Instructions for *ACL Proceedings

Anonymous ACL submission

Abstract

Neural Machine Translation (NMT) systems for low-resource languages like Mapudungun face significant challenges due to limited training data and linguistic complexities. We aim to improve translation between Spanish and Mapudungun through transfer learning, leveraging pre-trained models on Spanish-English and Spanish-Finnish language pairs. Our contributions include demonstrating the effectiveness of transfer learning in this context and providing a comparative analysis of different parent models. Our main findings show that transfer learning enhances translation performance, with not much of a difference between the Spanish-English and Spanish-Finnish pretrained model performance. This suggests that factors beyond morphological similarity, such as data quality or tokenization methods, play a crucial role in transfer learning success. These insights hope to pave the way for future research into optimizing translation tools for low-resource languages and involving communities in the development process.

1 Introduction

Mapudungun (ISO 639-2 code: `arn`), an indigenous language spoken by the Mapuche people of South America, particularly in Chile and Argentina, faces the threat of extinction due to its lack of recognition by the government and the century-long dominance of Spanish (ISO 639-1 code: `es`) brought about by colonization (Llancao, 2019). Creating a translator may help address its cultural, historical, and linguistic loss.

Given the scarcity of online parallel text data for Mapudungun, typical of "low-resource" languages, this project delves into exploring techniques tailored for such scenarios. While previous efforts for Mapudungun-Spanish translation have mainly focused on corpus development, statistical machine translation, or data augmentation, the explicit application of transfer learning remains largely unexplored (Levin et al., 2018; Duan et al., 2020).

Specifically, our two research questions coming in to the project were:

Part 1: Can transfer learning enhance existing neural machine translation systems between Spanish and Mapudungun?

Part 2: Does a higher level of similarity between the parent and child languages lead to a higher performance of transfer learning into Mapudungun?

To investigate Q1, we compare the results of a previously published baseline model (Duan et al., 2020) that did not utilize transfer learning to results of models we develop that apply transfer learning by fine-tuning from a parent model between Spanish and Finnish. We run these trials in both the Mapudungun source to Spanish (arn→es) and Spanish to Mapudungun (es→arn) translation directions.

To investigate Q2, we apply transfer learning only for the Spanish to Mapudungun (es→arn) direction, but in two trials. In the first trial, we use a Spanish to Finnish parent model, with Finnish (ISO 639-1 code: `fi`) being an agglutinative language like Mapudungun. In the second trial, we use a Spanish-English parent model, with English (ISO 639-1 code: `en`) being a non-agglutinative language. Our hypothesis was that, since Finnish and Mapudungun are both agglutinative languages, the Spanish to Finnish (es→fi) model would serve as a better parent model than the Spanish to English (es→en) model, in terms of higher BLEU and chrF scores signifying higher transfer learning.

2 Background

2.1 Transfer Learning

Machine translation (MT) converts text from a "source" language into output text in a "target" language, and consists of statistical (SMT), rule-based (RBMT), and neural (NMT) methods. While NMT systems excel with extensive training data (as in "high-resource" settings), the scarcity of such data for lower-resourced languages results in poorer NMT performance. Transfer learning is a technique that applies knowledge gained from one task to enhance performance on a related task, thereby reducing the amount of training data needed.

Zoph et al. (2016) provide a straightforward yet effective transfer learning framework for NMT. First, train an NMT model, termed the "parent" model, on a large parallel text corpus (e.g., French to English). Next, initialize another NMT model, known as the "child" model, with the pre-trained parent model and further train it on the small parallel corpus (e.g., Uzbek to English). This strategy improves NMT performance despite the parallel text resource constraints.

Choosing an appropriate parent model is one of the design decisions in order to apply transfer learning for low-resource NMT. Most research suggests that selecting a parent model based on language family is important (Wang et al., 2021). This means utilizing a parent model trained on a high-resource language within the same family as the target low-resource language. Furthermore, the conclusion from the original paper by Zoph et al. (2016) that introduced transfer learning for low-resource machine translation is that the more similar the parent model is to the child language, in terms of distributional characteristics or vocabulary, the better the expected performance.

2.2 Linguistic Considerations for Mapudungun

Linguists classify Mapudungun as an Araucarian language, along with other languages like Huilliche spoken in the same region (Smeets, 1989). Unfortunately, the other languages in this family are also low-resource in terms of their online data presence, limiting the common strategy of using a high-resource language from the same family. So, we aimed to explore the grammar aspect of language families as a similarity criterion. Mapudungun is a highly agglutinative, polysynthetic language, meaning that it makes use of suffixation, compounding and reduplication, with a complex verbal morphology (Smeets, 1989). For example, according to our dataset, the word "tünkülepürakatun" in Mapudungun means "ahora estoy más calmada" in Spanish ("now I am more calm" in English), effectively functioning as a whole sentence.

3 Approach

3.1 Part 1: Exploring the Effect of Transfer Learning

To investigate the effect of transfer learning on translation results in the Spanish to Mapudungun (es→arn) direction, we consider the Spanish-Mapudungun model published by Duan et al. (2020), which was trained from scratch without transfer learning. We refer to this model as the "es→arn baseline training" model.

For comparison, the corresponding model to this baseline is the Spanish-Mapudungun child model we developed by fine-tuning a parent model pre-trained on Spanish-Finnish translation. This model is referred to as the "es→arn transfer learning" condition.

Similarly, to explore the effect of transfer learning on translation results in the opposite direction (arn→es), we refer to Duan's Mapudungun-Spanish model as the "arn→es baseline training" model. The corresponding model for comparison is "arn→es transfer learning", the Mapudungun to Spanish child model we developed by fine-tuning a parent model that was pre-trained on Finnish to Spanish translation.

3.2 Part 2: Exploring the Effect of Parent Agglutinative Morphological Similarity

Next, we investigate the effect of morphological similarity in the parent model language on the performance of the child model. We explore this by comparing transfer learning for Spanish to Mapudungun translation when fine-tuned from two different parent models: one translating Spanish to Finnish (an agglutinative language like Mapudungun) and the other translating Spanish to English (a non-agglutinative language). The Spanish to Finnish parent model is the same pre-trained model accessed on Hugging Face as the one used in the es→arn transfer learning condition in the first part. The Spanish-English parent models are also Opus-MT models developed by the Helsinki NLP group and available on the Huggingface Transformers hub (Spanish-English).

Finnish shares these agglutinative properties and may thus be more related to Mapudungun than other non-agglutinative languages (Tóth, 2007). As an example of this agglutinative property, consider the morpheme split in the following sentence in Mapudungun, English, and Finnish (all translating to "in our houses" in English):

arn: rukameuengün = ruka ('house') + meu ('in') + engün (plural)

en: in our houses = in + our + house + s (plural)

fi: taloissamme = talo ('house') + i (plural) + ssa ('in') + mme ('our')

Note how while, in English, almost every morpheme is its own individual word, in both Mapudungun and Finnish, morphemes are merged together to form words. By choosing Finnish as the parent language, we hoped to exploit these morphological grammatical similarities to improve the translation performance of the Spanish to Mapudungun model.

3.3 Dataset

We construct a dataset using the Mapudungun-Spanish parallel corpus, sourced from the AVENUE project—a collaboration between Carnegie Mellon University and the Chilean Ministry of Education (Levin et al., 2018). This Creative Commons corpus consists of about 260,000 transcribed phrases from Mapudungun conversations and their corresponding Spanish translations (Duan et al., 2020).

The files in the corpus were not always uniformly formatted. Typically, each entry consists of a human-generated transcription phrase in Mapudungun, prefaced by "M:", followed by the human-generated Spanish translation prefaced by "C:". However, variations in formatting were present due to different translators' practices. To standardize the format, we developed a custom regex expression to consistently extract and pair the phrases.

For text cleaning, first, we removed the notes taken by the data transcribers. These included notes like "<*SPA>" to indicate that the Mapudungun speaker used the Spanish word instead of the Mapudungun word, or "<Noise>" for when the audio was not clear. Additionally, we removed duplicate punctuation, any remaining unclosed notes, diacritic typos, and capitalization. Part 1 text cleaning also included ASCII normalization, which removed any diacritics. Spacing was adjusted post-cleaning, and the cleaned dataset was saved in a TSV file with appropriate splits for ease of subsequent loading. The complete dataset consists of 268,256 parallel phrases. However, due to limited computational resources, we use a subset of 10,000 pairs for training and 1,250 pairs each for validation and testing.

We used the extracted cleaned monolingual Mapudungun data to train a new Byte-Pair Encoding (BPE) subword tokenizer (Sennrich et al., 2016) implemented via SentencePiece (Kudo and Richardson, 2018). Subword tokenization is particularly effective for languages like Mapudungun, where words are typically composed of subwords. New tokens discovered in the Mapudungun tokenizer were added to the shared vocabulary of the respective parent model, which is the common approach for low-resource NMT (Wang et al., 2021).

3.4 Model Architectures

Part 1, Baseline Training Model: The baseline results of part 1 are the previously published scores of the Spanish-Mapudungun NMT models by Duan and others (2020), which did not use transfer learning techniques. These models trained from scratch use the same data sizes (10,000) and corpus source (AVENUE). The baseline training models, as described in the paper by Duan and others, each utilized a sequence-to-sequence Transformer architecture (Vaswani et al., 2017), with 5 encoder and 5 decoder layers, an embedding size of 512, a feed-forward transformation size of 2048, 8 encoder and 8 decoder heads, a 0.4 dropout probability (Srivastava et al., 2014), label smoothing set to 0.1, and the Adam optimizer (Kingma and Ba, 2017).

Part 1, Transfer Learning Model: We apply the transfer learning NMT method in the Spanish to Mapudungun direction by fine-tuning from a pretrained Spanish-Finnish parent model, and in each direction of translation (Spanish to Mapudungun and Mapudungun to Spanish). The transfer learning child models we developed by fine-tuning pre-trained Spanish-Finnish and Finnish-Spanish models are Opus-MT models developed by the Helsinki NLP group, available via an Apache 2.0 license on the Huggingface Transformers (Wolf et al., 2019) hub [Spanish-Finnish [1] and Finnish-Spanish [2]], which we treat as the parent models. The transfer learning models are also sequence-to-sequence Transformers (Vaswani et al., 2017), with 6 encoder and 6 decoder layers, an embedding size of 512, a feed-forward transformation size of 2048, 8 encoder and 8 decoder heads, a 0.1 dropout probability (Srivastava et al., 2014), and the Adam optimizer (Kingma and Ba, 2017), total 60M parameters. Due to these slight architectural differences between the baseline training and transfer learning models, we consider our comparisons in part 1 preliminary ones.

[1] https://huggingface.co/Helsinki-NLP/opus-mt-es-fi
[2] https://huggingface.co/Helsinki-NLP/opus-mt-fi-es

Part 2, English Parent and Finnish Parent: The two models compared in answering part 2 are (1) the Spanish to Mapudungun child model fine-tuned from the pretrained Spanish to English [3] parent model and (2) the Spanish to Mapudungun child model fine-tuned from the pretrained Spanish to Finnish parent model. The models in part 2 have the same architecture as those in the part 1 transfer learning models. However, while in part 1, we normalize the characters to ASCII (remove diacritics like tildes), in part 2, we leave the diacritics inside.

3.5 Training Specifications

Part 1, Effect of Transfer Learning: We fine-tuned both the arn→es transfer learning model and the es→arn transfer learning model from their respective parent model on the Mapudungun-Spanish parallel corpus training subset size of 10,000, and with a learning rate of 3e-4, a weight decay of 0.01, and a batch size of 8. We trained for three epochs each on a single GPU on the Google Cloud Platform, which took about 21 minutes for the es→arn child model and 16 minutes for the arn→es child model. The metric for what the best model to save during training was validation set BLEU score, and chrF score was also reported.

Part 2, Effect of Parent Agglutinative Similarity: We fine-tuned the two es→arn models from the Spanish-Finnish pre-trained model and the Spanish-English pre-trained model each on the Mapudungun-Spanish parallel corpus training subset size of 10,000, and with a learning rate of 3e-4, a weight decay of 0.01, and a batch size of 8. We trained for up to 15 epochs each on a single Google Cloud Platform GPU, which took about 1 hour and 45 minutes for the Spanish-English pre-trained model and 1 hour and 53 minutes for the Spanish-Finnish pre-trained model. The metric for what the best model to save during training was validation set BLEU score, and chrF score was also reported.

4 Results

4.1 Part 1: Exploring the Effect of Transfer Learning

The following tables report the test BLEU and chrF scores for 10,000 training examples, comparing the Duan baseline model (trained from scratch for up to 200 epochs) with the transfer learning child model (fine-tuned from the Finnish-Spanish parent model for 3 epochs) for translation in the Mapudungun to Spanish (arn→es) direction and the Spanish to Mapudungun (es→arn) direction.

Model	Test BLEU	Test chrF
Baseline	6.26	10
Fine-tuning	10.03	30

Table 1: Effect of Transfer Learning on arn→es Translation

We observe that, with the same dataset size and original source, the application of transfer learning via a pre-trained parent model leads to higher BLEU (+3.77) and chrF (+20) scores compared to baseline training from scratch in the arn→es direction (Table 1). This aligns with our hypothesis, as we expect some aspects of translating into Spanish to be transferable from the Finnish-Spanish parent model to the child task of translating arn→es.

Model	Test BLEU	Test chrF
Baseline	1.09	10
Fine-tuning	9.77	31

Table 2: Effect of Transfer Learning on es→arn Translation

Similarly, in the es→arn direction (Table 2), transfer learning results in significantly higher BLEU (+8.68) and chrF (+21) scores. This supports our hypothesis that aspects of translating into Mapudungun are transferable from the Spanish-Finnish parent model to the child task of translating es→arn. This finding also aligns with Duan's study, which attributes the lower baseline BLEU score in the Spanish to Mapudungun direction to the difficulty of generating text in a polysynthetic, agglutinative language with limited training data (Duan, 2020).

4.2 Part 2: Exploring the Effect of Parent Agglutinative Morphological Similarity

Table 3 shows the test BLEU and chrF scores for the es→arn translation task using transfer learning from two different parent models: Spanish to English and Spanish to Finnish.

Parent Target	Test BLEU	Test chrF
English	8.07	29
Finnish	8.11	28

Table 3: Effect of Parent Morphological Similarity on es→arn Translation

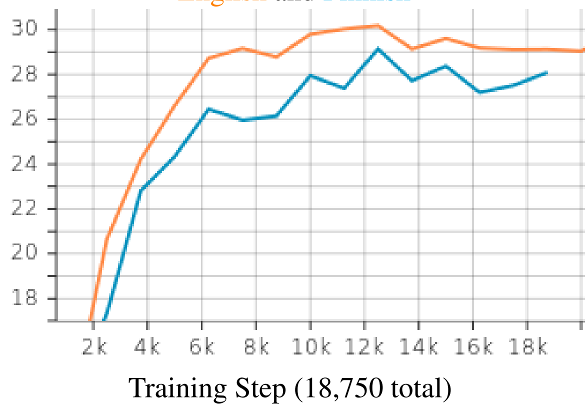
[3] https://huggingface.co/Helsinki-NLP/opus-mt-es-en

Figure 1: Validation BLEU over Time for

English and Finnish

Figure 2: Validation chrF over Time for

English and Finnish



There does not seem to be a significant difference in BLEU or chrF scores between using the Spanish-English pre-trained model and the Spanish-Finnish pre-trained model as the parent. In fact, the Spanish-English parent model performs slightly better in terms of chrF score.

Figures 3 and 4 show the trajectory of validation BLEU and chrF scores during training (with 1,250 steps equivalent to one epoch). The validation metrics indicate that the Spanish-English pre-trained model might be performing better than the Finnish model, with both models plateauing around 15 epochs (18,750 steps).

5 Limitations

We observe that the BLEU scores for the models using transfer learning via a pre-trained parent model are higher than their corresponding baseline models in both translation directions. Due to slight architectural differences between the baseline and transfer learning models, our comparison is preliminary. The baseline model and tokenizer developed by Duan et al. (2020) are not publicly available. Nevertheless, prior studies support the benefits of transfer learning (Zoph et al., 2016; Wang et al., 2021). Additionally, the baseline paper claims that the lower BLEU score in the Spanish to Mapudungun direction is due to the challenges of generating text in a polysynthetic, agglutinative language with limited training data (Duan et al., 2020). There might be a transfer of knowledge when translating into agglutinative languages, but we can not be sure of this claim until the architectural differences and tokenization methods are consistent as in a true experiment.

6 Discussion

6.1 Tokenization and Vocabulary Analysis

The method of tokenization and the corresponding vocabulary significantly impact translation quality. The text cleaning steps, such as removing diacritics, also play a crucial role. For instance, in part 1, removing diacritics from Spanish and Mapudungun texts and replacing them with normalized ASCII characters may have simplified translation leading to slightly higher scores. This is particularly relevant given the inconsistent use of diacritics in the original dataset (e.g., not using the acute "á" in "más").

The importance of the tokenizer vocabulary is evident when comparing the pre-trained tokenizer to a newly trained tokenizer on Mapudungun text. For example, consider the Mapudungun sentence "Tünkülepürakatun" ("Ahora estoy más calmada" in Spanish, "Now I am more calm" in English). The pre-trained Spanish-Finnish tokenizer produces the tokens ["_T", "ü", "nk", "ü", "le", "p", "ü", "ra", "katu", "n"], which include many single-character tokens. In contrast, the new tokenizer trained on monolingual Mapudungun data generates ["_T", "ün", "küle", "püra", "katun"], capturing more meaningful morphemes. According to other contexts in the dataset as well as an online Spanish-Mapudungun dictionary[4] and loanword database,[5] "tünküle" refers to "calmada" or "tranquila" ("calm"), "püra" comes from the infinitive "püran" approximating "subir" ("to go up") which may refer to "más" ("more"), and "atun" is a common verb suffix.

[4]https://www.conadi.gob.cl/storage/docs/Diccionario_mapudungun.pdf
[5]https://wold.clld.org/vocabulary/41

6.2 Example Translation

The pretrained Spanish to Finnish model's performance was unexpectedly lower compared to the Spanish to English model. For instance, consider the following test dataset example translation:

Spanish Source Text: `Toda la gente lo ve.` *English: All the people see it.*

Mapudungun Translation: `Kom che pey.`

Predicted Target (English Parent): `Kom che pe ngeafuy.`

Predicted Target (Finnish Parent): `Che pe pape feyu.`

Using online Mapudungun resources [6] and patterns from the parallel corpus, we infer that "Kom" translates to "Toda" (meaning "all") and "che" to "la gente" (meaning "the people"). The word "pey" is a third-person singular intransitive form of "pen" (meaning "to see"). The Spanish-English model uses the first-person singular intransitive form "pen" with the suffix "geafuy," which might relate to an impersonal pronoun or passive voice, as in "pigeafuy" ("it is said"). The Spanish-Finnish model may have produced the "pepapeyu" which we infer relates to "coming to see [something or someone]" ("venir" is "to come" and "ver" is "to see").

In some cases, the source Spanish text and target Mapudungun translation are similar due to the presence of Spanish words in Mapudungun utterances. For example:

Spanish Source Text: `y así nunca más` *English: and so never again*

Mapudungun Translation: `y femuechi nunca ma`

Predicted Target (English Parent): `fey tu rpu müten`

Predicted Target (Finnish Parent): `y asi nunca no ma`

In this instance, the child model from the Finnish parent produced a qualitatively similar translation to the reference compared to that from the English parent model.

6.3 Ethical Note

As the developer is not a native speaker of either Spanish or Mapudungun, there is a possibility of inaccuracies in translation that may not be evident through quantitative metrics like BLEU scores alone. These metrics do not always capture semantic accuracy, leading to potential mistranslations or cultural misrepresentations. Qualitative analyses we aimed to conduct were limited to patterns we found in the dataset and online resources about Mapudungun grammar. Not having a Mapuche community member or Mapundungun speaker on the team was a severe limitation since the aim of this translation project is to aid Mapundungun survival.

[6] https://en.wikipedia.org/wiki/Mapuche_language

7 Conclusion

Transfer learning on a parent model seems to improve on non-transfer learning for Spanish-Mapudungun translation, even with very limited epochs. There does not seem to be much of a difference in BLEU or chrF scores between English and Finnish as target languages for the parent model. Primary limitations were the difference between the model architecture of the baseline training and transfer learning models in part 1, and the difference in tokenization and vocabulary of the two trials in the Finnish vs. English comparison. Perhaps it is the similarity in vocabulary and tokenization that has a bigger influence than morphological similarity in grammar. Future work would investigate the effect of the tokenization method, extend training epochs, and compare transfer learning results using parent models trained on a variety of languages. Such parent languages could include those with different scripts and morphological structures, such as Arabic, which is also agglutinative but uses a different script, or Quechua, a similarly agglutinative and polysynthetic language that even shares some loan words, but is another low-resource language. Most importantly, future work should include active collaborations with Mapuche communities and Mapudungun speakers who can analyze the effectiveness of the translations. By engaging with the community, we can ensure that the tool aligns with their linguistic and cultural values and supports their goals for language preservation and revitalization.

Acknowledgments

Thanks to ChatGPT for debugging help, and thanks to XX and YY for checking in and providing support throughout the process.

References

Mingjun Duan, Carlos Fasola, Sai Krishna Rallabandi, Rodolfo Vega, Antonios Anastasopoulos, Lori Levin, and Alan W Black. 2020. A resource for computational experiments on Mapudungun. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2872–2877, Marseille, France. European Language Resources Association.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lorraine Levin, Rodolfo M Vega, Jaime G. Carbonell, Ralf D Brown, Alon Lavie, Eliseo Canulef, and Carolina Huenchullan. 2018. Data Collection and Language Technologies for Mapudungun.

Elizabeth Jacqueline Quintrileo Llancao. 2019. Technology's role in mapudungun language teaching and revitalization. *AlterNative: An International Journal of Indigenous Peoples*, 15(3):205–216.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ineke Smeets. 1989. *A Mapuche Grammar*. Ph.D. thesis, Rijksuniversiteit te Leiden, Leiden.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Dr. Alfréd Tóth. 2007. Are all agglutinative languages related to one another?

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4636–4643. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.