Universitat d'Alacant's Submission to the WMT 2024 Shared Task on Translation into Low-Resource Languages of Spain

Aarón Galiano-Jiménez,[†] **Víctor M. Sánchez-Cartagena,**[†]
Juan Antonio Pérez-Ortiz,[*†] **Felipe Sánchez-Martínez**[†]

[†]Dep. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain

[*]Valencian Graduate School and Research Network of Artificial Intelligence, ValgrAI

{aaron.galiano,vm.sanchez,japerez,fsanchez}@ua.es

Abstract

This paper describes the submissions of the Transducens group of the Universitat d'Alacant to the WMT 2024 Shared Task on Translation into Low-Resource Languages of Spain; in particular, the task focuses on the translation from Spanish into Aragonese, Aranese and Asturian. Our submissions use parallel and monolingual data to fine-tune the NLLB-1.3B model and to investigate the effectiveness of synthetic corpora and transfer-learning between related languages such as Catalan, Galician and Valencian. We also present a many-to-many multilingual neural machine translation model focused on the Romance languages of Spain.

1 Introduction

Spain is home to several languages, each with different levels of representation in neural machine translation (NMT) technologies and availability of training data. For example, Spanish (spa) has abundant data resources and is included in many multilingual translation models and large language models (LLM). Other languages, such as Catalan (cat) and Galician (glg), are relatively well-supported and have enough data to train NMT models from scratch. However, languages such as Asturian (ast), Aragonese (arg) and Aranese (arn) face significant challenges due to the limited availability of data needed to train these systems.

Despite these challenges, the linguistic similarity between some of these languages simplifies their integration into multilingual translation models, allowing them to benefit from transfer-learning from more widely represented languages; an example of this is the inclusion of Asturian in NLLB-200 (NLLB Team et al., 2022). In addition, shallow-transfer rule-based machine translation (MT) systems such as Apertium (Forcada et al., 2011), exist for some of these languages, including Spanish–Asturian[1], Spanish–Aragonese[2] and Occitan–Spanish[3]; the development of rule-based systems does not require large amounts of training data, but linguistic knowledge to construct dictionaries and translation rules.

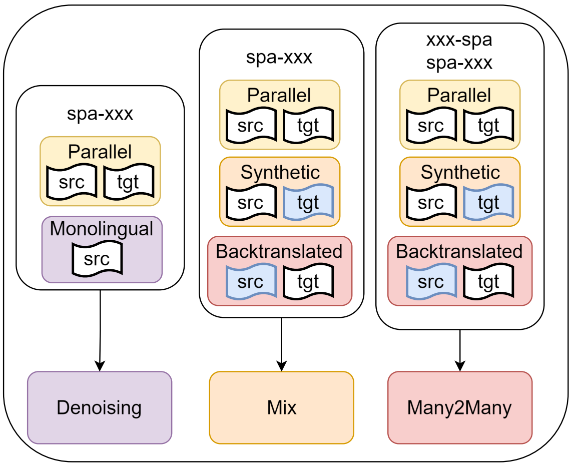

Figure 1: Submitted models for the shared task. Blue in *src* or *tgt* indicates text generated by MT. The Denoising and Mix models were trained for a single translation direction, whereas the Many2Many model was trained with multiple language pairs and in both translation directions for each pair. xxx represents any of the target languages: Aragonese, Aranese and Asturian.

Our approach to developing NMT systems for the WMT 2024 Shared Task on Translation into Low-Resource Languages of Spain (Sánchez-Martínez et al., 2024) for Aragonese, Aranese and Asturian is based on pre-trained models that include similar languages, such as NLLB-200, to incorporate languages that were not originally seen during training. Given the scarcity of corpora for Asturian, Aragonese and Aranese, we used Apertium to generate synthetic corpora. This involved translating monolingual corpora into Spanish and vice versa to generate additional resources for fine-

[1] https://github.com/apertium/apertium-spa-ast
[2] https://github.com/apertium/apertium-spa-arg
[3] https://github.com/apertium/apertium-oci-spa

Proceedings of the Ninth Conference on Machine Translation, pages 885–891
November 15-16, 2024 ©2024 Association for Computational Linguistics

tuning NLLB-200.

By combining pre-trained multilingual models, synthetic data generation and rule-based translation systems, we aim to improve the quality and accessibility of NMT for these low-resource languages.

All of our submissions, shown in Figure 1, are classified as *open* because they are based on the NLLB-200 model with 1.3B parameters, exceeding the 1B parameters limit for *constrained* submissions. However, we only used the corpora allowed for constrained submissions, as described in Section 3.

2 State-of-the-Art Methods Used

Our submission makes use of well-established techniques such as denoising pre-training (Lewis et al., 2020), transfer-learning (Zoph et al., 2016), back-translation (Sennrich et al., 2016) and sequence-level knowledge distillation (Kim and Rush, 2016). The languages involved in this shared task are related, making knowledge transfer by a multilingual model an effective solution. This transfer can be achieved by bilingual fine-tuning of a pre-trained model (Zoph et al., 2016), using the knowledge it already possesses, or by multilingual training from scratch (Bommasani et al., 2021), using multiple languages simultaneously during the training process. In our approach, we fine-tune the NLLB-200 model and investigate the effects of training with only one specific translation direction with different types of data, as well as adding more languages during training.

A common technique is to pretrain a system on monolingual corpora with the denoising task[4] to learn language generation, and then train the system on parallel bilingual corpora for translation (Lewis et al., 2020). It is well known that combining both tasks simultaneously improves the translation results (Kamboj et al., 2022). Since NLLB-200 was trained in this way (NLLB Team et al., 2022), it is reasonable to use the same technique for fine-tuning to leverage the available monolingual corpus.

Another method of exploiting monolingual corpora is to create synthetic parallel corpora. For this purpose, we used the rule-based MT systems built using the Apertium platform (Forcada et al., 2011). Specifically, we used Catalan–Spanish, Aragonese–Spanish, Aranese–Spanish, Spanish–Aragonese, Spanish–Aranese and Spanish–Asturian systems. Translating from source (spa) to target and using this synthetic corpus as a target for training is a type of sequence-level knowledge distillation (Lai et al., 2021; Yu et al., 2021). In contrast, translating from the target language and using this synthetic corpus as the source for training is called back-translation (Sennrich et al., 2016). The latter has the advantage that potential translation errors in the synthetic corpus do not affect the generation of the target language, as the synthetic corpus is used as input during training rather than as the desired output.

Corpus	Sentences	Src words	Tgt words
spa-arg	33,723	3,706,154	3,589,002
spa-arn	85,491	14,720,677	14,266,772
spa-ast	45,506	6,844,424	6,663,424
spa	500,000	62,004,331	—
arg	24,675	2,718,855	—
arn	229,886	29,110,670	—
ast	38,868	5,504,371	—
spa-cat	559,805	91,543,160	88,057,754
spa-glg	184,861	30,716,538	28,753,332
spa-val	287,403	52,836,299	53,137,411

Table 1: Number of sentences and words in each of the corpora used.

3 Data

We used only the corpora allowed for the constrained submissions: Opus[5] and PILAR (Galiano-Jiménez et al., 2024)[6]. For the development set, we used the FLORES+[7] (NLLB Team et al., 2022) dev versions (997 sentences) for Spanish (spa), Aranese (arn), Aragonese (arg) and Asturian (ast) (Pérez-Ortiz et al., 2024). The specific details of the corpora used are described below and shown in Table 1.

3.1 Parallel Corpora

Aragonese and Asturian: We used parallel corpora with Spanish available in OPUS,[8] consisting of 33,723 Spanish–Aragonese sentences and 45,506 Spanish–Asturian sentences.

[4] The denoising task is a self-supervised learning strategy that helps models learn effective representations from monolingual data by training them to restore original sentences from corrupted inputs.

[5] https://opus.nlpl.eu/

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

[7] https://github.com/openlanguagedata/flores

[8] https://opus.nlpl.eu/

Aranese: We used the parallel Catalan-Aranese corpus available in PILAR (85,491 sentences) and translated the Catalan part into Spanish using Apertium. This corpus was not included in the synthetic data description in section 3.3, as it was not automatically translated from or into Aranese.

3.2 Monolingual Corpora

We used the monolingual corpora available in PILAR, which contains 24,675 sentences in Aragonese, 229,886 in Aranese and 38,868 in Asturian.

3.3 Synthetic Corpora Generation

We used Apertium to generate different synthetic corpora. By translating the above monolingual corpora into Spanish, we created two corpora (there is no Asturian–Spanish system for Apertium) that were used for back-translation. Since these corpus sizes are relatively small for training NMT models, we also translated 500,000 Spanish sentences extracted from Paracrawl[9] into Aranese[10], Aragonese[11] and Asturian[12]. This resulted in synthetic corpora with the target language as the synthetic part.

3.4 Additional data

For our multilingual system, we used all the above corpora and added Wikimedia Spanish–Catalan (559,805 sentences) and Spanish–Galician (184,861 sentences) corpora from OPUS, as well as the Spanish–Valencian (val) corpus available in PILAR, making 287,403 sentences.

4 Methodology

We trained several models for each language pair to analyse the effects of different types of corpora and transfer-learning between different tasks and languages. Below we describe the methodology we used and the different systems we trained.

4.1 Model Architecture and Baseline

All our models are based on a fine-tuning of the NLLB-200 (NLLB Team et al., 2022) model with 1.3 billion parameters. We chose this model because it is a transformer (Vaswani et al., 2017) pretrained with 200 languages and specialised in translation tasks. These 200 languages include Spanish and Asturian, which correspond to one of the translation directions in this shared task, as well as other related languages, such as Catalan, Galician and Occitan.

As a baseline, we used Apertium to compare the effectiveness of a rule-based system with a neural-based system.

4.2 Training Approaches

In this section, we describe the different approaches we used to train the translation models for each language pair.

Bilingual parallel: For each translation direction, we trained a specific model using only the parallel corpus available for that language pair. These models correspond to the `Parallel` row in tables 2 and 3.

Bilingual Parallel + Monolingual: We trained a model for each translation direction by combining the translation task using the parallel corpus with a denoising task using monolingual data of the target language. This approach helps to improve translation quality by exploiting additional monolingual resources. These models correspond to the `Denoising` row in tables 2 and 3.

Bilingual Synthetic Generated with Apertium: For each translation direction, we trained a model using only synthetic parallel corpora generated by translating the Spanish monolingual corpora into the target language using Apertium. These models correspond to the `Synthetic` row in tables 2 and 3.

Bilingual Parallel with Synthetic and Back-translation: For each translation direction, we trained a model using a combination of parallel corpora, synthetic corpora generated with Apertium, and back-translation. For back-translation, we used Apertium to translate the monolingual target language data into Spanish. These models correspond to the `Mix` row in tables 2 and 3.

Multilingual Parallel with Synthetic and Back-translation: This model extends the previous approach by training a single model on all three translation directions. This multilingual training allows

[9] https://opus.nlpl.eu/ParaCrawl/es&ca/v9/ParaCrawl

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

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

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

the model to benefit from common linguistic features across languages. This model corresponds to the `Multilingual` row in tables 2 and 3.

Multilingual Many-to-Many: This model extends the multilingual approach by incorporating parallel corpora from related languages and including both translation directions for each language pair. This model can translate between Spanish, Aragonese, Aranese, Asturian, Catalan, Galician and Valencian. This model corresponds to the `Many2Many` row in tables 2 and 3.

4.3 Evaluation Metrics

To measure the quality of the translation models, we used BLEU (Papineni et al., 2002)[13] and chrF2 (Popović, 2015)[14] scores on the translation of the development set. For the NLLB-200 based models, we translated using beam search (Graves, 2012) with a beam size of 5.

5 Experiments and Results

In this section, we present the experimental setup, including hyperparameters and training configurations. We compare the performance of each system against the baseline provided by Apertium and discuss the results for each translation direction.

5.1 Experimental Setup

All our translation models were trained using the Transformers (Wolf et al., 2020) library on an Nvidia A100 GPU with 40 GB of memory. We extended the library to support simultaneous training with multiple tasks, including different datasets with different languages and combining translation and denoising tasks[15]. To balance the training data from datasets of different sizes, we used temperature upsampling with $1/T = 0.3$, following the approach used in NLLB-200 training (NLLB Team et al., 2022).

Due to GPU memory constraints, we used a batch size of 16 and accumulated gradients as many times as the number of different datasets used in the training to ensure that all tasks were seen before updating the model weights. The learning rate was set to 5e-5, and we used the AdamW optimizer (Loshchilov and Hutter, 2017) with β_1=0.9, β_2=0.999 and ε=10^{-8}. Models were trained for a maximum of 100 epochs with early stopping, and evaluations were performed every 1000 training steps. The stopping criterion was based on the BLEU score on the development set, with a patience of 6 evaluations.

Model	**spa-arg**	**spa-arn**	**spa-ast**
Apertium	66.0	38.0	17.1
Parallel	41.4	34.4	17.9
Denoising*	41.6	35.7	17.8
Synthetic	65.3	37.6	17.0
Mix*	65.1	37.8	17.0
Multilingual	64.8	37.5	17.0
Many2Many*	65.2	37.9	17.0

Table 2: BLEU scores on the FLORES+ dev. Models marked with an asterisk are those we submitted for the Shared Task.

We used the NllbTokenizer[16] class for corpus segmentation. This tokenizer uses the SentencePiece (Kudo and Richardson, 2018) model used by NLLB-200 and applies language tokens to both the source and target texts.

NLLB-200 uses language tokens in both the source and target sentences. When training with a new language, it is possible to use the language token of a similar language, but this eliminates the possibility of translating to or from the language of the original token. In our case, we added new language tokens for Aragonese, Aranese and Valencian. To avoid learning the embeddings for these tokens from scratch, we initialised them with the embeddings of the most similar languages included in NLLB-200. Specifically, we initialized the embeddings for Aragonese and Valencian with the Catalan embedding, and the embedding for Aranese with the Occitan embedding[15].

5.2 Results

Tables 2 and 3 show the results of the translation models in terms of BLEU and chrF scores, respectively. The first row corresponds to the Apertium baseline and the remaining rows show the results of each trained model.

For the shared task, we submitted specific Denoising and Mix models for each translation direction, and the Many2Many model for all three

[13] SacreBLEU BLEU signature: `nrefs:1 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.0.0`

[14] SacreBLEU chrF2 signature: `nrefs:1 | case:mixed | eff:yes | nc:6 | nw:0 | space:no | version:2.0.0`

[15] The code is available at https://github.com/transducens/transformers-multilingual-training

[16] https://github.com/huggingface/transformers/blob/v4.42.0/src/transformers/models/nllb/tokenization_nllb.py

Model	spa-arg	spa-arn	spa-ast
Apertium	82.2	60.0	50.7
Parallel	70.8	57.9	50.8
Denoising*	69.5	58.6	50.7
Synthetic	81.9	59.9	50.7
Mix*	81.8	59.9	50.8
Multilingual	81.8	59.8	50.8
Many2Many*	81.9	60.0	50.8

Table 3: chrF2 scores on the FLORES+ dev. Models marked with an asterisk are those we submitted for the Shared Task on Translation into Low-Resource Languages of Spain.

Model	spa-arg	spa-arn	spa-ast
Denoising	37.8	27.0	17.4
Mix	60.2	28.5	16.9
Many2Many	59.8	28.5	16.8

Table 4: BLEU scores on the WMT 2024 Shared Task on Translation into Low-Resource Languages of Spain test.

directions [17]. This allows us to compare a model trained only on the available corpus for each language, another that incorporates a synthetic corpus, and one that includes multiple translation directions and additional languages. The results on the test set of the task, which correspond to the FLORES+ devtest versions of these languages, are shown in Tables 4 and 5.

When analysing the results, it is important to consider how the development sets were created (Pérez-Ortiz et al., 2024). The Asturian sentences were first professionally translated from English by Meta (NLLB Team et al., 2022) and then revised by academics. In contrast, the Aragonese and Aranese sentences were first machine translated from Spanish using Apertium, then manually edited by language specialists and finally reviewed by academics. This means that the development sets for Aragonese and Aranese may be biased towards the results produced by Apertium.

We conducted paired significance tests to determine whether the submitted models outputs were significantly different despite the similarity of some results. Specifically, we calculated paired approximate randomisation (Riezler and Maxwell, 2005) as implemented by SacreBLEU on the devtest using BLEU and chrF2. The results indicated that the differences between the Mix and Many2Many models for Asturian and Aranese were not statistically significant, whereas the differences between all the other pairs of models were statistically significant.

Model	spa-arg	spa-arn	spa-ast
Denoising	67.5	48.3	50.7
Mix	78.9	49.3	50.9
Many2Many	78.8	49.3	50.9

Table 5: chrF2 scores on the WMT 2024 Shared Task on Translation into Low-Resource Languages of Spain test.

Effect of adding a monolingual corpus: The results show a minimal difference when adding a monolingual corpus compared to training only with parallel corpora (rows 2 and 3 of Tables 2 and 3). However, this could be due to the amount of training data available. The improvement in the quality of the translations for the Aranese direction is particularly remarkable, since it is the largest monolingual corpus.

Effect of using the synthetic corpus produced by Apertium: The increase in performace for the synthetic models compared to the Denoising models for Aragonese and Aranese can be explained both by the difference in data volume and by the bias of the development sets. Conversely, there is a decrease in the results for Asturian, suggesting a bias in the other sets.

The combination of parallel and synthetic corpora in both target and source (Mix models) shows minimal variation in the results. Again, this may be due to the difference in the proportion of the corpus generated by the spa-xxx translation and that generated by the xxx-spa translation.

Effect of multilingual training: Combining multiple translation directions in the same training session complicates the learning task for the model, but also increases the amount of data available. Adding more languages slightly improves the results compared to training with only the languages of the common task.

6 Conclusions

Overall, the results highlight the critical role of training data volume in the development of effective NMT models. The challenge with large neural models lies in the insufficient amount of training

[17] The Many2Many model is available at https://huggingface.co/Transducens/IbRo-nllb

data available for low-resource languages, which limits the full potential of these architectures.

However, rule-based systems remain a viable option for these languages, although they require linguistic expertise to build. The use of these systems to generate synthetic corpora is proving beneficial in integrating low-resource languages into neural translation models and exploiting the advantages they offer.

Acknowledgments

This paper is part of the R+D+i project PID2021-127999NB-I00 funded by the Spanish Ministry of Science and Innovation (MCIN), the Spanish Research Agency (AEI/10.13039/501100011033) and the European Regional Development Fund A way to make Europe. The computational resources used were funded by the European Regional Development Fund through project IDIFEDER/2020/003.

References

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez Felipe Sánchez-Martínez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144. Special Issue: Free/Open-Source Machine Translation.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024. Pan-iberian language archival resource.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711.

Samta Kamboj, Sunil Kumar Sahu, and Neha Sengupta. 2022. DENTRA: Denoising and translation pre-training for multilingual machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1057–1067, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

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.

Wen Lai, Jindřich Libovický, and Alexander Fraser. 2021. The LMU Munich system for the WMT 2021 large-scale multilingual machine translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 412–417, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

NLLB Team, 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. *arXiv*, abs/2207.04672.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Aarón Galiano-Jiménez, Antoni Oliver, Claudi Aventín-Boya, Cristina Valdés, Alejandro Pardos, and Juan Pablo Martínez. 2024. FLORES+ datasets for Aragonese, Aranese, Asturian and Valencian. In *Proceedings of the Ninth Conference on Machine Translation*, Miami. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz, Aarón Galiano-Jiménez, and Antoni Oliver. 2024. Findings of the WMT 2024 Shared Task on Translating into Low-Resource Languages of Spain: Blending rule-based and neural systems. In *Proceedings of the Ninth Conference on Machine Translation (WMT24)*, Miami, Florida, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

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.

Zhengzhe Yu, Daimeng Wei, Zongyao Li, Hengchao Shang, Xiaoyu Chen, Zhanglin Wu, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, Hao Yang, and Ying Qin. 2021. HW-TSC's participation in the WMT 2021 large-scale multilingual translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 456–463, Online. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.