TRIBBLE - TRanslating IBerian languages Based on Limited E-resources

Igor Kuzmin[1] **Piotr Przybyła**[1,2] **Euan McGill**[1] **Horacio Saggion**[1]
[1] LaSTUS Lab, TALN Group, Universitat Pompeu Fabra, Barcelona, Spain
[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
{igor.kuzmin, piotr.przybyla, euan.mcgill, horacio.saggion}@upf.edu

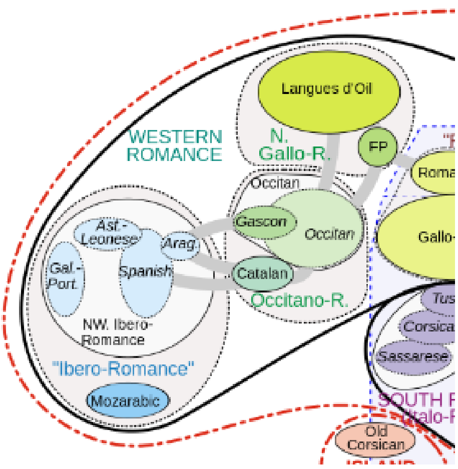

Figure 1: Language family tree diagram (partial) focusing on the Iberian peninsula

1 Introduction

In this short overview paper[1], we describe our system submission for the language pairs Spanish→Aragonese (spa-arg), Spanish→Aranese (spa-arn), and Spanish→Asturian (spa-ast)[2]. We train a unified model for all language pairs in the **constrained** scenario. In addition, we add two language control tokens for Aragonese and Aranese Occitan, as there is already one present for Asturian.

1.1 Linguistic background

The Iberian peninsula - which includes the territory of Spain, Portugal and Gibraltar - is a hotspot for linguistic diversity, especially among languages in the Romance family. Spanish, Portuguese and English have official status across these three respective territories.

Basque (a non-Indo-European language) has co-official status in the Spanish Autonomous Communities of the Basque Country and the northern portion of Navarre. In Galicia, Galician is co-official and in the Balaeric Islands, the Valencian Community and Catalonia Catalan/Valencian also enjoys this status.

This status ensures visibility of these languages in the socio-political space as well as a sizeable presence online. Catalan, Basque and Galician are included in many high-performing machine translation (MT) systems (and large language models (LLMs) capable of the task) (Armengol-Estapé et al., 2021) and benchmarks (Federmann et al., 2022).

This is not necessarily the case for the languages which are the focus of this challenge. They are a diverse set of languages, all from different subbranches of the Romance language family. Figure 1 shows their relation to other languages in the Romance family, and to each other, using the wave model (Heggarty et al., 2010) of linguistic evolution. Note the dialect continuum which appears to form between Portuguese → Asturian → Spanish → Aragonese → Catalan and Gascon Occitan.

Figure 2 provides a visual overview of the languages that are translated into from Spanish as part of this challenge. **Aranese**, a dialect of Gascon Occitan, also has co-official status in Catalonia but provision is only made in the Aran Valley for its use.

Aragonese and **Asturian** are spoken by larger numbers of people, but mostly as either second language learners or legacy speakers such as the elderly. It is for this reason that these languages all fall under the category of "Endangered" languages according to Ethnologue (Eberhard et al., 2024). All three languages are, however, considered "Vital" in terms of Digital Language Support (Simons et al., 2022). This is the second highest category behind "Thriving", meaning that there are extant corpora and resources available. However, this does not necessarily mean that there is decent quality technology such as MT available for these lan-

[1] Igor Kuzmin and Euan McGill are corresponding authors
[2] Submission IDs #622, #623, and #624 respectively

Proceedings of the Ninth Conference on Machine Translation, pages 955–959
November 15-16, 2024

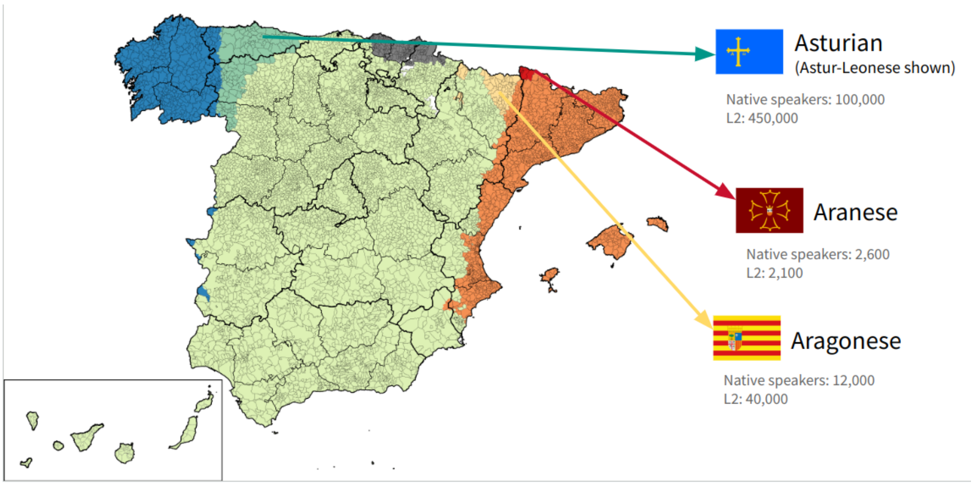

Figure 2: The languages involved in the WMT shared task and some demographic information

guages.

1.2 Extant technology for these languages

There is a recent increased push towards including languages with a small digital presence in language technology (Bapna et al., 2022), and effort has been made already to cover the languages of this challenge, including efforts to generate clean corpora from multilingual content from the internet (González and Álvarez, 2023; Ruder et al., 2023).

The first rule-based system to involve translation into and between the present languages is the open source Apertium (Forcada et al., 2011). Other systems and improvements have been built on top of this service such as Softcatalà (Ivars-Ribes and Sánchez-Cartagena, 2011) which focuses on translation into and out of Catalan, and a neural MT translator (NMT) between Spanish→Aragonese (Cortés et al., 2012).

In addition, the Spanish government-funded TAN-IBE project (Oliver et al., 2023) - of which this challenge is a part - seeks to apply modern techniques across NMT and LLM-based approaches to improve this low-resource MT task.

2 System description

We take the distilled NLLB-200 model (Costa-jussà et al., 2022) with 600M parameters and extend special tokens with 2 tokens that denote target languages (arn_Latn, arg_Latn) because Asturian was already presented in NLLB-200 model. After we initialized the weights of the new tokens using weights from existing tokens in the vocabulary. We used oci_Latn (Occitan) for arn_Latn (Aranese) and spa_Latn (Spanish) for arg_Latn (Aragonese) because this languages are from the corresponding language family.

2.1 Training and data filtering

To create our corpus, we sampled OPUS[3] and PILAR[4] FLORES+ (revised pairs), which contain Catalan->Aranese (from PILAR), Spanish->Aranese, Spanish->Occitan, Spanish->Asturian and Spanish->Aragonese directions. We used Apertium (Khanna et al., 2021) to translate Catalan to Spanish, but we kept both source languages in our training set. Additionally, for the Occitan target language, we used idiomata cognitor (Galiano-Jiménez et al., 2024) to keep only corresponding target languages. We applied the adapted MOSES Punctuation Normalizer provided by Meta Research group under the stopes library[5] for all language pairs because NLLB was trained on preprocessed texts. Further data filtering followed the NLLB paper (Costa-jussà et al., 2022). We used fastText[6] to delete all pairs with English examples. After that, we computed length ratios and kept all sentences where the length was from five to 1050 characters, with a max length ratio lower than 0.9 and a unique ratio higher than 0.125. Finally, we de-duplicated all translation language pairs, keeping a maximum of two source duplicates and three target duplicates. Additionally we kept all pairs where distance score was in [0.6;1.0]. The result distribution of the source and target languages in

[3] https://opus.nlpl.eu/
[4] https://github.com/transducens/PILAR
[5] https://github.com/facebookresearch/stopes/blob/main/stopes/pipelines/monolingual
[6] https://fasttext.cc/

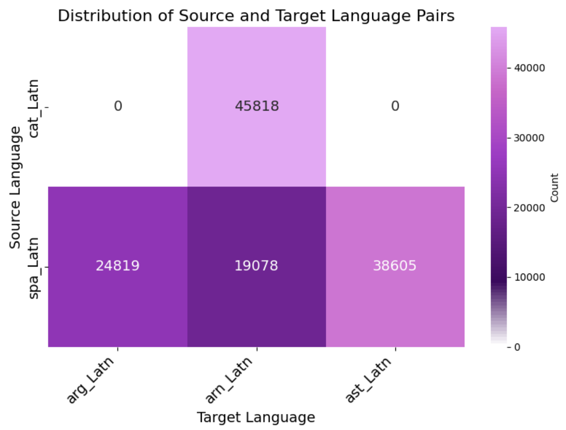

Figure 3: Distribution of language pairs from processed dataset.

our result corpora is captured at the Figure 3.

For the rest of the language pairs, we excluded all samples where the target language did not match the language predicted by idiomata cognitor.

2.2 Data augmentation

We adapt the model by training on a special regime of data augmentation with both monolingual and bilingual training data for the language pairs in this challenge.

The OPUS data were filtered in order to discard the spurious sentence pairs. We do that by performing translation of the Spanish sentence to the appropriate target language using Apertium and comparing the translation to the sentence present in the corpus. We assume that certain differences are possible due to imperfect performance of Apertium and natural variability of language, but the two variants should preserve some resemblance. To quantify that, we compute the Levenshtein (Levenshtein, 1966) edit distance ($d(s_1, s_2)$) between the two strings (s_1, s_2) and transform it into a similarity score defined as:

$$sim(s_1, s_2) = 1.0 - \frac{d(s_1, s_2)}{max(|s_1|, |s_2|)}$$

Based on manual analysis of the scores, we assume the similarity score of minimum 0.6 to be sufficient for the sentence pair to be used. Otherwise, it is discarded.

2.3 Fine tuning

The NLLB-200 model with 600M parameters, distilled from a 54B parameter Mixture-of-Experts model, demonstrated superior performance compared to the baseline version. Building on this foundation, we implemented a series of adaptation steps described above to further enhance the model's capabilities on a new target languages. In this sections, we detail our training methodology and the specific hyperparameters employed to optimize the model's performance across diverse linguistic tasks. The fine-tuning process was done with one T4 GPU using Hugging Face Transformers (Wolf et al., 2020) library with the following hyperparameters presented at the Table 1. Our result model is available at the Hugging Face repository[7].

Hyperparameter	**Value**
Learning Rate	1e-4
Weight Decay	1e-3
Train Batch Size	4
Eval Batch Size	4
Training Epochs	2
Optimizer	Adafactor
Clip Threshold	1.0
Warmup Steps	10% of total steps

Table 1: Hyperparameters for NLLB-200 Fine-tuning.

3 Results

Our results for the translation task from the Spanish language test set[8] to the target languages, as evaluated through OCELoT[9] submission system are reasonably positive, with respective BLEU and chrF+ scores of 49.2 and 73.6 for spa-arg, 17.9 and 15.5 for spa-arn, and 23.9 and 46.1 for spa-ast.

In terms of comparing the current approach with previous approaches such as Apertium and its successors, many of these studies only report word error rate whereas we used BLEU and chrF+. In those studies where BLEU is reported, it is known that BLEU favours SMT and NMT systems over rule-based ones. Moreover, this challenge introduces the present test set - so there is no previous work on the same data for direct comparison.

We find that this method of training is relatively efficient, with energy usage of 2.93kWh and emissions of approximately 1.81kg of CO_2[10].

[7] https://huggingface.co/igorktech/tribble-600m

[8] https://github.com/transducens/wmt2024-romance-tests

[9] https://ocelot-west-europe.azurewebsites.net/leaderboard/4

[10] https://wandb.ai/igorktech01/wmt24-tribble/runs/5z9r7tjt

Acknowledgements

The work of Piotr Przybyła is part of the ERINIA project, which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101060930. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the funders.

Euan McGill and Horacio Saggion would like to acknowledge that:

This work is part of Maria de Maeztu Units of Excellence Programme CEX2021-001195-M, funded by MCIN/AEI /10.13039/501100011033

Amb el suport del Departament de Recerca i Universitats de la Generalitat de Catalunya.

References

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online. Association for Computational Linguistics.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages. *Preprint*, arXiv:2205.03983.

Juan Pablo Martínez Cortés, Jim O'Regan, and Francis M Tyers. 2012. Free/open source shallow-transfer based machine translation for spanish and aragonese. In *LREC*, pages 2153–2157.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. Ethnologue: Languages of the World. Twenty-seventh edition.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

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.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024. Idiomata cognitor.

Antoni Oliver González and Sergi Álvarez. 2023. Filtering and rescoring the CCMatrix corpus for neural machine translation training. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 39–45, Tampere, Finland. European Association for Machine Translation.

Paul Heggarty, Warren Maguire, and April McMahon. 2010. Splits or waves? trees or webs? how divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1559):3829–3843.

Xavier Ivars-Ribes and Victor M. Sánchez-Cartagena. 2011. A widely used machine translation service and its migration to a free/open-source solution: the case of softcatalà. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 61–68, Barcelona, Spain.

Tanmai Khanna, Jonathan N Washington, Francis M Tyers, Sevilay Bayatlı, Daniel G Swanson, Tommi A Pirinen, Irene Tang, and Hèctor Alòs i Font. 2021. Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, 35(4):475–502.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

Antoni Oliver, Mercè Vàzquez, Marta Coll-Florit, Sergi Álvarez, Víctor Suárez, Claudi Aventín-Boya, Cristina Valdés, Mar Font, and Alejandro Pardos. 2023. TAN-IBE: Neural machine translation for the romance languages of the Iberian peninsula. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages

495–496, Tampere, Finland. European Association for Machine Translation.

Sebastian Ruder, Jonathan Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore. Association for Computational Linguistics.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.