A neural machine translation system for Galician from transliterated Portuguese text

Un sistema de tradución neuronal para el gallego a partir de texto portugués transliterado

John E. Ortega, Iria de-Dios-Flores, José Ramom Pichel and Pablo Gamallo

Centro de Investigación en Tecnoloxías da Información (CITIUS), Universidad de Santiago de Compostela, Spain

Abstract

We present a neural machine translation (NMT) system for translating both Spanish and English to Galician (*ES–GL* and *EN–GL*). Galician is a language closely related to Portuguese, with low to medium resources, spoken in northwestern Spain. Our NMT system is trained on large-scale synthetic $ES \rightarrow PT \rightarrow GL$ and $EN \rightarrow PT \rightarrow GL$ parallel corpora created by the spelling transliteration of Portuguese to Galician from a high-quality Spanish to Portuguese (*ES–PT*) and English to Portuguese (*EN–PT*) translation memories. The NMT system is then made available via a public web interface at https://demos.citius.usc.es/nos_tradutor.

Keywords

Galician Language, Neural Machine Translation, Transliteration

1. Introduction

Several systems have been compared and developed to perform machine translation (MT), ranging from rule-based systems to systems based on neural networks [1] Traditionally, rule-based systems like Apertium [2] are used for languages with a small amount of parallel data. That is because MT systems backed by neural networks, or neural machine translation (NMT) systems, require high amounts of data, typically on the order of millions of sentences or more [3, 4]. An interesting option for low-resource languages is the use of zero-shot translation techniques, that is, translating in multilingual settings between language pairs for which the NMT system has never been trained. However, as Gu et al. [5] state, training zero-shot NMT models easily fails as this task is very sensitive to hyper-parameter setting. The performance of zero-shot strategies is usually lower than that of more conventional pivot-based approaches.

We describe and implement an approach inspired by previous work [6] that uses the proximity of Portuguese and Galician to overcome the lack of resources problem and produces corpora to build an NMT system, similar to low-resource NMT systems found in previous work [7, 8], for translating both Spanish to Galician and English to Galician. Our system first uses high-quality Spanish–Portuguese (ES–PT) and English–Portuguese (EN–PT) parallel corpora to translate the target-sided (Portuguese) sentences (or segments) to Galician using *transliteration*, the conversion of text in one language to another through spelling. Transliteration between Portuguese and Galician works well due to the orthographic nearness of the two languages found in previous work [9]. Second, NMT systems with the transliterated Galician parallel text are created to form a Spanish–Galician (ES–GL) and English–Galician (EN–GL) MT system where both Spanish and English are the source languages and Galician is the target language. Two different neural-based architectures were tested: Long short-term memory (LSTM) and Transformers.

2. Method

Our translation strategy consists of two steps. The first step uses *transliteration* [10] to create parallel Galician segments from the Portuguese segments in the aligned corpus, by making using of the transliteration tool port2gal[1], which contains several hundreds of rules on characters and sequences of characters. Both training and validation sets are transliterated leaving a final parallel Galician corpus. Then, in the second step, the Galician (transliterated) cor-

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

john.ortega@usc.gal (J. E. Ortega); iria.dedios@usc.gal (I. de-Dios-Flores); jramon.pichel@usc.gal (J. R. Pichel); pablo.gamallo@usc.gal (P. Gamallo)
0000-0002-2328-3205 (J. E. Ortega); 0000-0002-5941-1707 (I. de-Dios-Flores); 0000-0001-5172-6803 (J. R. Pichel); 0000-0002-5819-2469 (P. Gamallo)

[1] https://github.com/gamallo/port2gal

system	pair	source	corpus size	bleu	ter	chrF2
lstm	es-gl	Europarl+CLUVI	2.35M	48.9	34.4	69.3
lstm	es-gl	Europarl+CLUVI+OpenSubt(part)	5M	**51.1**	**32.8**	**70.8**
lstm	es-gl	Europarl+CLUVI+OpenSubt	30M	46.0	37.2	66.5
transformer	es-gl	Europarl+CLUVI	2.35M	17.5	67.4	53.0
transformer	es-gl	Europarl+CLUVI+OpenSubt	30M	13.9	66.7	46.4
lstm	en-gl	Europarl+OpenSubt	27.M	26.6	50.3	45.5
transformer	en-gl	Europarl+OpenSubt	27.M	**29.3**	**49.7**	**51.0**

Table 1
Results obtained for the two language pairs (*ES–GL* and *EN–GL*) evaluated on two different systems, LSTM and Transformer, by making use of three quantitative measures: BLEU, TER and ChrF2. The corpus size is quantified in millions of sentences (M).

pus is used to train an NMT system with Spanish or English as the source language and Galician as the target language. For the first transliteration step, we also tested a more complex strategy by combining PT→GL Apertium translator [2], which uses a basic bilingual dictionary to translate word by word, with the transliteration tool for those words that are not in the bilingual dictionary.

The NMT system that we use for ES–GL and EN–GL translations was created using OpenNMT [11], a generic deep learning framework for creating sequence-to-sequence models in machine translation. In particular, we trained a LSTM (long short term memory) seq2seq model as well as a Transformer model for each language pair.

Concerning LSTM, we used the following default neural network training parameters: two hidden layers, 500 hidden LSTM units per layer, input feeding enabled, 13 epochs, batch size of 64. Alternatively, we modified the default learning step parameters to 100,000 training steps and 10,000 validation steps. Traditional tokenization was performed with Linguakit [12]

The Transformer implementation, described in Garg et al. [13], was configured with default training parameters: 6 layers for both encoding and decoding and batch size of 4096 tokens. We also modified the learning step parameters to the same values as the LSTM configuration. In this case, we used sub-word tokenization, performed with SentencePiece [14].

3. Corpora

The main parallel sources we used to train the NMT system come from Opus[2]. In particular we used the *ES–PT* and *EN–PT* partitions of both Europarl[3], with about 2 million sentences per language, and OpenSubtitles[4], containing about 30 million sentences in *ES–PT* and 25 in *EN–PT*. The Portuguese partition was transliterated to Galician so as to build *ES–GL* and *EN–GL* parallel corpora. In addition, we also added the Spanish-Galician partition of CLUVI[5], to the *ES–GL* corpus, containing 144 thousand sentences.

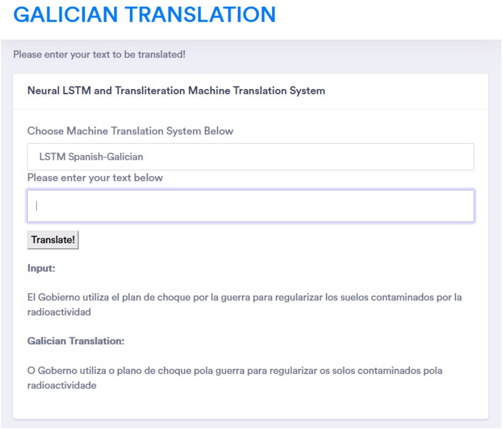
4. Test results

Table 1 show the results of different experiments for *ES–GL* and *EN–GL* combining the system, LSTM or Transformer, with the size of the corpus. We observe that LSTM works very well for close languages (*ES–GL*), but for the pair (*EN–GL*), two distant languages, the results are slightly better with Transformer. In addition, we also observe that the whole OpenSubtitles corpus hurts the performance in *ES–GL*. The best results in *ES–GL* combine Europarl with OpenSubtitles and are comparable to the state-of-the-art [15]. Let us note that the Movie and TV subtitles of OpenSubtitles are a highly valuable resource but the quality of the resulting sentence alignments is often lower than for other parallel corpora [16]. The results in Table 1 allow us to confirm that using transliteration between two closely aligned languages like Portuguese and Galician, favorable outcomes can be achieved.

5. Demonstration

Our demonstration is made up of a public-facing web page[6] that provides Galician translations for both Spanish and English inputs. Users will be able to test the system via an open web interface (see Figure 1) where they could select the language pair (*ES–GL* or *EN–GL*) and translation system

[2] https://opus.nlpl.eu
[3] https://opus.nlpl.eu/Europarl.php
[4] https://opus.nlpl.eu/OpenSubtitles.php
[5] https://repositori.upf.edu/handle/10230/20051
[6] https://demos.citius.usc.es/nos_tradutor


GALICIAN TRANSLATION
Please enter your text to be translated!
Neural LSTM and Transliteration Machine Translation System
Please enter your text below
Translate!

Choose Machine Translation System Below: LSTM Spanish-Galician

Input:

El Gobierno utiliza el plan de choque por la guerra para regularizar los suelos contaminados por la radioactividad

Galician Translation:

O Goberno utiliza o plano de choque pola guerra para regularizar os solos contaminados pola radioactividade

Figure 1: A screen capture of the web interface.

(LSTM or Transformer) to then enter text and generate translations.

In our demonstration, we plan to show where our system performs well and where it does not perform well. As an example, the sentence translated from Spanish to Galician using the LSTM system in Table 2 is an excellent translation despite its long length. Additionally, our system translations perform well with syntax and seem to generally translate better than previous systems tested on the same domain. Nonetheless, we have found that when comparing our system's performance for lexical and morphological quality, the Portuguese transliteration affect the performance, found to be better on other rule-based MT systems like Apertium [2] for example.

6. Future work

We plan to perform further work with a human-in-the-loop to increase the performance based on quality. This is outlined by a continuous improvement plan which insinuates the inclusion of translators for user functionality tests. For example, spelling and lexical issues such as *acidente* instead of *accidente*, formal Galician differences that need to be addressed are first to be solved using newly-developed heuristics as part of our future contingency plan. The aim will be to create the highest-quality system in order expand the language pairs to other languages such as Russian or Chinese.

Acknowledgments

This research was funded by the project "Nós: Galician in the society and economy of artificial intelligence", agreement between Xunta de Galicia and University of Santiago de Compostela, and grant ED431G2019/04 by the Galician Ministry of Education, University and Professional Training, and the European Regional Development Fund (ERDF/FEDER program).

References

[1] R. Knowles, J. Ortega, P. Koehn, A comparison of machine translation paradigms for use in black-box fuzzy-match repair, in: Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing, 2018, pp. 249–255.

[2] M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, F. M. Tyers, Apertium: a free/open-source platform for rule-based machine translation, Machine translation 25 (2011) 127–144.

[3] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[4] P. Koehn, R. Knowles, Six challenges for neural machine translation, arXiv preprint arXiv:1706.03872 (2017).

[5] J. Gu, Y. Wang, K. Cho, V. O. Li, Improved zero-shot neural machine translation via ignoring spurious correlations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1258–1268. URL: https://aclanthology.org/P19-1121. doi:10.18653/v1/P19-1121.

[6] J. R. P. Campos, P. M. Fernández, O. Gomez, P. Gamallo, A. C. García, Carvalho: English-galician smt system from europarl english-portuguese parallel corpus, Procesamiento Del Lenguaje Natural (2009) 379–381.

[7] J. E. Ortega, R. C. Mamani, K. Cho, Neural machine translation with a polysynthetic low resource language, Machine Translation 34 (2020) 325–346.

[8] J. E. Ortega, R. A. Castro-Mamani, J. R. Montoya Samame, Overcoming resistance: The normalization of an Amazonian tribal language, in: Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, Association for Computational Linguistics, Suzhou, China, 2020, pp. 1–13. URL: https://aclanthology.org/2020.loresmt-1.1.

Spanish	Galician
Debemos imponer el cumplimiento de los reglamentos y velar por que se aplique el principio de que “el que contamina paga” para que se utilicen sanciones y también incentivos financieros a fin de presionar a los propietarios de los buques y las compañías petroleras y lograr que se introduzcan los procedimientos mejores.	Temos de impor o cumpremento dos regulamentos e celar por que o principio do poluidor-pagador sexa aplicado para que sexan utilizadas sancións e tamén incentivos financeiros a fin de exercer presión sobre os proprietarios dos navíos e das compañías petrolíferas e conseguir que os procedementos mellores sexan introducidos.

Table 2
Translation using the best performing machine translation system (LSTM).

[9] J. R. Pichel, P. Gamallo, I. Alegria, M. Neves, A methodology to measure the diachronic language distance between three languages based on perplexity, Journal of Quantitative Linguistics 28 (2021) 306–336.

[10] K. Knight, J. Graehl, Machine transliteration, arXiv preprint cmp-lg/9704003 (1997).

[11] G. Klein, Y. Kim, Y. Deng, J. Senellart, A. Rush, OpenNMT: Open-source toolkit for neural machine translation, in: Proceedings of ACL 2017, System Demonstrations., Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 67–72. URL: https://www.aclweb.org/anthology/P17-4012.

[12] P. Gamallo, M. Garcia, C. Piñeiro, R. Martinez-Castaño, J. C. Pichel, LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction, in: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2018, pp. 239–244. doi:10.1109/SNAMS.2018.8554689.

[13] S. Garg, S. Peitz, U. Nallasamy, M. Paulik, Jointly learning to align and translate with transformer models, CoRR abs/1909.02074 (2019). URL: http://arxiv.org/abs/1909.02074. arXiv:1909.02074.

[14] T. Kudo, J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, arXiv preprint arXiv:1808.06226 (2018).

[15] M. D. C. Bayón, P. Sánchez-Gijón, Evaluating machine translation in a low-resource language combination: Spanish-galician., in: Machine Translation Summit XVII Vol. 2: Translator, Project and User Tracks, 2019, pp. 30–35.

[16] P. Lison, J. Tiedemann, M. Kouylekov, OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1275.