TAN-IBE: Neural Machine Translation for the Romance Languages of the Iberian Peninsula

Antoni Oliver[1], Mercè Vàzquez[1], Marta Coll-Florit[1], Sergi Álvarez[1], Víctor Suárez[1], Claudi Aventín-Boya[1], Cristina Valdés[2], Mar Font[3] and Alejandro Pardos[4]

[1]*Universitat Oberta de Catalunya (UOC). Rambla del Poblenou, 156 08018 Barcelona (Spain)*
[2]*Universidad de Oviedo. Campus de Humanidades "El Milán", C/ Amparo Pedregal, s/n, 33011 Oviedo (Spain)*
[3]*Universitat de Lleida. Plaça de Víctor Siurana, 1, 25003 Lleida (Spain)*
[4]*Universidad de Zaragoza. Pedro Cerbuna 12 50009 Zaragoza (Spain)*

Abstract

This paper describes the project TAN-IBE: Neural Machine Translation for the Romance Languages of the Iberian Peninsula, a three-year research project. Its main objective is to conduct research on techniques for training NMT systems for these languages, as there are high, medium and low resource languages among them. Particular attention will be paid to the languages with fewer resources: Asturian, Aragonese and Aranese.

Keywords
Romance languages, neural machine translation, parallel corpora

1. Funding institution and duration

The TAN-IBE project: Neural Machine Translation for the Romance Languages of the Iberian Peninsula is a research project funded by the Spanish Ministry of Science and Innovation in the call for proposals *Proyectos de generación de conocimiento 2021*. The project has a duration of 3 years and it started in September 2022.

2. Project participants

The following institutions are involved in the TAN-IBE project: Universitat Oberta de Catalunya[1] (UOC) which leads the project and is in charge of the training and evaluation of the neural systems; Universidad de Oviedo[2], which is mainly in charge of the compilation of the corpora for Asturian; Universidad de Zaragoza[3], which is mainly in charge of the compilation of the corpora for Aragonese and Universitat de Lleida[4] (UdL), which is mainly responsible for the compilation of the corpora for Aranese.

3. Motivation and background

3.1. Romance languages of the Iberian Peninsula

There is a large number of Romance languages on the Iberian Peninsula. In this project we will consider the following: Spanish, Portuguese, Catalan, Galician, Asturian, Aragonese and Aranese. This list could be extended by other languages and varieties. These languages are very disparate in terms of their official status and the number of speakers. These two factors, official status and number of speakers, correlate in most cases with the linguistic resources (especially for this project we are interested in parallel corpora) and the number and quality of the machine translation systems available. As far as officiality is concerned, we could distinguish three levels: state officiality (officiality in an entire state of the Iberian Peninsula), autonomous or regional officiality (officiality in an autonomous or regional region or at least part of it), and international officiality (officiality in international institutions such as the European Union or the United Nations). Table 1 shows the level of officiality and the approximate number of speakers of these languages on the Iberian Peninsula.

For example, Catalan is official in the state of Andorra and official in several autonomous communities

SEPLN-PD 2023: Annual Conference of the Spanish Association for Natural Language Processing 2023: Projects and System Demonstrations

aoliverg@uoc.edu (A. Oliver); mvazquezga@uoc.edu (M. Vàzquez); mcollfl@uoc.edu (M. Coll-Florit); salvarezvid@uoc.edu (S. Álvarez); vsuarezpi@uoc.edu (V. Suárez); caventinb@uoc.edu (C. Aventín-Boya); cris@uniovi.es (C. Valdés); mar.font@udl.cat (M. Font); apardoscalvo@gmail.com (A. Pardos)
0000-0001-8399-3770 (A. Oliver); 0000-0002-7983-4029 (M. Vàzquez); 0000-0002-8636-458X (M. Coll-Florit); 0000-0002-6355-4559 (S. Álvarez); 0000-0002-7631-7552 (V. Suárez); 0000-0001-6607-5281 (C. Valdés); 0000-0001-5323-655X (M. Font)

[1] https://www.uoc.edu
[2] https://www.uniovi.es/
[3] https://www.unizar.es/
[4] https://www.udl.cat/ca/

Languages	ISO	S.O.	A.O.	I.O.	Speakers
Spanish	spa	X	X	X	46.000.000
Portuguese	por	X	X	X	11.000.000
Catalan	cat	X	X		10.000.000
Galician	gal		X		2.500.000
Asturian	ast				200.000
Aragonese	arg				30.000
Aranese	oci*		X		1.500

Table 1
Level of officiality and number of speakers on the peninsula of the Romance languages of the Iberian Peninsula (S.O.: state officiality; A.O: autonomous or regional officiality; I.O.: international officiality).

and Aranese is official in the entire territory of the autonomous community of Catalonia.

3.2. Existing linguistic resources

In table 2 we can observe the approximate total number of segments in the parallel corpora available in the OPUS collection between Spanish and the other languages under study.

Language	Segments
Portuguese	249.7 M
Catalan	105.7 M
Galician	37.3 M
Asturian	7.4 M
Aranese	1.5 M

Table 2
Number of segments available in the parallel corpora available in the Opus Corpus collection between Spanish and other languages.

Another interesting resource for training machine translation engines are monolingual corpora, since there are techniques capable of training systems using monolingual corpora. For Spanish, Portuguese, Catalan and Galician, large amounts of text can be easily collected from Common Crawl, which periodically downloads all web content and makes the downloaded data available. A language detection algorithm is applied to this download to request the data for a given language. Unfortunately, for the rest of the languages under study no data is available, as the language detector used is not trained to detect these languages. Another possible source of monolingual corpora is Wikipedia, which has versions for all the languages in this project (with the exception of Aranese, which could experimentally use Occitan data). Table 3 shows the number of Wikipedia articles for each of the project languages.

Language	Articles
Spanish	402.430
Portuguese	429.730
Catalan	133.214
Galician	39.627
Asturian	11.734
Aragonese	10.552
Aranese*	14.584

Table 3
Number of articles in Wikipedia for each of the languages of the project (for Aranese the data corresponding to Occitan is provided).

With regard to the machine translation systems available between Spanish and the other languages, we will analyze three specific systems: Apertium, which is a shallow syntactic transfer system distributed under a free license; Google Translate, a very popular neural machine translation system that provides numerous language pairs; and DeepL, a commercial neural system that is also well known for its quality. In Table 4 we can observe the systems from Spanish to the rest of the languages in this study.

	Apertium	GoogleT	DeepL
Portuguese	X	X	X
Catalan	X	X	
Galician	X	X	
Asturian	X		
Aragonese	X		
Aranese	X		

Table 4
Availability of Spanish to other languages for three widely used machine translation systems.

As can be seen from Table 4, only three languages (Portuguese, Catalan and Galician) have a neural machine translation system with Spanish as the source language. Currently, the predominant methodology and the one that achieves better quality is neural machine translation

[1]. Thus, most of the Romance languages under study do not have machine translation systems using this methodology. Neural machine translation systems are trained using parallel corpora of good quality and large size. The data in Table 2 are not encouraging for languages that do not have neural machine translation systems, as there are no corpora of sufficient size for these languages. There is therefore an urgent need for larger parallel corpora for these languages.

3.3. Training strategies for under-resourced language pairs

In recent years, there has been considerable interest in the development of methodologies for training neural machine translation systems for language pairs with very few resources.

Four major groups of strategies can be distinguished: neural machine translation based on transfer learning; multilingual machine translation; self-supervised machine translation; and unsupervised machine translation. During the project we intend to explore the first two strategies.

3.3.1. MT based on transfer learning

We want to train a machine translation system from language A to language C, but this language pair has very few parallel segments available. But there is a language B, which is closely related to language C (for example, they are close languages of the same family, like the working languages of this project) and we have large parallel corpora between language A and B. Using so-called transfer learning, we start by training a neural system from language A to language B and, once the training is finished, we continue training it using a corpus of the language pair B-C [2]. In [3] a modification to this methodology using vocabulary overlap between these languages is introduced. To increase the overlap in the vocabulary, they split the words into subwords using BPE (Byte Pair Encoding) [4]. They then train the A-B system and transfer the parameters including word embeddings from the source language to another model and continue training the B-C system. In the TAN-IBE project a Spanish-Aranese system could be trained by first training a Spanish-Catalan with a large corpus and once trained continue training with the Catalan-Aranese corpus.

3.3.2. Multilingual MT

Mutilingual machine translation systems [5] allow us to train a single neural system that shares a single attention mechanism. Imagine we are working with languages A, B, C and D. If we have a parallel corpus for some of these language combinations (e.g. A-B, A-C, A-D, B-C and B-D) we can train a machine translation system that can translate between all pairs, regardless of the fact that for some of the language pairs there is no parallel corpus available (e.g. the C-D pair in our example). This is possible because the resulting system is able to use the similarities between the languages. This configuration can be very useful to train systems for language pairs with few resources while training language pairs with more resources. In our project the Spanish-Portuguese, Spanish-Catalan and Spanish-Galician pairs would be the resource-rich pairs; while Spanish-Asturian, Spanish-Aragonese and Spanish-Aranese would be the resource-poor pairs. This same configuration could produce translation systems for pairs without any parallel corpus, such as Asturian-Aranese. This is called *zero-shot translation*. In [6] it is shown that the quality of these *zero-shot* translations can be significantly improved if a few parallel segments of the C-D pair (Asturian-Aranese, in the example above) are available. In [7] it is emphasized that most multilingual systems take English as the core language, since they are trained only with parallel corpora consisting of texts that have been translated from English or into English. In their work they show that and improvement of up to 10 points in BLEU can be achieved by using non-English-centric models in the translation of non-English language pairs. This work is important for our project, as the core language will not be English, which is not the language we intend to work with in this project. Another aspect that has occupied the attention of researchers is the influence of typological differences between the languages involved in a multilingual system. In some studies [8] *backtranslation* is used in multilingual systems to improve the translation quality of language pairs for which no parallel corpus is available. The technique of *backtranslation* [9] consists of using monolingual corpora of the target language (B) to create a parallel corpus where sentences in the source language (A) are obtained using a machine translation system of the B-A language pair. This new synthetic parallel corpus is added to the available real A-B parallel corpus and both models are used to train the new A-B machine translation system. It is important to note that the only synthetic part of the synthetic parallel corpus obtained by *backtranslation* is the part corresponding to the source language (A), since the part corresponding to the target language (B) has been obtained from real language texts.

4. Goals of the project

The main objective of the project is the design, training and evaluation of neural machine translation systems between the Romance languages of the Iberian Peninsula.

This main objective can be divided into the following specific objectives:

- To compile parallel and monolingual corpora for the languages of the project, with a special effort for Asturian, Aragonese and Aranese.
- To explore new techniques for training neural translation engines.
- To train neural translation systems between Spanish and the other languages of the project, in both directions.
- To train multilingual systems capable of translating to and from all the languages of the project.
- To evaluate all trained systems using automatic metrics and compare them with existing machine translation systems.
- To perform human evaluations of the trained systems between Spanish and Asturian, Aragonese and Aranese.
- To create guides and scripts that facilitate the training of neural machine translation systems.
- To publish the results of the TAN-IBE project under free licenses.

5. Summary of results to date

During the first months, the activity has focused on the compilation of linguistic resources for Asturian, Aragonese and Aranese. Several scripts and programs have also been developed to facilitate the task of compiling parallel corpora.

5.1. Scripts and programs

Some of the larger parallel corpora for the languages of the project contain numerous errors: many segments are not in the required languages and many others are not translation equivalents. To filter out incorrect segments we have developed a script that reverifies the languages and applies a score based on SBERT [10] to detect misaligned segments. To facilitate the alignment of parallel corpora and the search for parallel segments in comparable corpora we have developed a set of programs that facilitate the process using Hunalign [11] and SBERT.

5.2. Corpora

We have developed the FLORES-200 [12] corpus for Aragonese and Aranese, and have thoroughly revised the Asturian version, because it contained errors.

For the creation of the parallel Spanish-Asturian corpus we are using various sources, mainly those available on the Internet such as legal texts, web pages and Wikipedia, and texts obtained through agreements with the media, publishers and institutions such as the Academia de la Llingua Asturiana, the Directorate General of Language Policy of the Principality of Asturias and the linguistic normalization services of the city councils of Gijón and Corvera. We would also like to highlight the ESLEMA material provided by researchers from the University of Oviedo and the compilation of various literary works.

The selection and preparation of the corpus for Aragonese has been conditioned by the fact that it is a minority language. Among other factors, we can highlight the lack of linguistic standardization, the absence of a reference institution regarding the proper use of the language or the diversity of orthographic rules used by the different associations and organizations. There is abundant literature on the early years of the *renaxedura de l'aragonés* (the rebirth of the language, mainly in the 1980s), in which a large number of books, magazines and journals were published, and a downward trend in the corpus observed from the second half of the 2000s until 2015. The lack of institutional recognition, internal discordance between associations and the limited presence of the language on the Internet or media can be pointed out as the main factors. However, the creation in 2015 of the Directorate General of Language Policy of the Government of Aragon has significantly increased the corpus by promoting the use of this language in education, literature, the Internet, the media, university and scientific research and reaching a better agreement on orthographic rules and linguistic standardization. The assistance of the Directorate General for Language Policy has been fundamental, since it has provided a large corpus, largely composed of monolingual texts, but also containing texts in Spanish and their translation into Aragonese. Most of them are translations of legal documents and laws, but also educational material and literature. The institution also provided a large database with the contents of the *Aragonario* (the reference dictionary of the Aragonese language), which contains the translation of practically all known words in Aragonese. Finally, it should be noted that the participation of three of the four most relevant publishers in the Aragonese language has been important in order to have a really limited corpus on the literary field published in recent years.

As for Aranese, the work carried out to date has involved starting the compilation from the normative documents up to the current approval and first standardization of this language, which date from the period after 1982, discarding the previous ones. For this reason, we have obtained texts in standardized Aranese from Aranese newspapers of the last thirty years. We have continued with the publications of the few existing Aranese writers who have offered us their entire bibliography, some monographs and online editions that have provided their material for open use: Associació Centre d'Estudis i Documentació de la Comunicació (UAB), Edicions deth Conselh (CGA), and other small publishers with whom we have collaborated, providing their writings in Aranese.

Acknowledgments

The project TAN-IBE: Neural Machine Translation for the Romance languages of the Iberian Peninsula is funded by the Spanish Ministry of Science and Innovation. Reference: PID2021-124663OB-I00 funded by MCIN /AEI /10.13039/501100011033 / FEDER, EU.

References

[1] S. Castilho, J. Moorkens, F. Gaspari, R. Sennrich, V. Sosoni, P. Georgakopoulou, P. Lohar, A. Way, A. V. Miceli-Barone, M. Gialama, A comparative quality evaluation of pbsmt and nmt using professional translators, in: Proceedings of Machine Translation Summit XVI: Research Track, 2017, pp. 116–131.

[2] B. Zoph, D. Yuret, J. May, K. Knight, Transfer learning for low-resource neural machine translation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1568–1575.

[3] T. Q. Nguyen, D. Chiang, Transfer learning across low-resource, related languages for neural machine translation, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2017, pp. 296–301.

[4] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1715–1725.

[5] O. Firat, K. Cho, Y. Bengio, Multi-way, multilingual neural machine translation with a shared attention mechanism, in: 15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016, Association for Computational Linguistics (ACL), 2016, pp. 866–875.

[6] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al., Google's multilingual neural machine translation system: Enabling zero-shot translation, Transactions of the Association for Computational Linguistics 5 (2017) 339–351.

[7] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, et al., Beyond english-centric multilingual machine translation, The Journal of Machine Learning Research 22 (2021) 4839–4886.

[8] B. Zhang, P. Williams, I. Titov, R. Sennrich, Improving massively multilingual neural machine translation and zero-shot translation, in: 2020 Annual Conference of the Association for Computational Linguistics, Association for Computational Linguistics (ACL), 2020, pp. 1628–1639.

[9] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, in: 54th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (ACL), 2016, pp. 86–96.

[10] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[11] D. Varga, P. Halácsy, A. Kornai, V. Nagy, L. Németh, V. Trón, Parallel corpora for medium density languages, in: Recent Advances in Natural Language Processing IV, John Benjamins, 2007, pp. 247–258.

[12] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, A. Fan, The flores-101 evaluation benchmark for low-resource and multilingual machine translation, Transactions of the Association for Computational Linguistics 10 (2022) 522–538.