Galician–Portuguese Neural Machine Translation System

Sofía García González
imaxin|software / Salgueiriños de Abaixo, N 11 L6, Santiago de Compostela
sofia.garcia@imaxin.com

Abstract

This paper presents the first Galician–Portuguese (GL–PT) bilingual neural machine translation (NMT) model. Due to the lack of Galician–Portuguese parallel data, this model was trained on synthetic data converting the Spanish part from original Spanish–Portuguese corpora to Galician using the RBMT system `Apertium`.

1 Introduction

In recent years, neural machine translation (NMT) has become the state-of-the-art in this natural language processing (NLP) area. It has shown promising results in various language pairs. However, developing efficient translation models for low-resource languages such as Galician is challenging due to the need for large training parallel corpus (Haddow et al., 2022).

O Proxecto Nós (The Nós Project) has currently developed neural MT models for Spanish–Galician[1] and English–Galician[2] pairs in both directions. These models were trained converting the Portuguese part from original English–Portuguese and Spanish–Portuguese corpora to Galician. (Ortega et al., 2022). However, there is currently no NMT system for Portuguese–Galician pair, except for multilingual models where Galician is included as M2M (Fan et al., 2021) or NLLB (Costa-jussà et al., 2022). Furthermore, despite the closeness of these two languages, both the RBMT system `Apertium` (Forcada et al., 2011) and the `port2gal`[3] transliterator perform poorly in both translation directions, particularly to put it into production as a company.

Therefore, this paper presents a Galician–Portuguese neural translation model tailored to the administrative domain, which **imaxin**|software provides to clients such as the *Xunta de Galicia* (Galician Gonvernment) with `GAIO`[4] or the Galician Parliament.

2 Methodology

2.1 Training Corpora

In accordance with the de Dios-Flores et al. (2022) strategy, the process was divided into two steps. Firstly, we gathered two Spanish–Portuguese parallel macrocorpora: CCMatrix,[5] and OpenSubtitles v2018.[6]; and a legal-domain corpus: the Spanish–Portuguese DGT v8[7] (see Table 1 for corpus sizes). Then, using the RBMT system developed for `GAIO`, we created synthetic corpora translating the Spanish part into Galician, in order to obtain synthetic Portuguese–Galician parallel corpora.

Domain	Dataset	Number of Sentences
General Domain	CCMatrix	25M
	OpenSubtitles	25M
Legal Domain	DGT v2019	3.5M

Table 1: Spanish–Portuguese training corpus sizes

2.2 Architecture

Regarding the training process, we have used the Transformer architecture from OpenNMT-py[8] open-source framework. For this initial model, we have assigned greater weight to the generic CCMatrix and OpenSubtitles corpora, with weights of 50 for both macrocorpora, while the DGT corpus had a weight of 20. The training parameters can be seen in Table 2.

[1] https://huggingface.co/proxectonos/Nos_MT-OpenNMT-es-gl
[2] https://huggingface.co/proxectonos/Nos_MT-OpenNMT-en-gl
[3] https://fegalaz.usc.es/~gamallo/port2gal.htm
[4] *Xunta de Galicia*'s MT system based on `Apertium`, https://tradutorgaio.xunta.gal/TradutorPublico/traducir/index.
[5] https://opus.nlpl.eu/CCMatrix-v1.php. We only used the half size of CCMatrix. Thus, we selected 25M random sentences
[6] https://opus.nlpl.eu/OpenSubtitles-v2018.php
[7] https://opus.nlpl.eu/DGT-v2019.php
[8] https://github.com/OpenNMT/OpenNMT-py

Parameters	Values
Model	Transformer
dropout	0.1
average_decay	0.0005
label_smoothing	0.1
optimization	adam
learning_rate	2
warmup_steps	8000
batch_size	8192

Table 2: Training Parameters

2.3 Evaluation

The corpora used to evaluate the NMT model were: Flores200-dev (Goyal et al., 2022)[9], News Test References for MT Evaluation (NTREX) (Barrault et al., 2019)[10] and a 1k corpus extracted from CC-Matrix. See Table 3 for sizes[11].

Evaluation Dataset	Size
Flores200-dev	1k
NTREX	2k
CCMatrix–test–dataset	1k

Table 3: Portuguese–Galician Evaluation test sizes

On the other hand, we used the Sacrebleu framework[12] as recommended by Post (2018). This framework includes: BLEU (Papineni et al., 2002), chrF (Popović, 2015) and TER (Snover et al., 2006) metrics. Moreover, we also used the current state-of-the-art COMET (Rei et al., 2022) [13].

3 Results

The following tables report the results for each evaluation dataset: Flores200-dev (Table 4), NTRIX (Table 5) and CCMatrix (Table 6). We have used *Apertium* as the baseline to compare our results.

MT Systems	BLEU	chrF	TER	COMET
Apertium	21.3	52	62.8	**0.824**
imaxin|software model	**24.2**	**54.3**	**61.2**	0.769

Table 4: Flores200-dev results in gl–pt systems

[9] https://github.com/facebookresearch/flores/tree/main/flores200

[10] https://github.com/MicrosoftTranslator/NTREX

[11] Because of the lack of legal-domain test datasets in this language pair, we have not been able to make a specific evaluation in this domain.

[12] https://pypi.org/project/sacrebleu/

[13] We have used the wmt22-comet-da model

MT Systems	BLEU	chrF	TER	COMET
Apertium	**23**	**53.4**	**63.3**	**0.810**
imaxin|software model	21.6	51.9	64.6	0.745

Table 5: NTRIX results in gl–pt systems

MT Systems	BLEU	chrF	TER	COMET
Apertium	**41.6**	**69.4**	**51.3**	0.848
imaxin|software model	32.7	69.1	52	**0.888**

Table 6: CCMatrix test results in gl–pt systems

4 Analysis

As shown in the tables, with the exception of the flores200-dev test (Table 4), Apertium continues to outperform our NMT model. The difference in results is particularly remarkable on the test taken from the CCMatrix corpus (Table 6), where Apertium outperforms the neural model by 10 BLEU points. However, both translation systems yield unsatisfactoryresults for two closely related languages. The absence of an authentic Galician-Portuguese corpus poses a challenge for developing good quality NMT models. In fact, one of the major issues with macrocorpora such as CCMatrix is that they mix variants of Portuguese from Portugal and Brazil, resulting in inconsistent language during translation. That is, they are unable to maintain the same variant throughout the translation process. On the other hand, Apertium does not present this issue, as it is a system designed to translate to and from the European variant of Portuguese. Therefore, in the future, a more in-depth analysis is necessary to determine how different varieties of Portuguese affect NMT models development.

5 Conclusions

This demo model provides a starting point for NMT between Galician and Portuguese. In the future, other strategies will be tested, such as deeper cleaning of the web-extracted corpora, distinguishing between Portuguese variants, or creating legal test corpora for this language pair, which currently does not exist and hinders accurate evaluation for this domain. The development of high-quality parallel corpora will be crucial for the future development of NMT models.

6 Demonstration

Our demonstration will be show on an **imaxin**|software webpage where users will be able to translate any text from Galician to

Portuguese to test this model.

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Iria de Dios-Flores, Carmen Magarinos, Adina Ioana Vladu, John E Ortega, José Ramom Pichel Campos, Marcos Garcia, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín Diz, Manuel González González, et al. 2022. The nós project: Opening routes for the galician language in the field of language technologies. In *Proceedings of the workshop towards digital language equality within the 13th language resources and evaluation conference*, pages 52–61.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

John E Ortega, Iria de Dios-Flores, Pablo Gamallo, and José Ramom Pichel. 2022. A neural machin e translation system for galician from transliterated portuguese text. In *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing. CEUR Workshop Proceedings*, volume 3224, pages 92–95.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.