Neural Machine Translation through Active Learning on low-resource languages: The case of Spanish to Mapudungun

María Begoña Pendas
National Center for Artificial Intelligence, Santiago, Chile

Andrés Carvallo
National Center for Artificial Intelligence, Santiago, Chile

Carlos Aspillaga
National Center for Artificial Intelligence, Santiago, Chile

Abstract

Active learning is an algorithmic approach that strategically selects a subset of examples for labeling, with the goal of reducing workload and required resources. Previous research has applied active learning to Neural Machine Translation (NMT) for high-resource or well-represented languages, achieving significant reductions in manual labor. In this study, we explore the application of active learning for NMT in the context of Mapudungun, a low-resource language spoken by the Mapuche community in South America. Mapudungun was chosen due to the limited number of fluent speakers and the pressing need to provide access to content predominantly available in widely represented languages. We assess both model-dependent and model-agnostic active learning strategies for NMT between Spanish and Mapudungun in both directions, demonstrating that we can achieve over 40% reduction in manual translation workload in both cases.

1 Introduction

Over the course of history, South America has been home to numerous indigenous cultures and languages (Campbell et al., 2012), reflecting the region's rich linguistic diversity and heritage. Unfortunately, the dominance of the Spanish language in this region has threatened many indigenous languages, often leading to their decline or even extinction. This has resulted in an immeasurable cultural and historical loss for humanity, as language diversity vanishes (Ostler, 1999). Among the last remaining native languages is *Mapudungun*, spoken in Chile and Argentina by nearly 1.8 million people (Mapuches), but only 10% of them handle the language correctly and barely another 10% understand it. In the same spirit, the Conadi Indigenous Languages Program[1] predicts that this language will become extinct in a few generations, mainly due to the lack of individuals that can speak this language. Despite this, there are still groups within Chile that only speak *Mapudungun*, leaving them sometimes excluded from the rest of society. Furthermore, the social tension over the past few years has raised native indigenous people to the forefront of discussion, attracting high interest in the community to find ways to include them in society as equals. Unfortunately, the availability of human translators fluent in those languages is minimal, and no automated translators exist today supporting those languages. In this work, we present an active learning setting to improve the efficiency and efficacy of machine translation for low-resource languages, in this case, *Mapudungun*. In other words, we aim to reduce the effort made by human translators given that the quantity of people fluent in *Mapudungun* is scarce. Given this, the task of translating and reviewing large amounts of text is unattainable. One of the main tasks of active learning is choosing the appropriate data points (texts) to be translated by human translators to train a neural machine translation (NMT) model with as few examples as possible. To evaluate our approach, we utilized an open-source corpus from the AVENUE project (Levin et al., 2000) and supplemented it by scraping the web for Spanish-Mapudungun sentence pairs. We assembled a dataset of approximately 30,000 pairs, creating a comprehensive corpus for our research. We simulate an offline active learning setting to measure the amount of work that can be reduced by using different active learning strategies. The main contributions of this paper are: (1) Proposing active learning training strategies to reduce low-resource language speaker translators workload by more than 40%, (2) Finetuning a *Mapudungun* NMT model capable of obtaining competitive results and (3) Sharing our code for research reproducibility[2].

[1] https://www.conadi.gob.cl/noticias/conadi-lanzo-aplicaciones-y-realizara-cursos-online-de-mapuzungun-para-que-miles-de-indigenas-aprend

[2] https://github.com/OpenCENIA/al4mt

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 6–11

2 Related work

Active learning

Active learning is an effective machine learning training approach where the algorithm actively selects informative data to learn from, resulting in improved performance with fewer labeled instances (Settles, 2009). While initially applied to text classification, information retrieval, classification, and regression tasks (Tong and Koller, 2001; Zhang and Chen, 2002; Carvallo et al., 2020; Carvallo and Parra, 2019; Houlsby et al., 2011), active learning has recently been extended to tasks such as Named Entity Recognition, Text Summarization, and Machine Translation (Shen et al., 2017; Zhang and Fung, 2012; Zhao et al., 2020; Zhang et al., 2018). This study investigates unexplored potential of active learning in machine translation for untranslated examples in Mapudungun, a low-resource language.

Machine translation for low-resource languages

Efforts to overcome resource scarcity in low-resource language translation have proposed pretraining strategies for data generation and performance improvement. Methods include cross-lingual language model pretraining on high-resource languages data, then finetuning on low-resource languages (Zheng et al., 2021), multilingual sequence-to-sequence pretraining (Song et al., 2019; Xue et al., 2020; Liu et al., 2020), dictionary and monolingual data augmentation (Reid et al., 2021), and back-translation data augmentation (Sugiyama and Yoshinaga, 2019). However, these strategies lack human-in-the-loop components and don't guarantee human approval of the model's iterative translations under active learning.

Data selection in NMT

The data selection problem in NMT has received attention from several authors. Some propose weighted sampling methods to improve performance and accelerate training (Van Der Wees et al., 2017; Wang et al., 2018a), while others focus on filtering noisy data (Wang et al., 2018b; Pham et al., 2018) or selecting domain-specific data for back-translation (Fadaee and Monz, 2018; Poncelas et al., 2023; Dou et al., 2020). Furthermore, Wang et al proposed a method to select relevant sentences from other languages to enhance low-resource NMT performance (Wang and Neubig, 2019). As in using data augmentation the task of selecting data for training a NMT model do not include a user in the feedback loop.

3 Methodology

In this section we describe in detail the active learning framework proposed for NMT on low-resource languages and the type of active learning strategies depending if there is or not a machine learning model involved in the selection of examples for being labeled. In Figure 1, we show the active learning setting used in this work. In the first step, we initialize an NMT model, then given a monolingual corpus in Spanish and an active learning strategy, it chooses examples for being translated by an oracle to *Mapudungun*. After obtaining the translated sentences, we fine-tune the NMT model, update its parameters, and then use this updated version to select new sentences for labeling. We use four active learning strategies to select sentences for an oracle's translation: entropy sampling, margin sampling, confidence sampling, and decay logarithm frequency. The strategies chosen are pertinent to both Spanish to Mapudungun and Mapudungun to Spanish translations in low-resource scenarios. They address key issues such as uncertainty, data diversity, and model reliance, thus optimizing translation models and aiding language preservation. The strategy's reliance on the model varies; model-agnostic strategies don't need it for selecting sentences, while model-related ones use its certainty level. The number of active learning iterations and oracle translation requests is user-determined at the start of training.

3.1 Model-related strategies

These strategies use the model to choose the examples for being labeled and rely on the model's confidence level in untranslated examples.

Entropy sampling

In this strategy we consider entropy as a measure of uncertainty, where the higher entropy indicates higher uncertainty and more chaos. Therefore this strategy consists in sampling examples with higher average entropy given by equation 1.

$$\frac{1}{m}\sum_{i=1}^{m} entropy(P_\theta(.|x, \hat{y}_{<i}) \quad (1)$$

Minimum margin sampling

This strategy calculates the average probability gap between the model's most confident word ($y^*_{i,1}$)

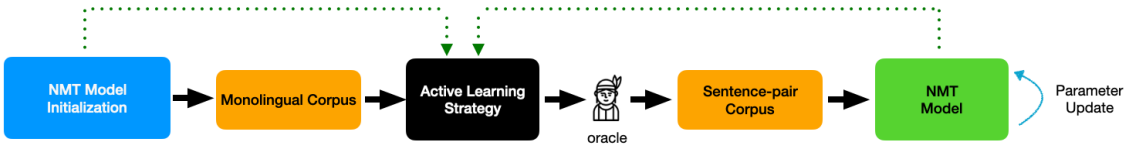


Figure 1: Illustration of the active learning approach.

and the second most confident word ($y^*_{i,2}$). If the margin is small, the model cannot identify the best translation from an inferior one, so we sample sentences with a lower margin as shown in the equation 2.

$$\frac{1}{m}\sum_{i=1}^{m}[P_\theta(y^*_{i,1}|x,\hat{y}_{<i}) - P_\theta(y^*_{i,2}|x,\hat{y}_{<i})] \quad (2)$$

Least Confidence sampling

This strategy estimates the model uncertainty by averaging the predicted probability of each word the translator generates. We sample those sentences with a lower level of confidence to force the model to learn harder sentences, as shown in equation 3.

$$\frac{1}{m}\sum_{i=1}^{m}[1 - P_\theta(\hat{y}_i|x,y_{<i})] \quad (3)$$

3.2 Model-agnostic strategy

In this case, we use the decay logarithm frequency strategy (Zhao et al., 2020) that does not require a NMT model to choose examples for being labeled by an oracle. The intuition behind this strategy is to choose sentences different from the ones that have already been translated in terms of linguistic features.

Decay logarithm frequency

We define two sets of sentences: U that are untranslated and L translated sentences on the current active learning iteration. In the first step, we define the logarithm frequency of a word w in U, namely $F(w|U)$ shown in equations 4 and 5.

$$G(w|U) = log(C(w|U) + 1) \quad (4)$$

$$F(w|U) = \frac{G(w|U)}{\sum_{w' \in U} G(w'|U)} \quad (5)$$

Where $C(w|.)$ measures the frequency of a word w in a given sentence set that can be U or L. Then we add a decay factor that favors the diversity of words and includes two hiper-parameters (λ_1 and λ_2) that allow giving more or less importance to words from the labeled (L) or the unlabeled sets (U). Also, we normalize by dividing the obtained score over the sentence length (K).

$$fy(s) = \frac{\sum_{i=1}^{K} F(s_i|U) \times e^{-\lambda_1 C(s_i|L)}}{K} \quad (6)$$

Equation 6 if used as threshold to obtain $\hat{U}(s)$ that is the set of all sentences that have a higher lf score than s. In this way, we tend to discard repetitive sentences and filter out insignificant function words. The obtention of the final delfy score is shown in equations 7 and 8.

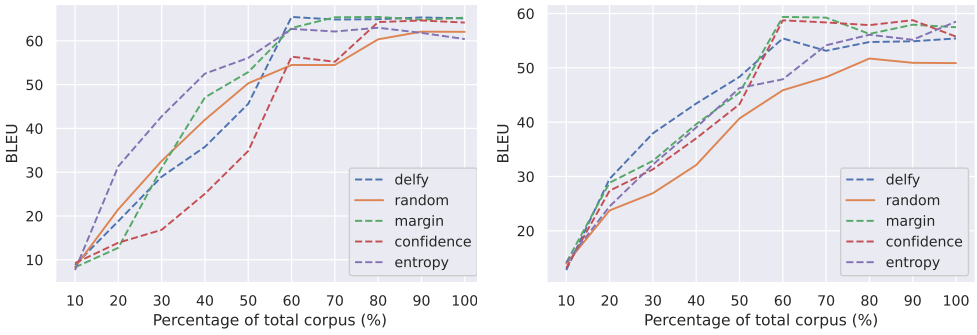
$$delfy(s) = \frac{\sum_{i=1}^{K} F(s_i|U) \times Decay(s_i)}{K} \quad (7)$$

$$Decay(s_i) = e^{-\lambda_1 C(s_i|L)} \times e^{-\lambda_2 C(s_i|\hat{U}(s))} \quad (8)$$

4 Experiments

4.1 Dataset, preprocessing and NMT model

The dataset consists of 29,829 Spanish to *Mapudungun* sentence pairs considering only sentences length higher than five words, with 50,840 unique words in Spanish, 67,757 unique words in *Mapudungun*, and a vocabulary size of 118,597. We do not remove stopwords, lemmatization, or low-case texts, since we aim to capture both languages' peculiarities, including punctuation and idioms. We used a MarianMT (Junczys-Dowmunt et al., 2018) translation model based on a transformer architecture consisting of 12 encoder layers, 16 encoder attention heads, 12 decoder layers, and 16 attention heads. For training on active learning, we use a learning rate of 0.0002 and a weight decay of 0.01. We train the necessary epochs in each active training round until the validation perplexity remains the same. λ_1 and λ_2 in the delfy are set to 1.0 each. For training on active learning, we



(a) ESP-MAPU (b) MAPU-ESP

Figure 2: NTM models training on Active Learning. X-axis indicate the percentage of the corpus used to train each model choosing examples based on each active learning strategy. Y-axis indicates the BLEU score.

finetune a MarianMT translator from Spanish to Deutsch. Despite the apparent oddity of linking an Indo-European language, Deutsch with Mapudungun, our approach harnesses shared agglutinative traits to enhance translation.

4.2 Active Learning for NMT

Concerning the active learning setting, we run ten iterations using the 10% of the train set. For evaluating active learning strategies, we used the SacreBLEU[3] library and evaluated the model's outputs with BLEU (Papineni et al., 2002). As we run an offline experiment, we assume the oracle is continuously right, extracting the correct translation each time and adding those examples to the train set. In our offline experiment, we used existing labeled training data to eliminate the need for human annotators. Our goal was to assess which strategy efficiently utilizes a smaller data proportion, reducing manual translation effort while preserving model performance. This approach enables optimization of active learning strategies without added annotation costs.

4.3 Results

The results of this study suggest that for Spanish to *Mapudungun* translation, the most effective active learning strategy is Delfy, which achieved a BLEU score of 65.45 when trained on 60% of the corpus. Margin and entropy sampling were also effective strategies, achieving BLEU scores of 62.92 and 62.72, respectively. For *Mapudungun* to Spanish translation, margin sampling was the most effective active learning strategy, achieving a BLEU score of 59.378. Both settings showed benefits of training on active learning, with a reduction in the workload of approximately 40%. However, there is space for improvement in further reducing workload, as other studies on high-resource or well-represented languages have reduced over 80% (Zhao et al., 2020) of manual translation work. This work demonstrated significant progress in translating a low-resource language such as *Mapudungun*, with both active learning strategies outperforming the baseline strategy of random sampling.

5 Conclusion

In conclusion, this study revealed that Delfy was the most effective active learning strategy for Spanish to *Mapudungun* translation, while margin sampling outperformed in *Mapudungun* to Spanish. In both cases, training with active learning strategies reduced workload by over 40%. Our comparative analysis, driven by the diverse approaches of the chosen strategies, identifies the most efficient methods for low-resource translation tasks. This research is crucial for languages particularly *Mapudungun*, as it fosters information access and reduces language barriers for indigenous communities. Future work will focus on designing active learning strategies specifically for low-resource languages.

Acknowledgements

National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

[3]https://github.com/mjpost/sacrebleu

References

Lyle Campbell, Verónica Grondona, and HH Hock. 2012. *The indigenous languages of South America*. de Gruyter.

Andres Carvallo and Denis Parra. 2019. Comparing word embeddings for document screening based on active learning. In *BIRNDL@ SIGIR*, pages 100–107.

Andres Carvallo, Denis Parra, Hans Lobel, and Alvaro Soto. 2020. Automatic document screening of medical literature using word and text embeddings in an active learning setting. *Scientometrics*, 125:3047–3084.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. *arXiv preprint arXiv:2004.03672*.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. *arXiv preprint arXiv:1808.09006*.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Lorraine Levin, Rodolfo M Vega, Jaime G Carbonell, Ralf D Brown, Alon Lavie, Eliseo Cañulef, and Carolina Huenchullan. 2000. Data collection and language technologies for mapudungun.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Rosemarie Ostler. 1999. Disappearing languages. *The Futurist*, 33(7):16.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Minh Quang Pham, Josep M Crego, Jean Senellart, and François Yvon. 2018. Fixing translation divergences in parallel corpora for neural mt. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2967–2973.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2023. Adaptation of machine translation models with back-translated data using transductive data selection methods. In *Computational Linguistics and Intelligent Text Processing: 20th International Conference, CICLing 2019, La Rochelle, France, April 7–13, 2019, Revised Selected Papers, Part I*, pages 567–579. Springer.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. Afromt: Pretraining strategies and reproducible benchmarks for translation of 8 african languages. *arXiv preprint arXiv:2109.04715*.

Burr Settles. 2009. Active learning literature survey.

Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44.

Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.

Marlies Van Der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. *arXiv preprint arXiv:1708.00712*.

Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2018a. Dynamic sentence sampling for efficient training of neural machine translation. *arXiv preprint arXiv:1805.00178*.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018b. Denoising neural machine translation training with trusted data and online data selection. *arXiv preprint arXiv:1809.00068*.

Xinyi Wang and Graham Neubig. 2019. Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. *arXiv preprint arXiv:1905.08212*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Cha Zhang and Tsuhan Chen. 2002. An active learning framework for content-based information retrieval. *IEEE transactions on multimedia*, 4(2):260–268.

Justin Jian Zhang and Pascale Fung. 2012. Active learning with semi-automatic annotation for extractive speech summarization. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(4):1–25.

Pei Zhang, Xueying Xu, and Deyi Xiong. 2018. Active learning for neural machine translation. In *2018 International Conference on Asian Language Processing (IALP)*, pages 153–158. IEEE.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806.

Francis Zheng, Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Low-resource machine translation using cross-lingual language model pretraining. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 234–240.