Lightweight Neural Translation Technologies for Low-Resource Languages

Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz, Víctor M. Sánchez-Cartagena, Andrés Lou, Cristian García-Romero, Aarón Galiano-Jiménez, Miquel Esplà-Gomis

Dep. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant
E-03690 Sant Vicent del Raspeig (Spain)

https://transducens.dlsi.ua.es/lilowla/
{fsanchez,japerez}@dlsi.ua.es

1 Project Overview

The LiLowLa[1] ("Lightweight neural translation technologies for low-resource languages") project, funded by the Spanish Government and the European Regional Development Fund, aims to enhance machine translation (MT) and translation memory (TM) technologies, particularly for low-resource language pairs,[2] where adequate linguistic resources are scarce.[3] Additionally, the project seeks to optimize web crawling methods to gather relevant data for low-resource languages effectively while avoiding unnecessary downloads, thereby reducing crawling times and enabling the acquisition of larger parallel corpora.

The scarcity of linguistic resources is often a result of the low commercial interest in the language pair in question, frequently stemming from the economic constraints of the speaking communities. This also implies that translation technologies developed for low-resource language pairs are likely to be utilized in environments with limited computing capabilities; for this reason, the project focuses on lightweight technologies.

2 Objectives

We define the following objectives:

1. The improvement of the efficiency, robustness and applicability of neural MT systems involving low-resource language pairs.
2. The improvement of web crawling methods to avoid downloading documents that end up being useless after their processing.
3. The widening of the applicability of TMs in professional computer-aided translation (CAT) tools by allowing them to exploit monolingual corpora when MT is not a viable option or when the database of existing translations is not sufficiently large.

To attain these objectives, the project focuses on investigating how to make neural MT significantly more robust and efficient by distilling the knowledge in large pre-trained neural models initially developed for high-resource language pairs, such as NLLB-200, and researching new lightweight data augmentation techniques to make the most of the scarce resources available. It also concentrates on the development of smart focus crawlers to improve current corpus crawling methods, and on the integration of cross-lingual sentence embeddings into CAT tools to permit the search of translation proposals in monolingual corpora.

3 Languages of interest

In addition to the improvement of translation technologies for low-resource language pairs, LiLowLa seeks to build corpora and translation models for a number of languages selected on the basis of social impact and the preservation of their cultural heritage:

- Pairs consisting of Spanish and another language of Spain, including Aragonese, Asturleonese, Aranese, Basque, Catalan and Galician.
- Pairs made of Spanish and Mayan languages spoken in Guatemala and Mexico, such as K'iche', Yucatec, Q'eqchi', Mam, Tzeltal and Kaqchikel.

[1] https://transducens.dlsi.ua.es/lilowla/

[2] The term *low-resource language pair* is commonly used to refer to a combination of two languages for which there are few bilingual resources.

[3] The project will run from September 2022 to August 2025.

4 Expected results

As a result of the execution of the project we plan to deliver:

1. A smart bilingual focus crawler guiding the crawling towards webpages more likely to contain parallel content; thus facilitating quicker discovery and reducing the time and bandwidth needed for acquiring parallel corpora.
2. A multi-task data augmentation method able to get the most of the available parallel corpora and that can be easily integrated in current training workflows.
3. A method for the generation of synthetic parallel sentences from large pre-trained models in the absence of training bilingual corpora, for the purpose of training small student models for low-resource language pairs.
4. A method overcoming the main limitation of TM software —the scarcity of in-domain translation memories— by allowing the retrieval of translation proposals from monolingual corpora, which are much more abundant.
5. Monolingual and bilingual corpora for the languages of interest to the project.
6. Standard test sets based on FLORES+ for targeted languages.
7. Translation models with a reduced size for the languages of interest to the project obtained by distilling the knowledge of large pre-trained models like NLLB-200.

5 Resources

Corpora and software developed as part of the project will be released under free/open-source licenses. In what follows we provide an incomplete list of software and corpora released so far:[4]

MATiLDA: Multitask data augmentation approach able to improve translation performance by generating synthetic training samples with non-fluent target segments (Sánchez-Cartagena et al., 2024).[5]

CrossLingualNeuralFMS: Method for obtaining translation proposal from target-language monolingual corpora in CAT tools (Esplà-Gomis et al., 2022).[6]

Tune 'n' distill: Pipeline to tune the mBART50 model to low-resource language pairs, and distill the resulting system to obtain a lightweight model (Galiano-Jiménez et al., 2023).[7]

PILAR: Collection of parallel and monolingual corpora for low-resource languages of the Iberian Peninsula.[8]

MayanV: Parallel corpora between several Mayan languages and Spanish (Lou et al., 2024).[9]

URL2lang: Tool to infer the language of a document from the URL linking to it.[10]

Parallel URLs Classifier: Tool to infer whether a pair of URLs link to parallel documents.[11]

Acknowledgments

Project (PID2021-127999NB-I00) funded by the Spanish Ministry of Science and Innovation, the Spanish Research Agency (AEI/10.13039/501100011033) and the European Regional Development Fund A way to make Europe.

References

Esplà-Gomis, M., V.M. Sánchez-Cartagena, J.A. Pérez-Ortiz, and F. Sánchez-Martínez. 2022. Cross-lingual neural fuzzy matching for exploiting target-language monolingual corpora in computer-aided translation. In *Proc. of the 2022 EMNLP Conference*, pages 7532–7543, December.

Galiano-Jiménez, A., F. Sánchez-Martínez, V.M. Sánchez-Cartagena, and J.A. Pérez-Ortiz. 2023. Exploiting large pre-trained models for low-resource neural machine translation. In *Proc. of the 24th EAMT Conference*, pages 59–68, June.

Lou, A., J.A. Pérez-Ortiz, F. Sánchez-Martínez, and V.M. Sánchez-Cartagena. 2024. Curated datasets and neural models for machine translation of informal registers between mayan and spanish vernaculars. In *Proc. of the 2024 NAACL Conference*, Mexico City, Mexico, June. In press.

Sánchez-Cartagena, V.M., M. Esplà-Gomis, J.A. Pérez-Ortiz, and F. Sánchez-Martínez. 2024. Non-fluent synthetic target-language data improve neural machine translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):837–850.

[4] For a complete list we refer the reader to https://transducens.dlsi.ua.es/lilowla/lilowla-resources/

[5] https://github.com/transducens/MaTiLDA

[6] https://github.com/transducens/CrossLingualNeuralFMS

[7] https://github.com/transducens/tune-n-distill

[8] https://github.com/transducens/PILAR/

[9] https://github.com/transducens/mayanv

[10] https://github.com/transducens/url2lang

[11] https://github.com/transducens/parallel-urls-classifier