TAN-IBE participation in the Shared task: Translation into Low-Resource Languages of Spain

Antoni Oliver
TAN-IBE team - Universitat Oberta de Catalunya (UOC)
aoliverg@uoc.edu

Abstract

This paper describes the systems presented by the TAN-IBE team into the WMT24 Shared task Translation into Low-Resource Languages of Spain. The aim of this joint task was to train systems for Spanish-Asturian, Spanish-Aragonese and Spanish-Aranesian. Our team presented systems for all three language pairs and for two types of submission: for Spanish-Aragonese and Spanish-Aranese we participated with constrained submissions, and for Spanish-Asturian with an open submission.

1 Introduction

The TAN-IBE team, consisting of participants from the project *TAN-IBE: Neural Machine Translation for the Romance Languages of the Iberian Peninsula* (Oliver et al., 2023), has developed systems for all language pairs in the task and participated with two types of submissions: Spanish-Asturian (open submission), Spanish-Aragonese (constrained submission), and Spanish-Aranese (constrained submission).

The principal concern of the team was not only the scarcity of resources for these language pairs but also the inadequate quality of the available resources. In order to address these issues, we decided to:

- Clean the existing parallel corpora using a tool developed during the project that rechecks the language of the segments and calculates the cosine similarity between source and target segments.
- Create parallel corpora from Wikipedia for all three language pairs.
- Experiment with the use of backtranslation.
- Experiment with the use of synthetic corpora.
- Experiment with the use of multilingual systems.

For some of the language pairs and direction of the shared task, a freely available rule-based MT system exists: Apertium[1] (Forcada et al., 2011; Khanna et al., 2021). In previous research (Oliver, 2020), it was demonstrated that Apertium achieves highly competitive quality, and that it is challenging for a neural system to achieve superior quality results. Consequently, this system has been employed to create backtranslated and synthetic corpora. Furthermore, Apertium will be employed as a reference system to facilitate the evaluation of the trained systems prior to submission to the shared task. The specific versions of Apertium used for each language pair are: Spanish–Aragonese 0.6.0,[2] Spanish–Aranese 1.0.8,[3] Spanish–Asturian 1.1.1.[4]

2 Tools

To train the NMT systems we have used the marian-nmt[5] (Junczys-Dowmunt et al., 2018). All the systems have been trained with a Transformer-big configuration. To calculate the subwords units, we have used SentencePiece[6] (Kudo and Richardson, 2018). Additional information on other training parameters can be found in the subsections for each system.

In order to create the parallel corpora and to clean and preprocess the corpora, several components of the MTUOC project[7] (Oliver and Alvarez, 2023) have been employed. It should be noted that several of these components have been developed during the course of the TAN-IBE project. The components that have been used are as follows:

- To create parallel corpora from the Wikipedia: MTUOC-WikipediaDump[8] and MTUOC-

[1] https://apertium.org/

[2] https://github.com/apertium/apertium-spa-arg/releases/tag/v0.6.0

[3] https://github.com/apertium/apertium-oc-es/releases/tag/v1.0.8

[4] https://github.com/apertium/apertium-spa-ast/releases/tag/v1.1.1

[5] https://marian-nmt.github.io/

[6] https://github.com/google/sentencepiece

[7] https://mtuoc.github.io/

[8] https://github.com/mtuoc/MTUOC-WikipediaDumps

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 871–877
November 15–16, 2024.

aligner.[9]

- To clean both the existing and the newly created parallel corpora: MTUOC-clean-parallel-corpus,[10] that performs several cleaning operations; and MTUOC-PCorpus-rescorer,[11] that rechecks the language of the segments and calculates the cosine similarity using a multilingual model, as SBERT, for example.
- To select from a large corpus the most similar segments from a large corpus: MTUOC-corpus-combination.[12]
- To preprocess the parallel corpora to train the systems: MTUOC-corpus-preprocessing.[13]

For some cleaning operations the language of the segments should be detected. As Asturian, Aranese and Aragonses are underrepresented in available language detection models, we decided to develop our own language detection model using fasttext[14] (Joulin et al., 2016). We trained a model able to detect the following languages: Aragonese, Aranese, Asturian, Catalan, English, French, Galician, Occitan, Portuguese and Spanish. For all the languages, except for Aranese, we used 400K segments from the Wikipedia. As no Wikipedia is available for Aranese, we used 297,557 segments from the PILAR corpus.[15]

The trained model performs similarly to Idiomata Cognitor[16] (Galiano-Jiménez et al., 2024), with the difference than Italian is included in this tool. In Table 1, the precisions calculated using the FLORES+ dev corpus for Idiomata Cognitor and our trained language detection model are presented.

We decided to train our own language detection model because fasttext models integrate seamlessly in our corpus cleaning scripts, and the same training strategy can be used in future experiments with language not present in Idomata Cognitor.

3 Existing resources

Table 2 provides an overview of the existing corpora that have been employed for system training,

[9] https://github.com/mtuoc/MTUOC-aligner

[10] https://github.com/mtuoc/MTUOC-clean-parallel-corpus

[11] https://github.com/mtuoc/MTUOC-PCorpus-rescorer

[12] https://github.com/mtuoc/MTUOC-corpus-combination

[13] https://github.com/mtuoc/MTUOC-corpus-preprocessing

[14] https://fasttext.cc/

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

[16] https://github.com/transducens/idiomata_cognitor

Language	Idiomata Cognitor	TAN-IBE fasttext model
Spanish	0.95	0.95
Catalan	1.00	0.99
Aragonese	0.96	1.00
Aranese	0.96	1.00
Occitan	0.94	0.93
Asturian	0.99	0.98
Galician	0.98	0.99
French	1.00	1.00
Portuguese	1.00	0.99

Table 1: Precision on language detection for Idiomata Cognitor and our trained fasttext model.

accompanied by the number of segments in the original corpora and the number of segments resulting from the cleaning process peformed with MTUOC-PCorpus-rescorer. As previously stated, this tool performs a second language detection of the segments using fasttext and calculates a cosine similarity between the source and target segments using SBERT. As the default language detection model used in fasttext have been trained with underrepresented texts for Asturian, Aragonese and Occitan (Aranese), we decided to retrain a language model for the cleaning of corpora for these languages, as explained in section 2. Please, note that there are no available parallel corpora for Aranese, and in the table we state the figures for the Spanish-Occitan parallel corpus used.

Langs	Corpus	Raw	Clean
spa-ast	NLLB	6,470,015	504,532
spa-arg	Wikimatrix	33,724	16,456
spa-oci	NLLB	925,448	108,440

Table 2: Size of the existing corpora in segments used for training the systems

4 Newly created resources

As the available corpora for the working language pairs are clearly insufficient to train NMT systems we have created a new parallel corpus from Wikipedia dumps. To this end, we have developed a series of scripts, which are freely available at the MUTOC-WikipediaDumps repository, that are capable of:

- Extract all the text from the Wikipedia dump, along with a file containing the titles of the articles. This process is performed for the smaller Wikipedias, in this case the Asturian, Aragonese and Aranese.
- Translate the list of titles of the extracted articles into Spanish using the langlinks database dump.

- Extract the text of the articles of the larger Wikipedia, in our case the Spanish one, restricting the extracted articles to the titles in the translated title list.

It should be noted that this process is carried out separately for Spanish-Asturian, Spanish-Aragonese and Spanish-Occitan. Once the text has been obtained, it is segmented and the resulting segments are deduplicated and a file is generated for each language pair, containing all the source segments and a file for all the target segments. For the creation of the Wikipedia corpora we have used the dumps of first of May of 2024. All the created corpora can be downloaded from Github.[17]

To align the files we use a bitext mining strategy using SBERT, implemented in the MTUOC-Aligner. The alignment process gives sets of aligned files that contain source segment, target segment and a margin score. The resulting aligned files are cleaned using MTUOC-PCorpus-rescorer, using a confidence of 0.75 for language detection and SBERT score. In table 3 illustrates the raw and cleaned number of parallel segments obtained through this process.

Langs	Raw	Clean	Backtrans.
spa-ast	2,194,031	1,193,264	4,023,140
spa-arg	153,863	32,374	386,666
spa-oci	295,476	55,112	1,151,205

Table 3: Size of the Wikipedia corpora created

Additionally, the extracted text from Wikipedia in Asturian, Aragonese and Occitan has been employed to generate back-translated Spanish-Asturian, Spanish-Aragonese and Spanish-Occitan corpora. In order to achieve this, it is necessary to have access to machine translation systems that are capable of functioning in the opposite direction. Fortunately, Apertium provides translation systems for Aragonese-Spanish and Occitan-Spanish, but not for Asturian-Spanish. To address this gap, we have trained a Transformer Neural system using the cleaned NLLB and the Wikipedia corpus created for this purpose. This system was then used for backtranslation. In subsection 5.1 more details on this system, along with evaluation figures are presented. Table 3 shows the number of segments in the backtranslated Wikipedia corpus.

As previously stated, there are no available corpora for the Spanish-Aranese language pair, and there is no Wikipedia version for Aranese either. Therefore, no parallel resources for this language pair can be employed. For this language pair, we have utilised the monolingual data available in the PILAR (*Pan-Iberian Language Archival Resource*) and backtranslated it into Spanish using Apertium. This process yielded 297,557 backtranslated parallel segments.

In order to augment the number of parallel corpora, we devised a method for the generation of synthetic data utilising the Spanish-Catalan Paracrawl corpus, which had been previously subjected to cleaning procedures. The 13 million Spanish segments were translated into Asturian, Aragonese and Aranese using Apertium. For each translated segment, a confidence score was calculated as the ratio of the unknown words (marked with an asterisk by Apertium) to the total number of words.

In the case of the multilingual system under experimentation, the parallel corpora presented in Table 4 were also employed. With regard to the Spanish-Catalan and Spanish-Galician languages, a parallel corpus was created from Wikipedia, as previously described.

Langs	Corpus	Raw	Clean
spa-cat	Paracrawl	17,238,953	13,931,594
spa-cat	Wikipedia	5,790,903	2,586,448
spa-fra	MultiParacrawl	39,026,138	-
spa-fra	WikiMatrix	905,761	-
spa-glg	Paracrawl	1,879,649	-
spa-glg	Wikipedia	1,697,307	729,840
spa-por	MultiParacrawl	26,181,054	-
spa-por	WikiMatrix	923,725	-

Table 4: Additional parallel corpora used to train the multilingual system

5 Trained systems and evaluation

In this section we will present all the trained systems for this shared task, along with evaluation figures using the FLORES+ devtest sets developed by the organisers of the shared task. To calculate the evaluation metrics (BLEU, TER and chrF will be presented), we have used Sacrebleu[18] (Post, 2018). In Appendix 8 we present the metric signatures for these metrics.

5.1 Neural Asturian-Spanish for backtranslation

In order to create back-translated corpora for Aragonese and Aranese, the Apertium system was

[17] https://github.com/mtuoc/WikipediaCorpora

[18] https://github.com/mjpost/sacrebleu

employed, given that it is available for these language pairs. However, the Asturian-Spanish pair is not available in Apertium. Consequently, a neural system was trained using Marian and the following parallel corpora: NLLB cleaned (see table 2) and the newly created Wikipedia corpus (see table 3). A transformer configuration using SentencePiece with a vocabulary size of 32K has been used.

Table 5 presents the evaluation metrics for the trained systems, along with the metrics for the Apertium system for Aragonese-Spanish and Aranese-Spanish, using the FLORES+ devtest sets.[19] For purposes of comparison, the evaluation figures for all reverse translation directions are also provided. It should be noted that Apertium systems are available for all reverse directions.

System	BLEU	chrF	TER
Marian ast-spa	24.0	53.5	57.4
Apertium arg-spa	61.4	77.9	26.7
Apertium oci-aran–spa	26.7	47.2	70.8
Apertium spa-ast	17.0	50.8	80.4
Apertium spa-arg	61.1	79.3	27.2
Apertium spa–oci-aran	28.8	49.4	72.3

Table 5: Evaluation figures for the systems used to create backtranslated corpora.

A number of conclusions can be drawn from Table 5. The Apertium systems for Aragonese-Spanish and Spanish-Aragonese achieve highly comparable results, as do the pairs Aranese-Spanish and Spanish-Aranese. Consequently, given that the Marian system for Asturian-Spanish attains superior outcomes compared to the reverse system Spanish-Asturian, it can be inferred that the quality of the training system is analogous to, and even surpasses, that of the Apertium systems utilised for backtranslation.

The Asturian-Spanish NMT system was employed to backtranslate the segments from the NLLB corpus that were identified as Asturian, but with translations in languages other than Spanish or with Spanish translations with SBERT scores below 0.75. The resulting backtranslated Spanish-Asturian corpus comprises 2,084,594 segments.

5.2 Basic neural Spanish-Asturian system

The basic neural system for Spanish-Asturian has been trained using the same configuration than the Asturian-Spanish for backtranslation, that is: NLLB cleaned (see table 2) and the newly created Wikipedia corpus (see table 3). A transformer configuration using SentencePiece with a vocabulary size of 32K has been used. In table 6 we can observe the evaluation figures for this system. It can be observed that this basic system attains inferior results in comparison to Apertium (see Table 5). The BLEU score is 15.3, whereas Apertium achieves 17.0.

5.3 Basic neural Spanish-Aragonese system

The basic neural system for Spanish-Aragonese has been trained using the same configuration than the Asturian-Spanish, using the following corpora: Wikimatrix cleaned (see table 2), the newly created Wikipedia corpus (see table 3, and the newly created backtranslated Wikipedia corpus (see table 3).

In table 6 we can observe the evaluation figures for this system. The evaluation results obtained by this system are significantly lower (18.8 BLEU) than the obtained by Apertium (see Table 5) (61.1 BLEU).

5.4 Multilingual system

As no parallel corpora are available for Aranese, and the available for Aragonese are very small, we experimented with multilingual systems to see whether the multilingual configuration may produce good results for these two languages, and may also improve the results obtained for Asturian.

We have trained a multilingual system from Spanish to the following languages: Aragonese, Aranese, Asturian, Catalan, Galician, French, Occitan and Portuguese. To train this system we have used the following corpora:

- **Spanish-Aragonese**: WikiMatrix cleaned (see table 2), newly created Wikipedia corpus (see table 3), Wikipedia backtranslated using Apertium (see table 3).
- **Spanish-Aranese**: Pilar backtranslated using Apertium.
- **Spanish-Asturian**: NLLB cleaned (see table 2), Wikipedia clean (see table 3) and the backtranslated corpus described in 5.1
- **Spanish-Catalan**: 10 M segments automatically selected form Paracrawl cleaned (see table 4)
- **Spanish-Galician**: all segments available in Paracrawl cleaned (see table 4)
- **Spanish-French**: 10 M segments automatically selected form MultiParacrawl cleaned (see table 4)

[19] Apertium v 3.9.4; linguistic data versions: spa-ast v.1.1.1-1; spa.arg v.0.6.0-1; es-oc v1.0.8-1

- **Spanish-Occitan**: NLLB cleaned (see table 2), newly created Wikipedia corpus (see table 3), Wikipedia backtranslated using Apertium (see table 3).
- **Spanish-Portuguese**: 10 M segments automatically selected form MultiParacrawl cleaned (see table 4)

To calculate the SentencePiece model for the multilingual system we have randomly selected 1M segments from each language pair, except for Aragonese and Aranese, for which we have used all the available segments. The final multilingual system uses an ensemble of the three best check points regarding BLEU DETOK.

Table 6 presents the evaluation figures for the multilingual system. With regard to Spanish-Asturian, the multilingual configuration exhibits an increase of 0.9 BLEU points in comparison to the basic system. However, this figure remains 0.8 BLEU points below the level achieved by Apertium. In the case of Spanish-Aragonese, the multilingual system exhibits a noteworthy enhancement of 13.3 BLEU points. Nevertheless, it remains considerably distant (29 BLEU points) from the performance of Apertium. In the case of Spanish-Aranese, the multilingual configuration yielded a BLEU score that was 8.7 points lower than that obtained with Apertium.

System	langs.	BLEU	chrF	TER
Basic	spa-ast	15.3	48.0	77.5
Basic	spa-arg	18.8	51.5	67.2
Multilingual	spa-ast	16.2	50.0	75.8
Multilingual	spa-arg	32.1	65.3	48.0
Multilingual	spa–oci-aran	20.1	44.5	77.8
Synth. val Flores	spa-ast	16.3	50.6	77.1
Synth. val Flores	spa-arg	57.2	78.1	29.4
Synth. val Flores	spa–oci-aran	26.9	48.8	72.7
Backt. val Flores	spa-ast	18.0	51.6	74.5
Apertium	spa-ast	17.0	50.8	80.4
Apertium	spa-arg	61.1	79.3	27.2
Apertium	spa–oci-aran	28.8	49.4	72.3

Table 6: Evaluation figures for the different systems trained and Apertium

5.5 Synthetic val Flores Spanish-Asturian

This systems uses the same configuration as the Basic Spanish-Aranese system, but adding the synthetic corpus from Paracrawl described in subsection 4. In this system we use the Flores dev corpus for validation. The final system uses an ensemble of the model corresponding the the 3 best checkpoints using the BLEU DETOK metric.

From table 6 we can observe that with this configuration we improve the basic system by 1 BLEU point, but we are still below Apertium.

5.6 Synthetic val Flores Spanish-Aragonese

This systems uses the same configuration as the Basic Spanish-Aragonese system, but adding the synthetic corpus from Paracrawl described in subsection 4. In this system we use the Flores dev corpus for validation. The final system uses an ensemble of the model corresponding the the 3 best checkpoints using the BLEU DETOK metric.

This configuration achieves an impressive improvement of 38.4 BLEU points in comparison with the basic configuration (see table 6), but Apertium keeps an advantage of 3.9 BLEU points. At this point, we can try to explain two key observations: why does Apertium achieve such strong evaluation metrics for Spanish-Aragonese? And why does the system trained with synthetic corpora created by Apertium show such a remarkable improvement? The answer may lie in how the Flores corpus was developed for this language pair. Since it was generated through machine translation from the Spanish Flores using Apertium, the system has a clear advantage when calculating automatic evaluation metrics.

This is the final submission for the Open systems for Spanish-Asturian, with id 568.

5.7 Synthetic val Flores Spanish-Aranese

This system have been trained using the backtranslated Pilar corpus and the synthetic corpus from Paracrawl described in subsection 4. For the validation set we have used the Flores dev corpus. The final system uses an ensemble of the model corresponding the the 3 best checkpoints using the BLEU DETOK metric.

This system achieves an improvement of 6.8 BLEU points in comparison with the multilingual configuration (see table 6), but it is still 1.9 BLEU points below Apertium.

This is the final submission for the Open systems for Spanish-Aranese, with id 610.

5.8 Backtranslation val Flores Spanish-Asturian

We have followed the same configuration than the Basic Spanish-Asturian, but using also the backtranslated Wikipedia corpus. For the validation set we have used the Flores dev corpus. The final system uses an ensemble of the models corresponding

the the 3 best checkpoints using the BLEU DETOK metric.

This system achieves an improvement of 2.7 BLEU points in comparison with the basic system, and 1.7 BLEU points in comparison with the synthetic one (see table 6), and also outperforms Apertium by 1 BLEU point.

This is the final submission for the Open systems for Spanish-Asturian, with id 568.

6 Final submissions

Table 7 contains a comprehensive overview of the systems that have been submitted to the shared task. With regard to the Spanish-Asturian task, an open system has been submitted, whereas for Spanish-Aragonese and Spanish-Aranese, constrained systems have been submitted.

7 Energy consumption report

The training scripts generate a log with the timestamp and the GPU consumption every second throughout the entirety of the training process. This enables the calculation of the total training time and the approximate energy consumption in kWh. The total training time and the consumption of each of the two GPU units utilized, along with the total consumption, are presented in Table 8. The final submitted systems are highlighted in bold.

As we can see, all the systems are trained in short times (from 1 h. 10 m. to 3 h. 23 m.) in a relativelly modest computer wit two NVIDIA RTX A5000 GPU units with 24 GB each, a AMD Ryzen Threadripper PRO 3945WX CPU with 12-Cores and 64 GB of RAM. As the training times are short, the energy consumption is very low in all trainings, ranging from 0.481 to 1.292 kWh.

8 Conclusions and future work

In this paper we have presented the systems that the TAN-IBE team have submitted to the WMT24 Shared Task Translation into Low-Resource Languages of Spain. We have presented an open system for the Spanish-Asturian language pair, and constrained systems for Spanish-Aragonese and Spanish-Aranese.

The primary challenge in completing the task was the unavailability of high-quality parallel corpora for the specified language pairs. Fortunately, all the language pairs in question have an Apertium system, and Apertium is also available for all the reverse language pairs except Asturian-Spanish. This enabled us to conduct experiments with synthetic and backtranslated corpora. To perform backtranslation experiments for Spanish-Asturian, we trained a basic neural Asturian-Spanish system.

Additionally, monolingual and parallel corpora were generated from Wikipedia dumps for Spanish to Asturian, Aragonese and Occitan (given the unavailability of an Aranese Wikipedia).

Regarding the training strategies, we experimented with bilingual and multilingual systems. While multilingual systems demonstrated enhanced performance relative to basic systems, the use of synthetic and backtranslated corpora yielded superior outcomes.

In the period preceding the conclusion of the TAN-IBE project in July 2025, it is our intention to undertake the following actions:

- During the course of the project, a corpus of monolingual and bilingual texts in Asturian, Aragonese, Aranese and Spanish has been compiled. The next stage is to process and align these texts in order to increase the number of available parallel segments.
- We plan to train new systems using the parallel corpora and the training techniques presented in this paper. The quality of the resulting systems will then be evaluated in order to ascertain whether the inclusion of the new parallel texts has had a positive impact.
- Furthermore, we intend to learn from the other participants in the shared task and attempt to reproduce the training techniques that have yielded the most favourable outcomes, utilising the newly created parallel corpora.
- We plan to develop neural systems for Spanish to the other languages of the TAN-IBE project, namely, Portuguese, Galician, Asturian, Argonese, Catalan and Aranese. These systems will be freely released.
- We also plan to train a multilingual system able to translate to and from all the languages of the TAN-IBE project.

After the completion of the TAN-IBE project, we plan to increase the size of the Apertium monolingual and transfer dictionaries for Spanish to Asturian, Aragonese and Aranese using automatic techniques that make use of monolingual and parallel corpora. The quality of the Apertium systems is noteworthy, and the enhancement of the dictionaries has the potential to further optimise the efficiency of the Apertium systems for the generation

Submission	Type	Section	langs.	BLEU	chrF	TER
#568	Open	5.8	spa-ast	18.0	51.6	74.5
#584	Constrained	5.6	spa-arg	57.2	78.1	29.4
#610	Constrained	5.7	spa–oci-aran	26.9	48.8	72.7

Table 7: Information about the systems submitted to the shared task.

System	langs.	Section	Time.	GPU0 (kWh)	GPU1 (kWh)	Total (kWh)
Backtranslation	ast-spa	5.1	2 h. 51 m.	0.594	0.598	1.191
Basic	spa-ast	5.2	3 h. 23 m.	0.704	0.709	1.414
Basic	spa-arg	5.3	1 h. 38 m.	0.332	0.339	0.671
Multilingual	spa-MULT	5.4	3 h. 7 m.	0.627	0.643	1.270
Synthetic	spa-ast	5.5	3 h. 13 m.	0.645	0.650	1.292
Synthetic	**spa-arg**	**5.6**	**1 h. 31 m.**	**0.302**	**0.305**	**0.606**
Synthetic	**spa–oci-aran**	**5.7**	**2 h. 53 m.**	**0.577**	**0.583**	**1.160**
Backtranslation	**spa-ast**	**5.8**	**1 h. 10 m.**	**0.240**	**0.241**	**0.481**

Table 8: Total time and energy consumption for all the trainings.

of synthetic and backtranslated corpora.

Acknowledgements

This work is supported by the project *TAN-IBE: Neural Machine Translation for the romance languages of the Iberian Peninsula*, project PID2021-124663OB-I00 founded by MCIN /AEI /10.13039/501100011033 / FEDER, UE.

References

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.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Tanmai Khanna, Jonathan N Washington, Francis M Tyers, Sevilay Bayatlı, Daniel G Swanson, Tommi A Pirinen, Irene Tang, and Hector Alos 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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Antoni Oliver. 2020. Traducción automática para las lenguas románicas de la península ibérica. *Studia Romanica et Anglica Zagrabiensia: Revue publiée par les Sections romane, italienne et anglaise de la Faculté des Lettres de l'Université de Zagreb*, 65:367–375.

Antoni Oliver and Sergi Alvarez. 2023. Training and integration of neural machine translation with MTUOC. In *Proceedings of the 1st Workshop on Open Community-Driven Machine Translation*, pages 5–13, Tampere, Finland. European Association for Machine Translation.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Appendix - Metric signatures

- BLEU nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.3.1
- chrF2 nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:yes | nc:6 | nw:0 | space:no | version:2.3.1
- TER nrefs:1 | bs:1000 | seed:12345 | case:lc | tok:tercom | norm:no | punct:yes | asian:no | version:2.3.1