TAN-IBE: Neural Machine Translation for the romance languages of the Iberian Peninsula

Antoni Oliver, Mercè Vàzquez, Marta Coll-Florit, Sergi Álvarez, Víctor Suárez, Claudi Aventín-Boya
Universitat Oberta de Catalunya
{aoliverg,mvazquega,mcollfl,salvarezvid,vsuarezpi,caventinb}@uoc.edu

Cristina Valdés
Universidad de Oviedo
cris@uniovi.es

Mar Font
Universitat de Lleida
mar.font@udl.cat

Alejandro Pardos
Universidad de Zaragoza
apardoscalvo@gmail.com

Abstract

This paper describes the project TAN-IBE: Neural Machine Translation for the romance languages of the Iberian Peninsula, a three year research project founded by the Spanish Ministry of Science and Innovation in the call *Proyectos de generación de conocimiento 2021* (Reference: PID2021-124663OB-I00). This project has started in September 2022.

1 Introduction

The main goal of this project is to explore the techniques for training NMT systems applied to Spanish, Portuguese, Catalan, Galician, Asturian, Aragonese and Aranese, a standardized subvariety of Gascon, which is a variety of Occitan. Aranese has the status of official language in the autonomous community of Catalonia. These languages belong to the same Romance family, but they are very different in terms of the linguistic resources available. Asturian, Aragonese and Aranese can be considered low-resource languages. These characteristics make this setting an excellent place to explore training techniques for low-resource languages: transfer learning and multilingual systems, among others.

The first months of the project have been dedicated to the compilation of monolingual and parallel corpora for Asturian, Aragonese and Aranese.

2 List of partners

- Universitat Oberta de Catalunya[1] (UOC), leading the project and in charge of the technical aspects regarding the training of the neural systems.
- Universidad de Oviedo[2], working in the compilation of corpora for Asturian.
- Universidad de Zaragoza[3], in charge of the compilation of resources for Aragonese.
- Universitat de Lleida[4] (UdL), working in the compilation of texts for Aranese.

3 Project objectives

The project's main objective is to design, train and evaluate NMT systems between the Romance languages of the Iberian Peninsula. This objective will be achieved through the following specific objectives:

- Compiling parallel and monolingual corpora for the languages included in the proposal, paying special attention to Asturian, Aragonese and Aranese.
- Exploring new techniques for training neural machine translation engines.
- Train neural machine translation systems between Spanish and the rest of the languages of the project, in both directions.
- Training neural multilingual systems capable to translate from and to all the languages of the project.
- Evaluating all the trained systems using automatic evaluation metrics and compare them with existing machine translation systems.
- Performing manual evaluations of the machine translation systems developed for Spanish to Asturian, Aragonese and Aranese.
- Creating guides and scripts that facilitate the training of neural machine translation en-

[1] https://www.uoc.edu

[2] https://www.uniovi.es/

[3] https://www.unizar.es/

[4] https://www.udl.cat/ca/

gines.

- Publishing the results of TAN-IBE with open licences.

4 Summary of partial results

The project started on September 2022 and during these first months we have concentrated the activity in the compilation of language resources for Asturian, Aragonese and Aranese. We have also developed several scripts and programs to assist in the tasks of compiling existing parallel corpora and creating new parallel corpora.

4.1 Scripts and programs

Some of the larger available parallel corpora for these languages contain numerous errors: many segment are not in the correct languages, and many parallel segments are not mutual translations. To filter out incorrect segments we have developed a script that rechecks the languages and apply a score based on SBERT (Sentence Embeddings using Siamese BERT-Networks) (Reimers and Gurevych, 2019) to detect misaligned segments. To facilitate the alignment of parallel and comparable corpora a set of programs to ease the process of automatic text aligment with Hunalign (Varga et al., 2007) and SBERT has been developed.

4.2 Corpora

We have developed the FLORES-200 corpus (Goyal et al., 2022) for Aragonese and Aranese, and we have also revised the Asturian version, as it contained errors.

For the creation of the new Spanish–Asturian parallel corpus, various sources were used, including those available on the Internet such as legal texts, Asturian web pages, the Wikidata database, Asturian Wikipedia articles, and literary texts. In addition, agreements were reached with media, publishers, associations, and institutions such as the Directorate-General for Language Policy of the Principality of Asturias or the Office of Language Services of the city councils of Gijón and Corvera.

The selection and the preparation of the corpus in Aragonese language were determined by the specific factors of other minority languages. Among other factors, we can highlight the existence of several orthographic norms and the fact that the official academy of the language has been very recently created. The aid of the Directorate-General for Language Policy has been essential, as they provided a wide corpus, consisting largely of a monolingual corpus, but also containing texts in Spanish and its translation into Aragonese. The greater part of them are translations of legal documents and laws, but it also contains educational material and literature as well. Finally, it's worth mentioning that some of the most important publishing companies in Aragonese language have provided us with literary texts.

Regarding Aranese, the work done to date involves starting the compilation from the normative documents to the current approval and first normalization of this language, which date back to the period after 1982, discarding previous ones. For this reason we have obtained texts in a standardized Aranese from Aranese periodicals of the last thirty years. We have continued with the publications of the few existing Aranese writers who have offered us their entire bibliography, few monographs and online editions that have posted their material online for open use: Associació Centre d'Estudis i Documentació de la Comunicació de la Universitat Autònoma de Barcelona (UAB), Edicions deth Conselh Generau d'Aran (CGA), and other small publishers with whom we have collaborated, providing their Aranese writings.

References

Goyal, Naman, 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.

Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing IV*, pages 247–258. John Benjamins.