Exploration of the CycleGN Framework for Low-Resource Languages

Sören Dréano
ML-Labs
Dublin City University
soren.dreano2@mail.dcu.ie

Derek Molloy
School of Electronic Engineering
Dublin City University
derek.molloy@dcu.ie

Noel Murphy
School of Electronic Engineering
Dublin City University
noel.murphy@dcu.ie

Abstract

CycleGN is a Neural Machine Translation framework relying on the Transformer architecture. Its approach is similar to a Discriminatorless CycleGAN, specifically tailored for non-parallel text datasets.

The foundational concept of our research posits that in an ideal scenario, retro-translations of generated translations should revert to the original source sentences. Consequently, a pair of models can be trained using a Cycle Consistency Loss only, with one model translating in one direction and the second model in the opposite direction.

One of the main advantages of such an approach is that it makes it possible to learn with non-parallel datasets, which are by definition rare and short for low-resource languages. In order to verify this hypothesis and as a contribution to the WMT24 challenge, CycleGN models were trained for both the "Translation into Low-Resource Languages of Spain" and "Low-Resource Indic Language Translation" tasks. These submissions fall under the "constrained" category, as no pre-trained translation model was used, and the models were trained using the provided datasets.

Given that the CycleGN architecture demonstrated its capacity to learn from non-parallel datasets, the authors anticipated that it would similarly be effective in learning from low-resource languages. However, preliminary results indicate that, for most low-resource language pairs, the models did not exhibit significant learning ability. This study explores this lack of learning.

1 Introduction

The introduction of the Transformer architecture (Vaswani et al., 2017) marked a significant advancement in Machine Translation, rapidly gaining widespread adoption. Its parallelized structure enhanced computational efficiency, allowing for the integration of a larger number of parameters.

Neural Machine Translation (NMT) relies on extensive text corpora, structured as aligned pairs, where sentences of equivalent meaning are available in at least two different languages. This alignment is crucial for initiating model training to establish linguistic connections. Ongoing efforts, such as OPUS (Tiedemann and Thottingal, 2020) and Tatoeba (Tiedemann, 2012), focus on providing public access to these datasets. However, parallel datasets represent only a small fraction of the total data available in monolingual datasets.

While large parallel corpora exist for many language pairs, the ability to utilize monolingual datasets alone would greatly increase the available training data. This approach is particularly advantageous for low-resource languages, with limited parallel text corpora.

Back-translation (Sennrich et al., 2016) is a technique that enhances training data by using a pre-trained machine translation (MT) model to translate sentences from a monolingual dataset, creating synthetic parallel pairs. This method allows for the generation of additional training examples in situations where parallel corpora are scarce.

This research builds on the concept that translating a sentence from a source language to a target language, and then back-translating it to the source language, provides a means to evaluate the effectiveness of the translation models. By comparing the original sentence with the machine-generated back-translation, the discrepancy is then quantified using a Cycle Consistency Loss, which serves as a metric for model performance and guides the backpropagation of gradients within the neural networks. This approach is analogous to techniques used in Image-to-Image Translation, such as the CycleGAN framework proposed by Zhu et al. (2017).

2 Previous work

The TextCycleGAN model (Lorandi et al., 2023), although not based on the Transformer architecture

Proceedings of the Ninth Conference on Machine Translation, pages 756–761
November 15-16, 2024 ©2024 Association for Computational Linguistics

or focused on Machine Translation (MT), introduced a novel approach for text style transfer. This method applied a CycleGAN to the Yelp dataset, enabling the model to learn mappings between positive and negative textual styles without the need for paired examples.

Shen et al. (2017) demonstrated the potential of training two encoder-decoder networks in an unsupervised manner, allowing for the sharing of a latent space and facilitating style transfer. Similarly, Lample et al. (2018) extended this technique to the MT domain, proving that effective translation can be achieved without relying on parallel datasets.

3 Definitions

Machine Translation models are most commonly trained using "parallel" datasets, which are structured collections of text pairs. Each pair comprises a segment of text in a source language and its translation in the target language. A non-parallel dataset on the other hand does not consist in pairs of text segments, consequently the source and target sentences do not share any explicit correspondence.

In the context of this study, the datasets are "permuted". A permuted dataset is defined as a parallel dataset wherein the sentences of one language have been systematically rearranged. Consequently, this results in a non-parallel corpus where it is guaranteed that each sentence has a corresponding translation located at an unspecified index within the dataset.

4 Datasets

The PILAR dataset (Galiano-Jiménez et al., 2024) has been used exclusively for the low-resource languages of Spain. Using a parallel curated dataset as a starting point ensures that the dataset is non-parallel by permuting the sentences. For each Iberian language, both a literary and a crawled versions were available in the PILAR datasets and have been merged for training. The development sets of the PILAR dataset are translations of the development sets of the FLORES dataset (NLLB Team et al., 2022), which is an evaluation benchmark for multilingual machine translation.

The Low-Resource Indic Language Translation task was also part of the WMT23 (Pal et al., 2023). The datasets were kept the same between the two editions.

Table 1 references the number of sentences used for each language-pair.

Language Pair	Number of lines	Number of epochs
Spanish-Aragonese	84,703	10
Spanish-Asturian	38,869	10
English-Assamese	2,624,715	1
English-Khasi	182,737	3
English-Manipuri	2,144,897	1
English-Mizo	1,909,823	1

Table 1: Number of sentences for each language pair and number of epochs during training

5 Training

For clarity and consistency, the mathematical notations from the original CycleGAN framework will be adopted in this study. The objective is to develop two Neural Machine Translation (NMT) models for two languages, $\mathcal{X}$ and $\mathcal{Y}$, using their respective datasets. Specifically, we aim to construct models $\mathcal{G} : \mathcal{X} \mapsto \mathcal{Y}$ and $\mathcal{F} : \mathcal{Y} \mapsto \mathcal{X}$ such that, in the ideal scenario of perfect translation, the relationships $\mathcal{G}(\mathcal{F}(y)) = y$ and $\mathcal{F}(\mathcal{G}(x)) = x$, with $x \in \mathcal{X}$ and for $y \in \mathcal{Y}$.

To achieve this, the Cross-Entropy Loss (CEL) (Zhang and Sabuncu, 2018) is utilised as the Cycle Consistency Loss (CCL), which measures the distance between the original sentence and its doubly translated counterpart, thereby guiding the computation of gradients.

Furthermore, similar to the original CycleGAN implementation, our study also incorporates an Identity Loss (IL) to enhance training stability. This loss, also based on CEL, ensures that when the model $\mathcal{G}$, which maps $\mathcal{X} \mapsto \mathcal{Y}$, receives an input $y \in \mathcal{Y}$, the output remains unchanged, i.e., $\mathcal{G}(y) = y$. The same loss function is applied to $\mathcal{F}$, ensuring that $\mathcal{F}(x)$ remains equal to x, as illustrated in Figure 1.

Further details of the training process, including the specific methodologies, vocabulary organization and pretraining, are comprehensively discussed in the CycleGN submission for the WMT24 main translation task. Readers interested in the full technical details are encouraged to refer to that publication for a more complete understanding of the training framework.

5.1 Model architecture

The architecture used for both models, $\mathcal{G}$ and $\mathcal{F}$, is the Marian framework (Junczys-Dowmunt et al., 2018) implemented by Huggingface's Transformers library (Wolf et al., 2020), which is licensed under the Apache Licence. While most parameters

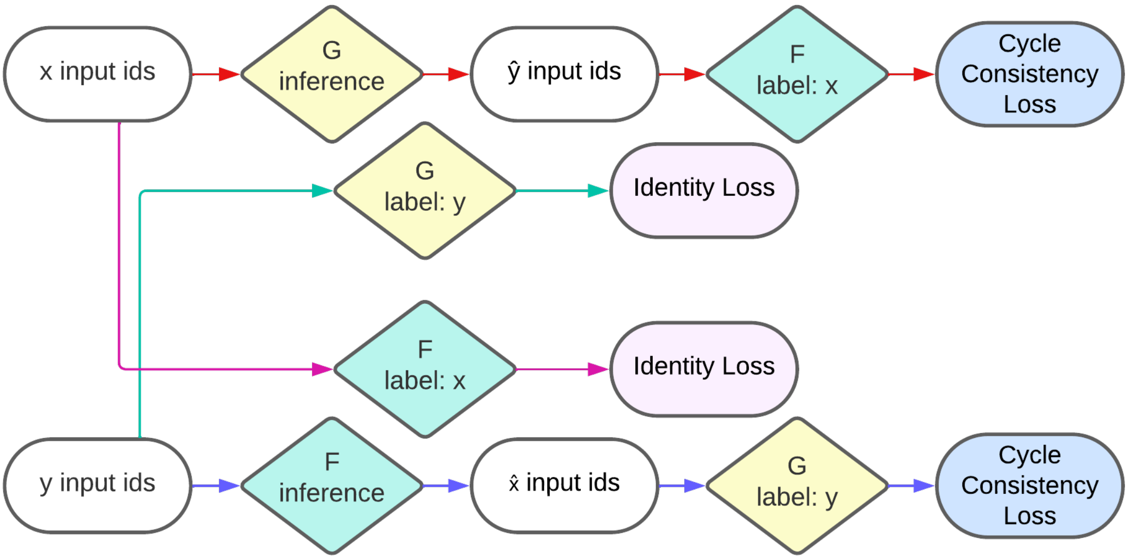

Figure 1: CycleGN training process

follow the default configuration, Table 2 references the changes that were made in order to reduce the computational cost of the architecture.

Parameter	Huggingface	Current work
Vocabulary size	58,101	32,000
Encoder layers	12	6
Decoder layers	12	6
Encoder attention heads	16	8
Decoder attention heads	16	8
Encoder feed-forward	4096	2048
Decoder feed-forward	4096	2048
Position embeddings	1024	128
Activation function	GELU	ReLU

Table 2: Non-default parameters in the configuration of Marian Transformer models

6 Results

Even if tracking the CCL is an inexpensive manner to estimate the progress of the training of the CycleGN architecture, a low loss value can also hide an absence of translation. Indeed, as there is no Discriminator to ensure that $\hat{x}$ belongs to $\mathcal{X}$ and $\hat{y}$ belongs to $\mathcal{Y}$, $\mathcal{G}$ and $\mathcal{F}$ will converge towards $x = \hat{y} = \hat{\hat{x}}$ and $y = \hat{x} = \hat{\hat{y}}$, as this approach achieves an optimal outcome on the CCL function, registering a value of zero. This is why an evaluation metric such as COMET is crucial to assess the progression of the CycleGN framework. To measure the performances of CycleGN, every 1,000th batch the CCL was averaged.

6.1 Indic Languages

Tracking the evolution of the CDC clearly shows the absence of learning in the four language pairs examined. The evolution of the CCL is particularly chaotic, which is partly due to an imbalance of class. Table 3 displays the average number of tokens in the Indic datasets depending on the language. In 3 of the 4 cases, the difference is large, i.e. sentences where the difference in the number of tokens is more than 10%.

Language pair	Length of source	Length of target
English-Assamese	33.10	22.81
Encoder layers	24.09	75.27
English-Manipuri	24.09	26.07
English-Mizo	32.05	17.55

Table 3: Average number of tokens in sentences

Figures 2, 3, 4 and 5 display the respective evolution of the Cycle Consistency Loss during the training of the language-pairs English-Assamese, English-Khasi, English-Manipuri and English-Mizo.

Contrary to what the authors had hoped for on the basis of previous results obtained for the main task of the WMT24 challenge, no model followed the expected learning curve, i.e. G and F models with a close and slowly decreasing Cycle Consistency Loss.

To reduce this imbalance of class, it may be necessary to manually adjust the size of the sentences.

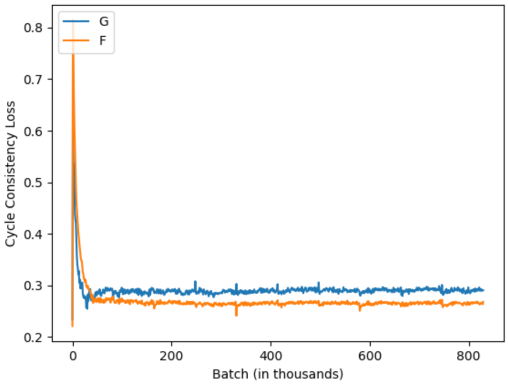

Figure 2: Evolution of the Cycle Consistency Loss during the training of the English-Assamese model

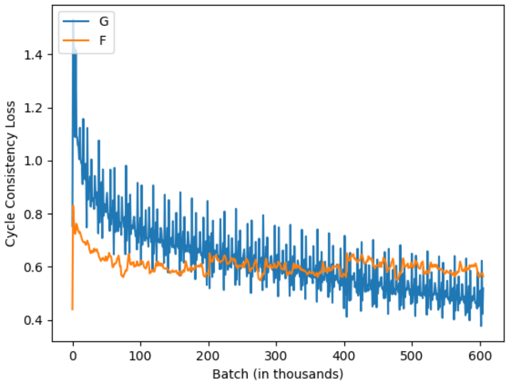

Figure 3: Evolution of the Cycle Consistency Loss during the training of the English-Khasi model

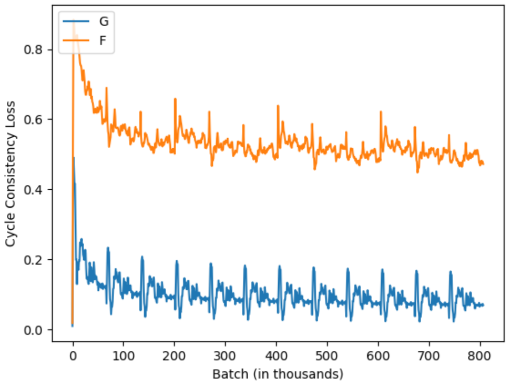

Figure 4: Evolution of the Cycle Consistency Loss during the training of the English-Manipuri model

This can be done by choosing another tokenization method, selectively choosing phrases to keep only those of a similar size, or by trimming sentences to lengthen or shorten them as required.

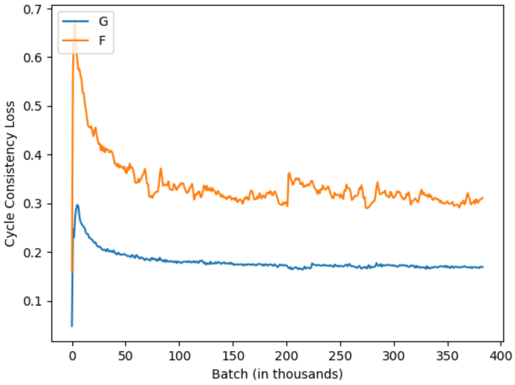

Figure 5: Evolution of the Cycle Consistency Loss during the training of the English-Mizo model

6.2 Iberian Languages

As with Indic Languages, CycleGN was unable to learn from the datasets provided. However, it was not due to an imbalance of classes in this case, but rather because the classes were too close together, as the Iberian languages are very close to the source language, Spanish.

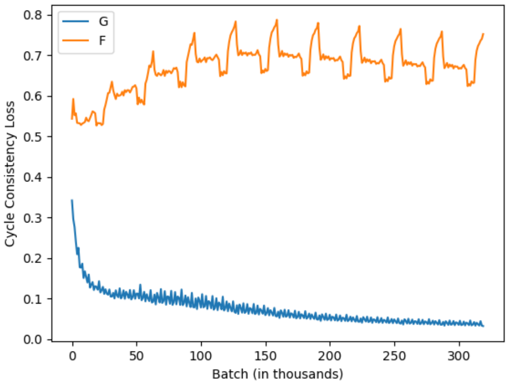

Figure 6: Evolution of the Cycle Consistency Loss during the training of the Spanish-Aragonese model

Rather than translating directly from Spanish into Aranese or Asturian, it is possible that translation can be achieved by using a different intermediate language such as English. Thus, two CycleGN models would have to be trained, the first to translate from Aranese or Asturian into English, and the second from English into Spanish. This would double the training time for an already expensive framework.

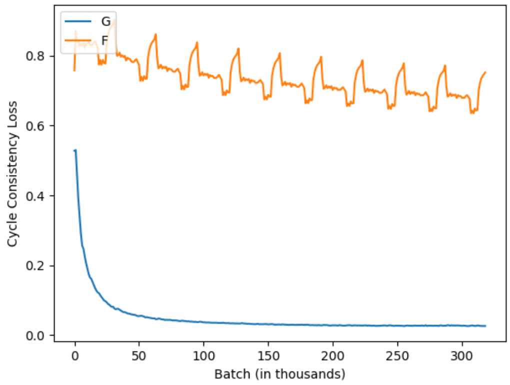

Figure 7: Evolution of the Cycle Consistency Loss during the training of the Spanish-Asturian model

7 Conclusion

In conclusion, while the training process demonstrated significant progress and effective translation capabilities in the main study, the results presented in this paper reveal several challenges that prevented similar success. The issues identified, particularly in relation to both class imbalance and class proximity, indicate that further refinement and investigation are necessary. Future research should focus on addressing these challenges, with the aim of optimizing the training process and overcoming the outlined issues. Resolving these problems is crucial for realizing the full potential of the framework within the context discussed in this paper.

8 Future Work

Further investigations will benefit from the incorporation of a more extensive dataset and an exploration of larger model architectures. Future work also include methods discussed in Section 6 to allow translation training.

8.1 Larget dataset

The current work has been trained on a small dataset compared to MT standards. Future work should try to see how convergence progresses with more iterations. Further computational optimizations are probably necessary to shorten the training time required.

8.2 Larger models

The current architecture relies on a total of 158,769,152 parameters, which is only about a third of the size of the default in the Huggingface library.

9 Source Code

The source code of CycleGN is available at https://github.com/SorenDreano/CycleGN.

Limitations

The investigation acknowledges certain inherent limitations which may impact the generalizability and applicability of the findings.

Language diversity

Another issue that arises from the computing cost of CycleGN is the lack in language diversity. Indeed, our current work only used the English-German and Chinese-English language pairs. Consequently, it cannot be certain that the approach presented can be applied to other languages and all alphabets. This is why CycleGN is taking part in WMT24, to explore the framework's performance on a wide range of language pairs.

Training limitations

Due to time constraints and the fact that CycleGN is a computationally expensive architecture, it was not possible to train the Spanish-Aranese pair. Similarly, the training of all models was stopped early, before reaching performance stagnation.

Ethics Statement

This study, focusing on the training of NMT models using non-parallel datasets, adheres to the highest ethical standards in research. We recognize the critical importance of ethical considerations in computational linguistics and machine learning, especially as they pertain to data sourcing, model development, and potential impacts on various linguistic communities.

Our research utilizes publicly available, non-parallel linguistic datasets. We ensure that all data is sourced following legal and ethical guidelines, respecting intellectual property rights and privacy concerns.

In our commitment to scientific integrity, we maintain transparency in our research methodologies, model development, and findings. We aim to make our results reproducible and accessible to the scientific community, contributing positively to the field of machine translation.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024. Pilar.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only.

Michela Lorandi, Maram A.Mohamed, and Kevin McGuinness. 2023. Adapting the CycleGAN Architecture for Text Style Transfer. *Irish Machine Vision and Image Processing Conference*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks.