Low-Resource Machine Translation for Moroccan Arabic

Alexei Rosca **Abderrahmane Issam** **Gerasimos Spanakis**
Department of Advanced Computing Sciences
Maastricht University
{alexei.rosca, abderrahmane.issam, jerry.spanakis}@maastrichtuniversity.nl

Abstract

Neural Machine Translation (NMT) has achieved significant progress especially for languages with large amounts of data (referred to as high resource languages). However, most of the world languages lack sufficient data and are thus considered as low resource or endangered. Previous research explored various techniques for improving NMT performance on low resource languages, with no guarantees that they will perform similarly on other languages. In this work, we explore various low resource NMT techniques for improving performance on Moroccan Arabic (Darija), a dialect of Arabic that is considered a low resource language. We experiment with three techniques that are prominent in low resource Natural Language Processing (NLP), namely: back-translation, paraphrasing and transfer learning. Our results indicate that transfer learning, especially in combination with back-translation is effective at improving translation performance on Moroccan Arabic, achieving a BLEU score of 26.79 on Darija→English and 9.98 on English→Darija. [1]

1 Introduction

Neural Machine translation (NMT) has achieved impressive results for high resource languages, supported by extensive linguistic data and resources that facilitate model training and optimization (Johnson et al., 2017; Vaswani et al., 2017). However, due to limited data availability and inherent linguistic complexity, significant performance gaps remain for low-resource languages (Lakew et al., 2020). This work addresses this critical gap by developing MT solutions for Moroccan Arabic (commonly referred to as Darija), a dialect characterized by unique linguistic features and a lack of training data. By focusing on Moroccan Arabic, this work seeks to advance natural language processing (i.e. NLP) methods for underrepresented languages and contribute to a more inclusive landscape in machine translation.

The main objective of this paper is build a Darija-English translation model using state-of-the-art low resource NMT techniques. We experiment with three techniques that have been shown to be effective in low resource scenarios (Haddow et al., 2022), namely: back-translation, paraphrasing and transfer learning. More specificaly, we apply back-translation by training a model to translate from the target to the source language and using it to generate translations. We then translate monolingual target sentences to the source language then use them for training. For paraphrasing, we use a paraphrasing model to generate synthetic copies of the input sentence and use them to augment the training set. In combination with paraphrasing and back-translation, we experiment with transfer learning from publicly available pretrained models. We fine-tune a multilingual model that supports Moroccan Arabic, and given that Arabic is the mother language of Moroccan Arabic, we also fine-tune a bilingual model that is trained for Arabic-English translation. Furthermore, we compose an encoder-decoder translation model from the checkpoints of a BERT model pretrained on Moroccan Arabic and an Arabic to English translation model and we fine-tune it for Darija→English translation.

Our results demonstrate the effectiveness of transfer learning on training an NMT model for Moroccan Arabic, especially when combined with back-translation, achieving a BLEU score of 26.79 on Darija→English and 9.98 on English→Darija. However, when evaluating on datasets from a different domain, we find the improvements are unstable, which questions the generalization of these techniques to other domains. Furthermore, we observe a significant disparity between translation directions: translating into English from Darija achieves more than twice the performance of translating

[1] https://github.com/RoscaAlex00/lowresource_mt

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 32–38, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-005

from English→Darija. This calls for further research into democratizing NLP for low resource languages such as Moroccan Arabic.

2 Related Works

Moroccan Arabic NLP: NLP resources for Moroccan Arabic are characterized by scarcity. Previous work introduced resources for different NLP tasks (Samih and Maier, 2016; Issam and Mrini, 2021; Boujou et al., 2021; Moussa and Mourhir, 2023). For machine translation, Tachicart et al. (2014) introduce the Moroccan Dialect Electronic Dictionary (MDED): a bilingual dictionary for Moroccan and Modern Standard Arabic (MSA). Mrini and Bond (2017) introduce Moroccan Darija WordNet (MDW): an extension of the Open Multilingual WordNet (Bond and Foster, 2013) to Moroccan Arabic. Unfortunately, these works are limited to word level translations. Outchakoucht and Es-Samaali (2021) introduce Darija Open Dataset (DODa), a collaborative dataset of word and sentence level translations between Darija and English. In this work, we leverage this dataset for training and evaluating machine translation models.

Low Resource NMT: Low resource NMT is an active area of research with real-world impact. Various techniques were introduced to deal with data and resource scarcity and were shown to be effective (Haddow et al., 2022). The most straightforward technique is data augmentation (Feng et al., 2021), where rule based or neural based techniques can be used to generate more data that can be used for training. In this work, we study two successful data augmentation techniques, namely, back-translation (Sennrich et al., 2016, 2017; Hoang et al., 2018; Edunov et al., 2018) and paraphrasing (Callison-Burch et al., 2006; Wang et al., 2016; Mallinson et al., 2017). Back-translation translation target sentences to the source language using an available model, while paraphrasing creates synthetic copies of the source sentences. Furthermore, previous work has shown the effectiveness of transfer learning especially in low resource scenarios (Zoph et al., 2016; Howard and Ruder, 2018; Devlin et al., 2019), where a model is pretrained on large amounts of data and fine-tuned for a target task or domain. We similarly leverage pretrained models in combination with data augmentation and evaluate their performance when fine-tuned on Moroccan Arabic.

3 Methodology

3.1 Back-Translation

Back-translation is a frequently used technique for data augmentation in machine translation (Sennrich et al., 2016, 2017; Hoang et al., 2018; Edunov et al., 2018). It helps to overcome the lack of parallel corpora, particularly for languages with limited resources such as Moroccan Arabic. This method takes advantage of the greater availability of the target language (i.e. English) to generate new synthetic sentence pairs. The process begins with the training of a target-to-source model (i.e. English-to-Moroccan Arabic), which is used to translate the target language text into the source language text. This newly generated dataset is then included in the training process of the source-to-target translation model.

3.2 Paraphrasing

Similar to back-translation, paraphrasing aims at augmenting the training data and diversifying the linguistic structure and vocabulary of the input sentences (Callison-Burch et al., 2006; Wang et al., 2016), while preserving their original meaning. Paraphrasing can be achieved either using rule based techniques such as synonym replacement or using neural models. Previous work shows that neural based models generate better paraphrases (Mallinson et al., 2017). In this work, we generate paraphrases of the source sentences using BART-paraphrase [2] which is a BART model (Lewis et al., 2020) fine-tuned for paraphrasing. We use this model to paraphrase the English sentences either on source or the target side depending on the translation direction (i.e. English→Darija or Darija→English). A single paraphrase for each example is added to the dataset.

3.3 Transfer Learning

Transfer learning leverages knowledge learned during pretraining to improve performance on closely related tasks. It has been particularly effective in low resource scenarios, since the pretrained model is often trained on larger amount of data than is available in the target task. We similarly apply transfer learning by fine-tuning models that are trained on massive amounts of data on translating Moroccan Arabic.

[2] https://huggingface.co/eugenesiow/bart-paraphrase

4 Experiments

4.1 Datasets

For training and evaluation, we use the Darija Open Dataset (i.e. DODa) (Outchakoucht and Es-Samaali, 2021, 2024). It is of the largest dataset for Darija-English translation with more than 45000 translation pairs. We split this dataset randomly into a training, test and validation set in a ratio of 80%, 15% and 5% respectively.

To assess the generalization of our experiments, we evaluate the models on two different datasets that contain Darija-English translation pairs. Specifically, we use translations from the New Testament that were collected for Moroccan Arabic (Sajjad et al., 2020) which we refer to as BIBLE, and MADAR (Bouamor et al., 2018), which is a dataset that contains translations between English and 26 Arabic dialects including Moroccan Arabic. These two test sets contain 500 and 5500 examples respectively.

For both backtranslation and paraphrasing, we generate one copy of the training set and include it in the training.

4.2 Models

No Language Left Behind (NLLB) (Team et al., 2022): NLLB is a massively multilingual model that supports more than 200 languages including Moroccan Arabic. We fine-tune this model on Darija→English and English→Darija. We use the small distilled version of NLLB [3] for both directions.

OPUS-MT (Tiedemann and Thottingal, 2020): Is an open source initiative that has released a collection of resources and models. Although there is no OPUS-MT translation model that supports Moroccan Arabic, there are models that support Arabic. Since Arabic is the mother language of Darija, we experiment with fine-tuning OPUS-MT English to Arabic (i.e. OPUS-MT-En-Ar [4]) and Arabic to English (i.e. OPUS-MT-Ar-En [5]) models on translating between English and Darija.

Encoder-decoder checkpointing (Rothe et al., 2020): We experiment with composing a translation encoder-decoder from different pretrained encoder and decoder checkpoints . Specifically, we use a BERT (Devlin et al., 2019) model that is pretrained for Moroccan Arabic (i.e. DarijaBERT [6] (Gaanoun et al., 2023)) to initialize the encoder, and the decoder part of OPUS-MT Arabic to English to initialize the decoder. Although the two models are different, previous research (Rothe et al., 2020) has shown that this can be effective for fine-tuning.

For training and evaluation details, please see Appendix A.

5 Results and Discussion

5.1 Main Results

Table 1 and 2 show the BLEU score results of our experiments on Darija→English and English→Darija respectively on DODa, BIBLE and MADAR test sets (we include chrF score results in Appendix B). We first notice a significant disparity between the performance on Darija→English versus English→Darija, especially of NLLB, which is significantly better at translating to English than to Darija. Furthermore, the results show that fine-tuning consistently improves performance over the base model especially on the in-domain DODa test set. However, when looking at BIBLE and MADAR test sets, we notice that fine-tuning negatively affects the performance of NLLB on Darija→English translation.

Back-translation leads to better results than paraphrasing especially when translating Darija→English. Paraphrasing is even worse than fine-tuning on this direction. In the English-to-Darija direction, paraphrasing consistently outperforms fine-tuning. We attribute this disparity to the direction of paraphrasing. In the case of Darija→English translation, paraphrasing is applied to the target sentences, which can negatively impact model outputs. In contrast, paraphrasing the source sentences, as in English→Darija, tends to be more robust and beneficial.

BERT-OPUS achieves the best results on DODa Darija→English translation, with a slight improvement over NLLB. This is still significant given that BERT-OPUS is smaller than NLLB (i.e. 200M parameters in BERT-OPUS compared to 600M parameters in NLLB), and is trained on significantly less data. This shows the advantage of training language specific models, where DarijaBERT is pretrained on Darija sentences mined from the web.

Although NLLB supports Darija, the results of translating English→Darija are very low, even lower than 1 BLEU point on DODa and BIBLE test

[3] https://huggingface.co/facebook/nllb-200-distilled-600M
[4] https://huggingface.co/Helsinki-NLP/opus-mt-en-ar
[5] https://huggingface.co/Helsinki-NLP/opus-mt-ar-en
[6] https://huggingface.co/SI2M-Lab/DarijaBERT

Model	Dataset	Base	FT	Para	BT
NLLB	DODa	8.66	26.50	20.57	**26.65**
	BIBLE	**20.43**	13.53	11.84	13.48
	MADAR	**29.31**	27.44	28.27	28.19
OPUS-MT	DODa	2.03	14.39	10.52	**15.89**
	BIBLE	4.05	4.30	4.42	**4.80**
	MADAR	7.03	13.81	**16.56**	15.32
BERT-OPUS	DODa	0.00	26.78	20.54	**26.79**
	BIBLE	0.05	1.94	1.46	**2.32**
	MADAR	0.01	15.53	15.87	**17.33**

Table 1: We provide the BLEU score results on Darija→English. *Base* shows the results of the pretrained model, *FT* the results after fine-tuning, *Para* the results of fine-tuning on paraphrased dataset, and *BT* shows the results of fine-tuning on the dataset with back-translated data.

sets. Fine-tuning is effective especially on DODa test set, increasing BLEU score by more than 7 BLEU points. Fine-tuning OPUS-MT is not as effective as fine-tuning NLLB (2.58 vs 8.10 after fine-tuning respectively). This illustrates the effectiveness of multilingual pretraining, while OPUS-MT-En-Ar struggles to generalize to Moroccan Arabic given its divergence from MSA.

Model	Dataset	Base	FT	Para	BT
NLLB	DODa	0.82	8.10	9.68	**9.98**
	BIBLE	0.04	0.34	0.82	**0.94**
	MADAR	4.42	5.89	4.67	**6.63**
OPUS-MT	DODa	0.25	2.58	5.02	**5.11**
	BIBLE	**0.35**	0.30	0.21	0.29
	MADAR	1.30	2.05	2.16	**3.20**

Table 2: We provide the BLEU score results on English→Darija. *Base* shows the results of the pretrained model, *FT* the results after fine-tuning, *Para* the results of fine-tuning on paraphrase dataset, and *BT* shows the results of fine-tuning on the dataset with back-translated data.

5.2 Discussion

Disparity between Darija→English and English→Darija performance: There is a significant performance disparity between Darija→English and English→Darija, where the best BLEU score performance on Darija→English is more than 16 BLEU points higher than the performance on English→Darija. We expalin this independently for the two models as follows: In the case of NLLB, we attribute this discrepancy to the amount of English data compared to Darija data. NLLB was trained on significantly more English than Darija data or even MSA data, which makes translating into English easier, this can be seen in the difference in performance of the Base model on the two directions. In the case of OPUS-MT, we explain the performance by the linguistic difference between MSA and Darija, where the decoder of OPUS-MT-EN-AR is trained on generating MSA and struggles to translate into Darija. This means that even after fine-tuning the model will struggle to learn to generate in a new vocabulary and linguistic structure.

Out of distribution generalization: We find that fine-tuning on DODa dataset lacks generalization on the BIBLE dataset, while the performance on MADAR in general improves except for NLLB on Darija→English (Table 1). We attribute this to the fact that MADAR sentences are closely similar to DODa sentences, while BIBLE data is significancy different in domain, and uses a high number of rare MSA words that are not used in Darija due to its vernacular nature, which we suggest explains the significant decrease in performance of NLLB on BIBLE after fine-tuning on DODa (Table 1).

6 Conclusion

In this work, we apply three low resource techniques to train machine translation models for English and Moroccan Arabic. Namely, we experiment with paraphrasing, back-translation and transfer learning. Our results show that combining back-translation with transfer learning achieves the best results, especially when fine-tuning a massively multilingual model such as NLLB, or an encoder that is pretrained on the source language such as DarijaBERT. Furthermore, our results raise concerns about the generalization of these techniques to out-of-domain datasets such as the BIBLE, where fine-tuning can even degrade the performance. Across all the techniques and models, we see a significant disparity between the performance on translating to English versus translating from it to Darija. Overall, our work contributes to the understanding of low-resource MT strategies in real-world scenarios and highlights the need for more equitable approaches to multilingual NLP. Future work should focus on improving robustness to domain shift, developing techniques that work well in both translation directions and investing in better resources and benchmarks for dialectal and underrepresented languages like Moroccan Darija.

Broader Impact

This work contributes to the broader goal of making language technologies more inclusive by advancing machine translation for low-resource languages and dialects such as Moroccan Darija. MT plays an important role in enabling access to information, public services, education and communication, especially in linguistically diverse regions where speakers may have limited proficiency in high-resource languages such as English or even MSA. For many speakers of Moroccan Darija, MT systems can support everyday tasks such as understanding online content, communicating across language barriers and participating in digital platforms that otherwise would be inaccessible. By evaluating practical low-resource MT techniques and revealing key challenges such as out-of-domain generalization and translation direction asymmetry, our work encourages the development of more robust and equitable NLP systems.

At the same time, it is essential to recognize limitations of current models and ensure that MT systems are used responsibly, especially in high-stakes domains like healthcare, law, and public policy where human oversight is critical. Finally, evaluating translation quality solely through automated metrics remains a limitation, therefore future work should include human evaluations by native Darija speakers to better assess usefulness, fluency and cultural relevance.

References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and 1 others. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

ElMehdi Boujou, Hamza Chataoui, Abdellah el Mekki, Saad Benjelloun, Ikram Chairi, and Ismail Berrada. 2021. An open access nlp dataset for arabic dialects : Data collection, labeling, and model construction.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA*, pages 17–24. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2023. Darijabert: a step forward in nlp for the written moroccan dialect.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24. Association for Computational Linguistics. 2nd Workshop on Neural Machine Translation and Generation, WNMT 2018 ; Conference date: 15-07-2018 Through 20-07-2018.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Abderrahmane Issam and Khalil Mrini. 2021. Goud.ma: a news article dataset for summarization in moroccan darija. In *3rd Workshop on African Natural Language Processing*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, and 1 others. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Surafel M Lakew, Stephan Gouws, Yulia Tsvetkov, Vukosi Marivate, and Stefan Weber. 2020. Low-resource neural machine translation: A benchmark for five african languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2602–2611.

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.

Jonathan Mallinson, Rico Sennrich, and Maria Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893. Association for Computational Linguistics (ACL). The 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 ; Conference date: 03-04-2017 Through 07-04-2017.

Hanane Nour Moussa and Asmaa Mourhir. 2023. Darnercorp: An annotated named entity recognition dataset in the moroccan dialect. *Data in Brief*, 48:109234.

Khalil Mrini and Francis Bond. 2017. Building the moroccan darija wordnet (mdw) using bilingual resources.

Aissam Outchakoucht and Hamza Es-Samaali. 2021. Moroccan dialect -darija- open dataset. *Preprint*, arXiv:2103.09687.

Aissam Outchakoucht and Hamza Es-Samaali. 2024. The evolution of darija open dataset: Introducing version 2. *Preprint*, arXiv:2405.13016.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. Arabench: Benchmarking dialectal arabic-english machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107.

Younes Samih and Wolfgang Maier. 2016. An Arabic-Moroccan Darija code-switched corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4170–4175, Portorož, Slovenia. European Language Resources Association (ELRA).

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399, Copenhagen, Denmark. 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.

Ridouane Tachicart, Karim Bouzoubaa, and Hamid Jaafar. 2014. Building a moroccan dialect electronic dictionary (mded).

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, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

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 *Advances in neural information processing systems*, pages 5998–6008.

Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2016. Source language adaptation approaches for resource-poor machine translation. *Computational Linguistics*, 42(2):277–306.

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.

A Training and Evaluation

In all experiments, we keep the same data split to ensure the robustness of the results. We experimentally tune the hyperparameters of each model. In Table 3, we list the final hyperparameters we used for fine-tuning the models in both directions (i.e. Darija→English and English→Darija). We use HuggingFace transformers [7] library for training

[7] https://huggingface.co/docs/transformers

and evaluation. For more details, we release our code publicly [8].

Hyperparameter	NLLB	OPUS-MT	BERT-OPUS
Learning Rate	1e-5	1e-6	8e-5
Batch Size	4	16	16
Weight Decay	0.01	0.01	0.0
Number of Epochs	3	5	7
Warmup Steps	0	0	500

Table 3: Hyperparameters for fine-tuning each model.

B chrF Results

In this section, we provide the results of using chrF metric to compute translation performance in Table 4 and 5.

Model	Dataset	Base	FT	Para	BT
NLLB	DODa	31.35	43.24	43.63	**44.68**
	BIBLE	**42.54**	36.69	34.32	37.08
	MADAR	**47.43**	45.66	45.67	46.70
OPUS-MT	DODa	20.03	33.31	34.02	**35.32**
	BIBLE	**25.94**	24.17	23.01	25.39
	MADAR	25.46	31.58	32.89	**32.94**
BERT-OPUS	DODa	5.06	44.24	42.78	**44.80**
	BIBLE	12.38	19.52	17.01	**20.41**
	MADAR	6.89	33.29	32.95	**35.27**

Table 4: We provide the chrF score results on Darija→English. *Base* shows the results of the pre-trained model, *FT* the results after fine-tuning, *Para* the results of fine-tuning on paraphrase dataset, and *BT* shows the results of fine-tuning on the dataset with back-translated data.

Model	Dataset	Base	FT	Para	BT
NLLB	DODa	14.98	32.49	35.05	**35.59**
	BIBLE	13.24	19.57	24.41	**25.92**
	MADAR	26.27	33.50	32.40	**35.35**
OPUS-MT	DODa	13.27	22.55	26.48	**26.74**
	BIBLE	**20.84**	19.56	17.32	20.77
	MADAR	21.54	23.17	24.34	**26.16**

Table 5: We provide the chrF score results on English→Darija. *Base* shows the results of the pre-trained model, *FT* the results after fine-tuning, *Para* the results of fine-tuning on paraphrase dataset, and *BT* shows the results of fine-tuning on the dataset with back-translated data.

[8] https://github.com/RoscaAlex00/lowresource_mt