Model for Real-Time Subtitling from Spanish to Quechua Based on Cascade Speech Translation

Abraham Alvarez-Crespo, Diego Miranda-Salazar and Willy Ugarte[a]
Universidad Peruana de Ciencias Aplicadas, Lima, Peru

Keywords: Speech Recognition, Machine Translation, Quechua Revitalization.

Abstract: The linguistic identity of many indigenous peoples has become relevant in recent years. The speed with which many of these have been lost faces many countries of the world with a serious reduction of their cultural heritage. In South America, the critical situation of vulnerability of many of its native languages is alarming. Even languages such as Quechua, widely spoken in the region, face an early disappearance due to a low rate of inter-generational transmission. Aware of this problem, we have proposed the development of a translation and subtitling system for short film videos from Spanish to Quechua. The proposal contemplates the use of cinema to promote and retain the language. For its realization, we have built a solution that combines a Spanish Voice Recognition system and our proposal for a Quechua Machine translation model. This will be integrated with a desktop application that will also subtitle the film videos. In the tests we carry out, we have obtained better translation indicators than past proposals; in addition to the validation of Quechua-speaking users of the tool's value. Aware of this problem, we have proposed a speech-to-text translation model that could be used as a resource for language revitalization. For its realization, we developed a cascade architecture that combines a Spanish speech recognition module and our proposal of a Quechua machine translation module, fine-tuned from a Turkish NMT model and a parallel public dataset. Additionally, we developed a subtitling algorithm to be joined with our solution into a real-time subtitling desktop application for clips of films. In the tests we carry out, we have obtained better BLEU and chrF scores than previous proposals; in addition to the validation of the translation returned in the subtitles by the Quechua speakers consulted.

1 INTRODUCTION

Numerous nations and international organizations have recently stepped up their efforts to safeguard, preserve, and promote the thousands of indigenous languages that risk extinction by the turn of the century. Proof of this can be found in UNESCO's announcement of an exclusive decade for the rescue of several of these. However, the conflict's diverse nature deteriorates in some regions. The risk scenario for numerous languages on the American continent was already covered in (Nordhoff and Hammarström, 2012). Due to this, many South American countries have started taking steps to conserve the linguistic diversity of their citizens. These haven't, however, had a significant influence on lowering vulnerability in nations like Peru. Furthermore, recent research[1] reveals how poorly these are promoted and revitalized. With a populace that is only familiar with less than a fifth of the 47 indigenous languages spoken in their nation and that is increasingly inclined to give up their native languages in favor of Spanish. As a result, the situation is severe for many minority indigenous languages. Even for one of the largest in the region such as Quechua, which is close to 10 million speakers in the Andean regions of countries such as Colombia, Ecuador, Bolivia, Argentina and Peru (Rios, 2011). And it's because of this that it faces a significant disappearance problem due to its low rate of intergenerational transfer.

This makes the preservation of this and other languages important. A significant loss of its cultural and historical value results from the population abandoning its native language. Therefore, it is essential to develop methods that allow the revitalization of Quechua. For this research, we will focus on the standard version of Southern Quechua, specifically the Quechua Chanka variety. We will take advan-

[a] https://orcid.org/0000-0002-7510-618X

[1] Encuesta Nacional: Percepciones y actitudes sobre diversidad cultural y discriminación étnico-racial, IPSOS, 2018

Alvarez-Crespo, A., Miranda-Salazar, D. and Ugarte, W.
Model for Real-Time Subtitling from Spanish to Quechua Based on Cascade Speech Translation.
DOI: 10.5220/0011783300003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 837-844
ISBN: 978-989-758-623-1; ISSN: 2184-433X

tage of the power of computational translation models to break the state of linguistic isolation to which the Quechua population has been exposed in such important aspects as education, health, or entertainment.

The work is challenging, though, because there have only been a limited number of studies done in languages with a nature similar to their own. Due to the fact that Quechua is an agglutinative language, which means that it heavily relies on morphemes added as suffixes (Rios, 2011). And with subject-object-verb (SOV) grammatical ordering, it has a disadvantage over more widely used languages like Spanish or English. Because of this, our approach is built on a less promising area of research and with a lower threshold for success. Another point is the low-resource pain of many indigenous languages, Quechua included (Mager et al., 2018). Which refers to languages with parallel corpora that contain fewer than 0.5 million parallel sentences (LeCun et al., 2015). This characteristic has a negative effect on the results and the capability to explore contemporary solutions that required big amount of data.

There have been small-scale non-computational and computational initiatives in the past aimed at reviving Quechua. Regarding the first group: There are various educational or social programs run by the Peruvian government to assimilate the Quechua community (Hornberger and Coronel-Molina, 2004; Sumida Huaman, 2020). However, they are unable to provide written or spoken resources in Quechua, organize plans with precise goals or implement policies that encourage inclusion rather than segregation. Regarding the second group, recent years have seen the growth of datasets and limited machine learning techniques addressing Quechua (Oncevay, 2021; Chen and Abdul-Mageed, 2022; Edgar et al., 2012; Ortega and Pillaipakkamnatt, 2018) speech synthesis, text translation systems, and voice recognition.

To overcome the task we will incorporate natural language processing (NLP) technologies for the Quechua subtitling of cinematic content as educational tools that support the language's preservation. Our suggested proposal converts the video's dialogues into waveform representations in the first place. Then a speech recognition model(SR) based on Wav2Vec (Baevski et al., 2020) will analyze to determine the content and convert it into written discourse representations. The data will then be numerically transformed using the SentencePiece tokenizer before our suggested machine translation model (MT) that uses the Transformer-base model (Vaswani et al., 2017) translates each dialogue and is then applied to the process of subtitling the original movie.

Our main contributions are as follows:

- We design an integrated cascade architecture for speech-to-text (S2T) translation between Spanish and Quechua languages.
- We train a new machine translation model using the transfer learning method.
- We develop a desktop application to subtitle clips of movies to Quechua.

This paper is organized as follows. Therefore, in Section 2, an analysis of the state of the art considered for this work will be made. Section 3, first, introduces the technologies used for the development of the proposed solution and then, describes the whole process of deploying the software. Finally, Section 4 show the experimental protocol, the results obtained, and the discussion. To conclude with Section 5

2 RELATED WORKS

The challenge of converting acoustic voice signals into text or even audio is intricate and multifaceted. The main method has been the cascaded model built upon machine translation (NMT) and voice recognition (ASR) systems. As a consequence, for a long time, the advancement in the literature has been driven by general advances in the ASR and NMT models as well as a transition from the loosely coupled cascade's most fundamental form to one with a tighter coupling. Although end-to-end trainable encoder-decoder models and other new modeling techniques have lately generated a number of changes in the field's perspective (Jia et al., 2019). Despite this, the literature's effectiveness and strength are correlated with dominant languages. The field hasn't been studied or developed enough for many indigenous dialects.

In (Nayak et al., 2020), the authors take an application-focused approach to the field. Its application to film dubbing is discussed, and the aspect of source and destination conversation length alignment suggested by another work is addressed. Its encoding module is where it makes a contribution to literature, composed of attention modules that execute the temporal alignment of the word with the time sequence of the video and recurrent networks that catch audio sequences. However, in our work, we make use of a part of the voice recognition module that records the temporality of the dialogues and subsequently permits the subtitling. By doing this, we can completely ignore the video processing and concentrate on the audio.

In (Tjandra et al., 2019), it is suggested to train speech-to-speech translation systems without language supervision using complicated transformer networks. The suggestion alludes to works where attention-layer networks bypass a phase of the traditional cascade to do direct translations with fewer modifications and, thus, a greater awareness of non-lexical voice metadata. His contribution concentrates on making them more effective while concentrating on words that can't be translated linguistically. For this, a red transformer that performs direct translation is trained using unsupervised discrete representations of the vocal signals of the two languages.

In (Kano et al., 2020), the authors examine end-to-end translation in languages with different syntactic structures. The experiment highlights the extra effort needed to translate words in a different order and how it affects the outcomes when compared to pairings of syntactically related languages. The work's recommendation is to create an encoder-decoder architecture for ASR and NMT systems trained on curriculum learning, in order to take into account the syntactic distinctions between the two studied languages (en - jp). In this instance, we've used transfer learning on a model that was previously trained using pairings of syntactically various languages.

In (Oncevay, 2021), the authors propose neural machine translation modules for indigenous languages of South America. The proposal considers the tokenization and translation procedure. Through testing and fine tweaking, SentencePiece and the transformer basis model are utilized to identify the ideal parameters for many languages. They do not, however, account for the syntactic variation of the model that was previously trained in en-es. In this instance, we employ a trained model of language pairs that are more syntactically related to es - quy.

3 CASCADE SPEECH TRANSLATION FROM SPANISH TO QUECHUA

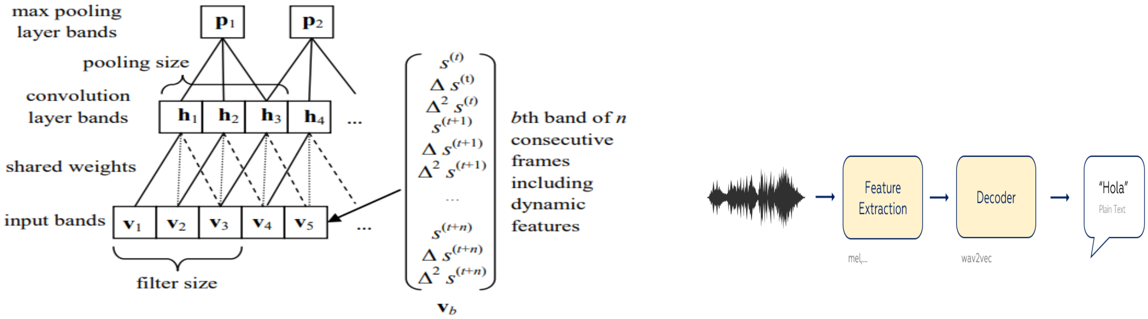
3.1 Preliminary Concepts

We have read up on speech translation (ST) systems in order to design the proposed solution. These systems use deep learning algorithms to take the speech as a starting point and, depending to the final representation could perform a speech-to-text(S2T) or speech-to-speech(S2S) translation. In recent years, the state of the art has been turning between the cascade's levels reduction and the addition of paralinguistic information in the translation process. For instance, traditional ST models examined in (Baldridge, 2004), used to divide the translation task into speech recognition (SR), machine translation (MT), and text-to-speech (TTS); while newer approaches try to avoid one side of the text conversion (Tjandra et al., 2019) or even train the models only with parallel audio corpora (Jia et al., 2019). Although, this is possible only for the most widely spoken languages. The least studied languages still need to improve each level of the cascade; with even higher priority, increase data collection.

Deep Learning: The concept refers to models that are made up of several processing layers, to learn representations of data at various levels of abstraction. With the use of this technique, the state-of-the-art in important areas for the creation of voice translation models, such as speech recognition, automatic translation, or speech synthesis, has greatly advanced. By employing the backpropagation technique to suggest changes to a machine's internal parameters, deep learning may uncover detailed structures in massive data sets (LeCun et al., 2015). By employing the backpropagation technique to suggest changes to a machine's internal parameters that are used to calculate the representation in each layer from the representation in the previous layer, deep learning may uncover detailed structures in massive data sets.

Convolutional Neural Network (CNN): One or more pairs of convolutional and maximum pooling layers make up a CNN. A convolution layer applies a series of filters that duplicate themselves throughout the whole input space and process discrete local portions of the input. By extracting the maximum filter activation from various points inside a given window, a max-pooling layer produces a lower resolution version of the convolution layer activations (Abdel-Hamid et al., 2012). This increases tolerance for little variations in the locations of an object's component pieces and translation invariance. Higher layers handle more complicated portions of the input using broader filters that operate on lower-resolution inputs. In order to classify the total inputs, the top fully linked layers aggregate inputs from all locations.

Speech Recognition: The idea alludes to AI's capacity to identify speech-related vocalizations in human speech. Replicate voice signals as a string of words using computational models, in other words. Speech Recognition (SR) evolved and has been made possible by the use of statistical and neural modeling techniques, which have made it possible to recognize and comprehend lexical speech data in a variety of languages and practical settings. The emphasis of these systems has also been shifted to include features of the speaker's identity in addition to interpreting speech content. In any case, Figure 1 illustrates the general methodology used for speech recognition, which starts with feature extraction by making use



(a) A pair of CNN convolution layers and max-pooling layers are displayed in the diagram, where weights shared by all convolution layer bands are indicated by the same line style (Abdel-Hamid et al., 2012).

(b) General Architecture of Speech recognition.

Figure 1: Architectures of CNN and Speech recognition.

of feature extraction methods such as Wavelets, etc. Once they're obtained, a decoder is employed to detect the speech (Haridas et al., 2018)

Machine Translation: The phrase alludes to the capacity of computers to automatically translate human languages. It was based on simple rule systems for a long time. Today, there exist systems with probabilistic or neural models that are more effective thanks to new parallel and non-parallel data sources. As shown in Figure 1, a solution for Machine Translation is the application of a neural machine translation encoder-decoder framework. The encoder network converts each input token of the source-language phrase into a low-dimensional real-valued vector (word embedding) and then encodes the sequence of vectors into distributed semantic representations, which the decoder network creates token by a token from left to right (Zhang and Zong, 2020)

Transformer Model: The Transformer is a model architecture that foregoes recurrence in favor of drawing global relationships between input and output via an attention mechanism, it allows for substantially higher parallelization and can achieve new levels of translation quality after a few hours. The Transformer is the first transduction model to generate representations of its input and output using just self-attention rather than sequence-aligned RNNs or convolution (Vaswani et al., 2017).

Figure 2 illustrates the encoder-decoder structure of most neural sequence transduction models; the encoder in this example translates an input series of symbol representations (x1,..., xn) to a sequence of continuous representations z = (z1, ..., zn). The decoder, given z, produces a symbol output sequence (y1,..., yn) one element at a time. This fundamental architecture is followed by the Transformer (see Figure 2), which, as shown in the image, utilizes layered self-attention and point-wise, entirely connected layers for both the encoder and decoder.

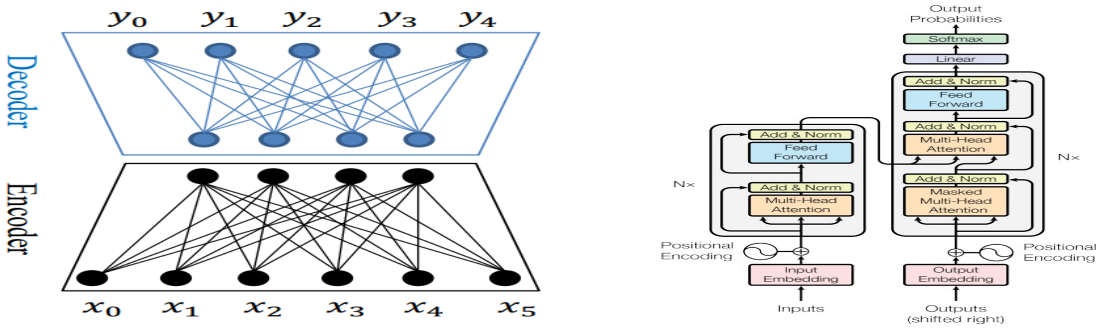
3.2 Method

Our work's key contribution is the cascade structuring of NLP systems for speech-to-text (S2T) translation between Spanish and Quechua (shown in Figure 3) and the application of learning transfer to previously trained machine translation models. The first was with the Spanish (sp) - Finnish (fi) language pair, and the second was a multilingual model for translating from Italic languages(itc), Spanish included, to Turkish (tr).

Cascade Architecture: The cascade approach to Speech Translation (ST) is based on a pipeline that concatenates an Automatic Speech Recognition (ASR) system followed by a Machine Translation (MT) system. This setup generates efficiency in solving morphological problems.

Speech Recognition: In this manner, we spread the work in a Spanish speech recognition module that converts voice inputs into a string of words while maintaining the lexical richness of the original inputs. We used the Spanish-speaking wav2vec supervised model of automatic voice recognition for this challenge. As is shown in Figure 4, audio is converted into a two-dimensional waveform (a signal representative of the audio). Then, this source entered a convolutional network (CNN) to be converted into a latent representation. Finally, a supervised red transformer that forecasts a result using linear projections will contextualize the latent data.

Machine Translation: It is necessary to translate a discussion into Quechua after we have identified it. Therefore, an attention-based neural network translation module is now applied to the expected text. SentencePiece is utilized to tokenize the Quechua datasets for this task. After this, we used transfer learning to the Transformer-base model (Vaswani et al., 2017) with the default configuration in Marian NMT (Junczys-Dowmunt et al., 2018) to fine-tunned



(a) A neural machine translation encoder-decoder framework. The encoder converts the input sequences into distributed semantic representations, which the decoder uses to generate the output sequence (Zhang and Zong, 2020).

(b) Transformer model architecture (Vaswani et al., 2017).

Figure 2: Architecture models.

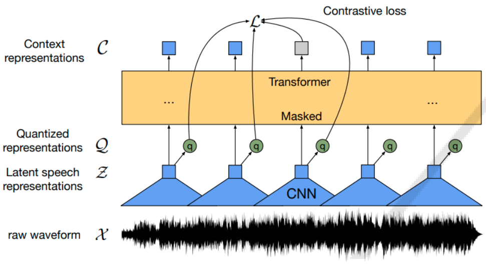


Figure 3: Wav2Vec Architecture: Speech Recognition model.

the network, in order to get the contextual representation which will be aligned to the most propensity Quechua result. The fundamental principle of transfer learning is to use the parent model's parameters as the starting point for our proposal rather than beginning from scratch and randomly initializing the parameters (Zoph et al., 2016; Nguyen and Chiang, 2017).

4 EXPERIMENTS

We will go over the experiments that our project has gone through in this section. The datasets were obtained from many domains and freely available sources.

The public corpus of Spanish (es) and Quechua Ayacucho (quy) offered by AmericasNLPs[2] has shown to be the most pertinent (Agic and Vulic, 2019).

4.1 Experimental Protocol

In order for the studies to be repeated by others, we will describe the setup and procedures used.

Development Environment: About the environment in which the experiments were developed, we used our own computers. They had an Intel Core I5 processor, 32GB of RAM, and one GPU on an NVIDIA GeForce RTX 3060. We used python to create the models, along with Pytorch, its deep-learning framework for Nvidia graphics cards. A large number of additional libraries were also used, including sentencepiece's tokenizer, evaluate's evaluation metrics, and transformers for the transfer learning application.

Dataset: The dataset that was used for the experiments came from open sources. They were all parallel corpora for machine translation. The data set includes Spanish text that has been translated into southern Quechua standard (Quechua chanka). A variant that is spoken in several parts of Peru. The AmericasNLP Shared Task included the two primary training datasets. The first, JW300 (Agic and Vulic, 2019) is made up of Jehovah's Witnesses writings and is available in OPUS, while the second contains official dictionaries from the Ministry of Education and other sources. We also, clean the datasets because they were noisy and not cleaned. Lines are reduced according to several heuristics: removal of URLs, numbering of sentences or book appendages, a sentence that has more symbols or numbers than actual words, or it may have a word ratio where one side's words are five times longer or shorter than the other, etc. The result of this process is shown in Table 1.

Models Training: All model experiments were run using Pytorch in a physical GPU environment with 32 GB of RAM, for 8 epochs using batches of 16 tuples per dataset, with the Adam optimizer and try-

[2] Available in https://github.com/AmericasNLP/americasnlp2021

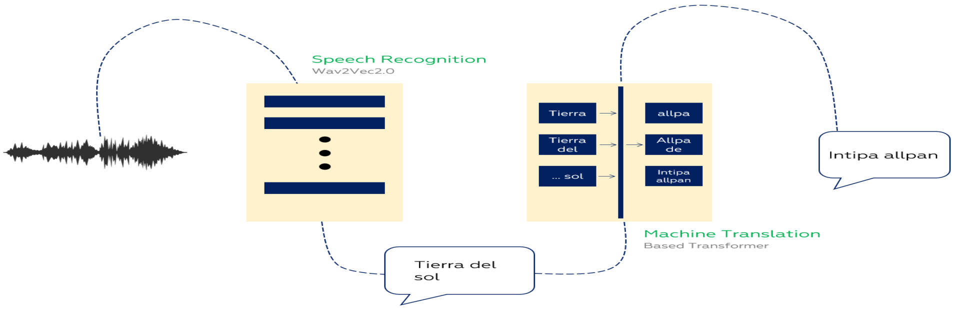


Figure 4: Proposed S2T Translation Architecture.

Table 1: Statistics and cleaning for all parallel corpora.

Corpus	S (orig.)	S (clean)	% clean
JW300	125,008	103,293	-17.4%
DICT	9,643	5,925	-38.6%

Table 2: Number of parallel sentences per split.

Corpus	Train	Test
JW300	92,964	10,329
DICT	5,333	593

Table 3: Testing Results.

		BLEU	chrF
	JW300		
(a)	TL Bilingual	10.48	49.63
(b)	TL Multilingual	12.78	51.20
(c)	Baseline	10.04	44.60
	DIC		
(d)	TL Bilingual	3.48	30.22
(e)	Baseline	9.55	40.33

ing to minimize the loss value, taking an average of 80 minutes per epoch with the most extensive dataset. JW300 training typically took 9 hours, whereas DIC training only required 3 hours.

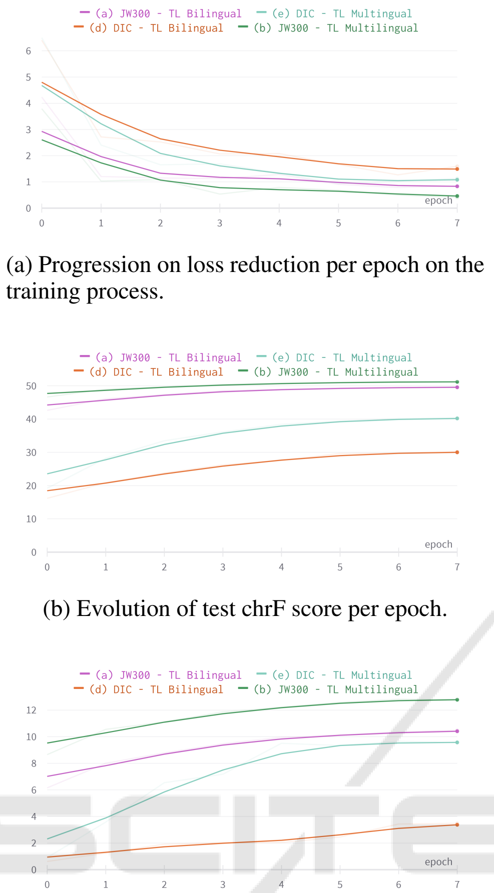
Evaluation: To qualify the performance of the models, we utilized the BLEU (Papineni et al., 2002) and chrF (Popovic, 2015) scores to evaluate them. The goal is to reach the highest score to validate an efficient translation. We used Wandb to get a full review of the executions and their particular configurations.

Models Training: The entire code and dataset accord the proposed solutions and experiments is publicly available at the following repository: https://github.com/Malvodio/CinemaSimi.

4.2 Results

To compare our results we set a baseline model that takes the best scores obtained on (Oncevay, 2021). Because to the best of our knowledge, this is the one who achieve the highest performance in evaluation metrics of Quechua machine translation and they also used the datasets provided by AmericasNLP Shared Task. This baseline was built with a 6 self-attention layer (Vaswani et al., 2017) and 8 heads to first pretrain it with a Spanish-English parallel corpus and then finetune it with 10 indigenous language pairs; to produce a multilingual MT model. For our models, we leverage publicly accessible pre-trained models from Huggingface (Wolf et al., 2020) as provided by Helsinki-NLP (Tiedemann and Thottingal, 2020). The pre-trained MT models released by Helsinki-NLP are trained on OPUS, an open-source parallel corpus (Tiedemann, 2012). Underlying these models is the Transformer architecture of Marian-NMT framework implementation (Junczys-Dowmunt et al., 2018). Each model has also 6 self-attention layers in the encoder and decoder parts, and each layer has 8 attention heads. The models we specifically use were pretrained with OPUS Spanish-Finnish data and Spanish and other italic languages-Turkish data. We choose this model because their source language is Spanish so they will have good Spanish subword embeddings. And for the Quechua, we take the considerations mentioned by (Ortega and Pillaipakkamnatt, 2018). They point out that the task needs a similar agglutinate language with a similar word order. In our investigations, the Finnish and Turkish fit with it. Following recommended practices, we used a uniform sampling on our datasets to avoid under-fitting the low-resource condition of our pair language. For the experiments, each dataset was divided into 90% training data and 10% testing data. In Table 2, the volumes are shown.

The results are shown in Table 3 and as is seen our model gets better scores than the selected baseline. We also add a graph that shows the loss reduction of



(a) Progression on loss reduction per epoch on the training process.

(b) Evolution of test chrF score per epoch.

(c) Evolution of test BLEU score per epoch.

Figure 5: Evolution of metrics per epoch.

the model around the training process (see Figure 5). While Figures 5 and 5 show the gain over the BLEU and chrF score by epochs of the validation process.

4.3 Discussion

One of the most interesting outcome on the performance of the models was the impact of the dataset. The lower scores on the experiments over the DICT dataset were not expected. The results show that, the training end with a high underfitting. We wonder if this was caused for the reduced number of sentences. And obviuous the extremely lower resource state of this dataset don't let the model learn over a whole domain of the language. But we think that the reason why the model can't even learn in its same domain could be the low average words per sentence, less than a third of the JW300 average.

On other hand, the translation scores obtain over the JW300 dataset with both proposed models, the multilingual (itc - tr) and the bilingual (es - fi), are better than the base model choosed, as shown in Table 3. This positive results could indicated that the strategy used to approach this work was successful. Concentrate our effors on fine-tunning existing solutions that have similarities with the problem addressed work highly enough to sustain our hipotesys. Being aware of the limitations that Quechua or other indigenous languages have, allows us to optimize learning on models that have already effectively learned similar aspects of these languages. The based model also applies transfer learning, but between a pair of languages (es - en), which are syntactically and typographically distinct from Quechua. These approuch might don't let his model learn well the nature of the language. Another interesting point is the distinctions between the bilingual and multilingual models results over JW300. We think the lower results of the first is caused for the syntactic ordering type of these languages. Because if we noticed, both Finnish and Turkish are agglutinating languages. But just the last has the same syntactic ordering as Quechua.

Quechua translation scores are still well below what is expected for solutions involving widely spoken languages. This keeps the task very challenging and highlights the need for a more complete and extensive corpus of Quechua. Finally, the differences between the BLEU and chrF scores are seen in the way they measure translation efficiency, while the former is a word-level standard, the latter allows us to assess the character level. Very worthy considering the agglutinative nature of Quechua.

5 CONCLUSION

We conclude that the results of our best model validate the approach proposed, even despite the ”low resource” limitation of the language. Using transfer learning strategy allowed us to retrain models that share similarities with the quechua speech translation instead of developing them from scratch. There is a chance to improved the tokenization with a cleaner corpus and define a estrategy for out-of-domain values could also improved the results. In future works, we hope to be able to close the translation cycle by adding the speech synthesis module to the proposed architecture; so that we can get the dubbing of a movie. We also want to explore the translation of other Quechua variants, using robotic interfaces (Burga-Gutierrez et al., 2020) or generating text in Quechua (de Rivero et al., 2021).

REFERENCES

Abdel-Hamid, O., Mohamed, A., Jiang, H., and Penn, G. (2012). Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *IEEE ICASSP*.

Agic, Z. and Vulic, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *ACL*.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*.

Baldridge, J. (2004). *Verbmobil: Foundations of Speech-to-Speech Translation*. *Nat. Lang. Eng.*, 10(2).

Burga-Gutierrez, E., Vasquez-Chauca, B., and Ugarte, W. (2020). Comparative analysis of question answering models for HRI tasks with NAO in spanish. In *SIMBig*.

Chen, W. and Abdul-Mageed, M. (2022). Improving neural machine translation of indigenous languages with multilingual transfer learning. *CoRR*, abs/2205.06993.

de Rivero, M., Tirado, C., and Ugarte, W. (2021). Formalstyler: GPT based model for formal style transfer based on formality and meaning preservation. In *KDIR*.

Edgar, J., Muñoz, V., Antonio, J., Tello, C., Alexander, R., and Castro Mamani, R. (2012). Let's speak quechua: The implementation of a text-to-speech system for the incas' language. In *IberSPEECH*.

Haridas, A. V., Marimuthu, R., and Sivakumar, V. G. (2018). A critical review and analysis on techniques of speech recognition: The road ahead. *Int. J. Knowl. Based Intell. Eng. Syst.*, 22(1).

Hornberger, N. H. and Coronel-Molina, S. M. (2004). Quechua language shift, maintenance, and revitalization in the andes: the case for language planning. *International Journal of the Sociology of Language*, 2004(167).

Jia, Y., Weiss, R. J., Biadsy, F., Macherey, W., Johnson, M., Chen, Z., and Wu, Y. (2019). Direct speech-to-speech translation with a sequence-to-sequence model. In *ISCA INTERSPEECH*.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *ACL*.

Kano, T., Sakti, S., and Nakamura, S. (2020). End-to-end speech translation with transcoding by multi-task learning for distant language pairs. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28.

LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep learning. *Nat.*, 521(7553).

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza-Ruíz, I. V. (2018). Challenges of language technologies for the indigenous languages of the americas. In *ACL COLING*.

Nayak, S., Baumann, T., Bhattacharya, S., Karakanta, A., Negri, M., and Turchi, M. (2020). See me speaking? differentiating on whether words are spoken on screen or off to optimize machine dubbing. In *ACM ICMI Companion*.

Nguyen, T. Q. and Chiang, D. (2017). Transfer learning across low-resource, related languages for neural machine translation. In *AFNLP IJCNLP(2)*.

Nordhoff, S. and Hammarström, H. (2012). Glottolog/langdoc: Increasing the visibility of grey literature for low-density languages. In *ELRA LREC*.

Oncevay, A. (2021). Peru is multilingual, its machine translation should be too. In *Workshop on NLP for Indigenous Languages of the Americas*. ACL.

Ortega, J. and Pillaipakkamnatt, K. (2018). Using morphemes from agglutinative languages like quechua and finnish to aid in low-resource translation. In *AMTA LoResMT@AMTA*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Popovic, M. (2015). chrf: character n-gram f-score for automatic MT evaluation. In *ACL WMT@EMNLP*.

Rios, A. (2011). Spell checking an agglutinative language: Quechua. In *5th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Fundacja Uniwersytetu im. A. Mickiewicza.

Sumida Huaman, E. (2020). Small indigenous schools: Indigenous resurgence and education in the americas. *Anthropology & Education Quarterly*, 51(3).

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *ELRA LREC*.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT - building open translation services for the world. In *EAMT*.

Tjandra, A., Sakti, S., and Nakamura, S. (2019). Speech-to-speech translation between untranscribed unknown languages. In *IEEE ASRU*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *ACL EMNLP (Demos)*.

Zhang, J. and Zong, C. (2020). Neural machine translation: Challenges, progress and future. *CoRR*, abs/2004.05809.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *ACL EMNLP*.