(1+\beta^2) \cdot \text{precision}n(r,t) \cdot \text{recall}n(r,t)}{\beta^2 \cdot \text{precision}n(r,t) + \text{recall}_n(r,t)}$$

where β is a parameter that controls the trade-off between precision and recall, n is the length of the character n-grams, r is the reference text, and t is the translation.

One advantage of chrF is that it takes into account the character-level information, which can be helpful for agglutinative languages like Quechua. These languages often have complex morphological structures, where words are formed by concatenating morphemes that can change depending on the context. chrF can capture these morphological differences, as it looks at the character-level similarities between the reference and the translation, rather than relying on exact word matches.

Another advantage of chrF is that it is language-independent, which means that it can be used to evaluate the translation quality of any language pair. This is particularly useful for low-resource languages like Quechua, where there may not be enough training data to develop language-specific metrics.

5.3 Experimental details

For all the experiments, we used a Transformer-based model. Our first set of experiments involved training a neural machine translation model to translate Spanish to Quechua using transfer learning. We started with a pre-trained model and preprocessed the data with MarianTokenizer and fine-tuned the model using Seq2SeqTrainingArguments. We initially tried to use different tokenizers for Spanish and Quechua, but later we met errors in the code and it turns out we need to use the tokenizer from the original Spanish to Finnish model. We didn't change much about the transformer architecture and we used the original openNMT model with a small variation of hyperparameter settings. Later we experimented with different parameters such as learning rate and batch size to find the best ones. We used a learning rate of 2e-5, a weight decay of 0.01, and a batch size of 32. We did not use the backtranslation technique common for low resource language translation because there isn't a Quechua to Spanish NMT model available in the mean time. The running time for each experiment is about four hours.

For the second set of experiments, we froze the encoder during the first half of training epochs and later turned on the encoder. This is common known as gradual unfreezing or progressive unfreezing. Surprisingly, this approch has some exciting results for both the BLEU score and chrF score. We will discuss the result of this approach in the Analysis section.

5.4 Results

The score of each experiment is described in this table.

Model	BLEU Score	chrF Score	Epochs
Baseline	1.2937	0.1932	10
Pretrained model with encoder frozen over all epochs	11.0789	0.3702	10
Pretrained model with gradual unfreezing	12.473	0.3869	10

Table 2: Comparison of Models.

6 Analysis

6.1 Transfer Learning

The result shows that using transfer learning from Spanish to Finnish NMT model has a big improvement compared to the baseline (+9.7852 in BLEU and 0.177 in chrF). The increase in chrF means that we have a great improvement because chrF is considered to be most important for evaluating Quechua translation. Even though the BLEU score is still pretty low for the translation to be considered a good translation, we will provide a detailed analysis in the third section of why some translations performed better and some not so.

We think that the transfer learning performs well because Finnish, like Quechua, is an agglutinative language, which means that words are formed by stringing together morphemes. By learning how to translate Spanish to Finnish, the NMT model can gain experience with handling this type of morphology, which could also be helpful for translating Spanish to Quechua. Overall, transfer learning on Spanish to Finnish translation can provide a useful starting point for training an NMT model for Spanish to Quechua translation, as it can help the model learn how to handle some of the linguistic features that are common to both Finnish and Quechua, but are different from Spanish.

6.2 Gradual Unfreezing

The evaluation metrics for the model with unfrozen encoder after 5 iterations showed a lower evaluation loss, higher BLEU score (+1.3941), higher chrF score(+0.0167), and higher generated sequence length compared to the model with a frozen encoder for the entire 10 iterations. These metrics suggest that the unfrozen model was able to generate more accurate and diverse translations compared to the frozen model.

We think the reason why it performs better is because freezing the encoder during the entire training may limit the model's ability to learn new representations from the training data, especially when the fine-tuning dataset is relatively small. By unfreezing the encoder after some initial iterations, the model can continue to learn from the training data and improve its performance.

6.3 Error Analysis

For the Spanish to Quechua translation task, the metric used for evaluating the quality of translation, BLEU score, has been consistently low. The low BLEU score suggests that the machine translation systems have been struggling to achieve high translation accuracy for this particular task. Possible reasons for the low BLEU score and low translation accuracy could be the complex grammar and vocabulary of the Quechua language, the scarcity of parallel data for training machine translation models, or the limitations of the machine translation algorithms themselves. However, there are some translations that are close to correctness (judged by us) and we want to discuss the possible reasons why the model has made the error. Here are two examples from the validation set.

Source Sentence: El 10 de septiembre de 1823 comenzó la traducción propiamente dicha. (Google Translation to English: On September 10, 1823, the translation properly began.)
Reference Translation: *1823 wata 10 punchaw septiembre killapim tikrayta qallaykurqaku.*
NMT Translation: *1823 watapi 10 de septiembre killapim qallaykurqa.*

The error here is that the NMT cannot translate the work "tikrayta" which means translation (no pun intended). Therefore it only translates the verb "began" or "commence" but did not recognize the noun. The error here might because the training dataset is too small and the NMT hasn't learned the work translation.

Source Sentence: ¿Verdad que eso nos hace sentir felices? (Google Translation to English: Doesn't that make us feel happy?)
Reference Translation: *¿Manachu kayqa kusikunapaq?*
NMT Translation: *¿Imaynatam kusichiwanchik kusisqa tarikusunchik?*

The error here is that the reference sentence is posting a rhetorical question. However, the NMT translation is asking how instead of what. The error here might because the NMT hasn't recognized sufficient sentence structures such as rhetorical questions and therefore whenever it sees a question mark (or two as one inverted and another question mark for two in all together are used in Spanish and Quechua), it will output "how".

7 Conclusion

To summarize, our study proposes a transfer learning method that utilizes Finnish, a related agglutinative language, to enhance machine translation performance for the Spanish-Quechua language pair. Our experiments indicate that transfer learning can increase BLEU and chrF scores for low-resource languages such as Quechua-Spanish, and progressive unfreezing can further improve the fine-tuned model's performance.

However, we were unable to implement data augmentation techniques such as backtranslation due to the lack of available translation models. In future work, we plan to single out Bible texts from the training data and evaluate the model's performance solely on this data, as it represents a significant portion of Quechua's data. Additionally, we only experimented with unfreezing the encoder for half the time, and we intend to test the effectiveness of unfreezing earlier or later. Our hypothesis is that unfreezing later would be more effective, as unfreezing too early or too frequently could lead to overfitting on the fine-tuning dataset.

References

[1] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation, 2016.

[2] John E. Ortega, Richard Castro Mamani, and Kyunghyun Cho. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346, dec 2020.

[3] Oscar Moreno. The REPU CS' Spanish–Quechua submission to the AmericasNLP 2021 shared task on open machine translation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 241–247, Online, June 2021. Association for Computational Linguistics.

[1] [2] [3] [4]Hugging Face. "hackathon-pln-es/spanish-to-quechua." Hugging Face Datasets, 2021, https://huggingface.co/datasets/hackathon-pln-es/spanish-to-quechua.