AngOFA: Leveraging OFA Embedding Initialization and Synthetic Data for Angolan Language Model

Osvaldo Luamba Quinjica
Department of Computer Science
Beijing Institute of Technology
Beijing, China
aosiwaduo@outlook.com

David Ifeoluwa Adelani
Department of Computer Science
University College London
United Kingdom
d.adelani@ucl.ac.uk

Abstract

In recent years, the development of pre-trained language models (PLMs) has gained momentum, showcasing their capacity to transcend linguistic barriers and facilitate knowledge transfer across diverse languages. However, this progress has predominantly bypassed the inclusion of very-low resource languages, creating a notable void in the multilingual landscape. This paper addresses this gap by introducing four tailored PLMs specifically finetuned for Angolan languages, employing a Multilingual Adaptive Fine-tuning (MAFT) approach. In this paper, we survey the role of informed embedding initialization and synthetic data in enhancing the performance of MAFT models in downstream tasks. We improve baseline over SOTA AfroXLMR-base (developed through MAFT) and OFA (an effective embedding initialization) by 12.3 and 3.8 points respectively.

1 Introduction

Significant advancements have marked the progress of language models and evaluation datasets across various global languages (Devlin et al., 2019; Conneau et al., 2020; Workshop et al., 2023; Xue et al., 2021). Nevertheless, this progress has often bypassed numerous African languages, creating a significant gap. Simultaneously, the majority of African-centric language models have overlooked the inclusion of Angolan languages (Dossou et al., 2022; Alabi et al., 2022; Ogueji et al., 2021). Efforts within the AfricaNLP community have been commendable in broadening downstream evaluation datasets (Adelani et al., 2021; 2022; Muhammad et al., 2023; Ma et al., 2023). However, despite these initiatives, Angolan languages still lack representation.

In the pursuit of developing a multilingual pre-trained language model (PLM), there are two primary approaches. The first entails building a model from scratch, training it directly on multiple languages, employing a specific self-supervised learning such as masked language modeling (Devlin et al., 2019). An alternative approach is multilingual adaptive fine-tuning (MAFT) which involves adapting an existing multilingual pretrained language model to a new set of languages (Alabi et al., 2022; Wang et al., 2022; ImaniGooghari et al., 2023). MAFT gains favor for its resource efficiency, especially in scenarios where computational budgets pose constraints amid the escalating model sizes (Tay et al., 2022; Gupta et al., 2023). The performance of MAFT can be further enhanced by introducing new vocabulary tokens for the additional languages and employing non-Gaussian embedding initialization (Minixhofer et al., 2022; Dobler & de Melo, 2023; Liu et al., 2023a).

In this paper, we introduce the first set of multilingual PLMs tailored for five Angolan languages using the MAFT approach. We compare PLMs developed through MAFT with and without informed embedding initialization, denoted as AngOFA and AngXLM-R , respectively. Leveraging OFA approach to perform embedding initialization before performing MAFT, our findings reveal that AngOFA significantly outperforms AngXLM-R and OFA, underscoring the substantial performance improvements achievable through the incorporation of informed embedding initialization and synthetic data.

Language	Bantu Zone	No. Speakers	NLLB Corpus (MB)	Synthetic Corpus (MB)	Combined Corpus (MB)	Combined No. Sentences
Chokwe (cjk)	Zone K	0.5M	11.3	108.2	119.5	878,824
Kimbundu (kmb)	Zone H	1.7M	10	98.5	108.5	800,603
Kikongo (kon)	Zone H	2M	112.1	107.9	220	2,189,413
Luba-Kasai (lua)	Zone L	0.06M	133.2	113.9	247.1	2,415,794
Umbundu (umb)	Zone R	6M	15.1	98. 5	113.6	902,961
Total		10.2M	281.6	527	808.6	7,187,595

Table 1: **Language Information and Statistics**: Summary of language, language family, number of speakers, number of sentences. All languages belongs to the Niger-Congo/Bantu group, we state the Bantu Zones according to (Smith, 1949). Synthetic corpus was generated using NLLB-600M machine translation model

2 Angolan Languages

Boasting a rich linguistic landscape comprising more than 40 languages and a population of 32 million people, Angolan languages include Portuguese, some Khoisan languages, and mostly Bantu languages from the Niger-Congo family. Despite this linguistic diversity, there is a notable scarcity of literature, radio, or television programming in native Angolan languages. All languages in Angola are written using the Latin script, and many share common digraphs. Due to data scarcity, our focus will primarily revolve around the five most spoken Angolan languages: Umbundu, Kimbundu, Kikongo, Chokwe, and Luba-Kasai. See Table 1 for more details.

3 Approaches to improve MAFT

3.1 Vocabulary Expansion

PLMs are proned to Out-of-Vocabulary (OOV) tokens for languages or scripts uncovered during pre-training. The situation is more pronounced for unseen scripts (Adelani et al., 2021; Pfeiffer et al., 2021), one of the most effective way of dealing with this is to expand the vocabulary of the PLM to cover new tokens (Wang et al., 2019). Glot-500 (ImaniGooghari et al., 2023) was created by first expanding the vocabulary of XLM-R from 250K to 400K before MAFT. However, the new tokens added were randomly initialized.

3.2 OFA: Embedding Factorization

OFA addresses two problems of adapting PLMs to new languages (1) the random initialization of embeddings for new subwords fails to exploit the lexical knowledge encoded in the source model (2) the introduction of additional parameters poses potential obstacles to the efficient training of the finetuned model (Liu et al., 2023a). OFA solves these problems by leveraging both external multilingual embeddings and embeddings in the source PLM to initialize the embeddings of new subwords. In its approach, OFA factorizes the embeddings matrix of the source PLM into two smaller matrices as replacements. Within a lower-dimensional space, the embeddings of non-overlapping new subwords are expressed as combinations of source PLM subword embeddings. These combinations are weighted by similarities derived from well-aligned external multilingual embeddings, i.e., ColexNet+ (Liu et al., 2023b), covering more than one thousand languages. Overlapping subword embeddings are directly copied. This approach ensures that embeddings for subwords shared between the source PLM and the extended vocabulary are integrated, preserving continuity in representation. To complete the process, OFA duplicates all non-embedding parameters from the source PLM model, and the source tokenizer is substituted with the target tokenizer post-vocabulary extension.

3.3 Synthetic data for language modeling

For languages lacking sufficient pre-training data, synthetic data can be generated through dictionary augmentation (Reid et al., 2021) or machine translation (MT) model—an approach very popular in

MT research known as back-translation is an effective way to improve MT model for low-resource languages (Sugiyama & Yoshinaga, 2019; Xia et al., 2019). In this paper, we utilize synthetic data obtained through machine translation as described in (Adelani et al., 2023). The authors generated machine-translated data for 34 African languages(including Angolan languages) with less than 10MB of data, using the English news commentary dataset (Kocmi et al., 2022), which contains over 600K sentences.

4 DATA

4.1 TRAINING DATA

We leveraged the NLLB dataset (NLLB-Team et al., 2022), excluding English translations, and focused solely on Kimbundu, Umbundu, Kikongo, Chokwe, and Luba-Kasai. These languages were concatenated into a single file as our pre-training corpus. Additionally, we added synthetic data generated through NLLB. Table 1 shows the details of the monolingual data.

4.2 EVALUATION DATA

In our work, we evaluated on SIB-200 (Adelani et al., 2023), a text classification dataset that provides train/dev/test sets with 7 classes in more than 200 African languages and dialects. The distribution of the classes are: science/technology (252), travel (198), politics (146), sports (122), health (110), entertainment (93), geography (83). SIB-200 is the only benchmark dataset that covers Angolan languages. We evaluated only on the subset of Angolan languages covered in this work.

5 EXPERIMENTAL SETUP

We utilized the cross-lingual capabilities of XLM-R (Conneau et al., 2020) for training, resulting in the creation of a novel set of PLMs[1]: ANGXLM-R and ANGOFA. These models, underwent distinct fine-tuning processes. Specifically, ANGXLM-R underwent fine-tuning using the MAFT approach outlined in Alabi et al. (2022), with two variants—one trained solely on monolingual data (281.6 MB), and the other incorporating both monolingual and synthetic data (808.7 MB).

Similarly, ANGOFA also underwent two variations of fine-tuning, utilizing the datasets in the same manner as ANGXLM-R . However, ANGOFA followed the configurations outlined for `ofa-multi-768`, as described in (Liu et al., 2023a). We opted to maintain 768 as the only latent dimension in our experiments based on insights from (ImaniGooghari et al., 2023; Liu et al., 2023a) and further supported by preliminary results from our own experiments. These findings revealed evidence of information loss in lower dimensions, particularly noticeable in tasks such as text classification. This dataset partitioning approach aimed to investigate the effects of the MAFT and OFA approaches, both with and without synthetic data, on model performance.

We compared our new models to the following baseline models:

1. XLM-R (Conneau et al., 2020): an encoder-only model that underwent pre-training on 100 languages through a masked language model objective. XLM-R does not cover any language evaluated in this work.
2. Serengeti (Adebara et al., 2023): trained on 500 African languages, including 10 high-resource ones. It includes Kimbundu, Umbundu, and Chokwe.
3. Glot-500 (ImaniGooghari et al., 2023): derived from XLM-R, was extended to cover 500 languages by expanding its vocabulary from 250K to 400K, thus accommodating new tokens representing 400 languages previously absent in XLM-R. Glot-500 covers all Angolan languages used in our evaluation.
4. AfroXLMR-base (Alabi et al., 2022): developed using the MAFT approach, it covers 20 languages with a monolingual corpus of at least 50MB. Angolan languages are not included.

[1]Models available at `https://github.com/zuela-ai/ANGOFA`

	Pre-trained (scratch)		*MAFT*			*MAFT* + syn. data		*OFA*		*OFA + syn*
Lang.	**XLM-R**	**Serengeti**	**Glot 500**	**Afro XLMR**	**ANG XLM-R**	**Afro XLMR76**	**ANG XLM-R**	**ANG OFA**	**OFA 500**	**ANGOFA**
cjk	41.3	43.2	42.9	51.3	43.6	55.6	51.7	46.3	52.8	**58.4**
kmb	44.8	46.9	43.5	50.6	50.2	58.5	56.6	58.5	63.2	**64.7**
kon	67.8	69.1	72.6	65.7	72.5	77.2	76.1	78.8	76.9	**82.4**
lua	54.5	57.9	54.7	62.5	65.4	64.4	73.2	69.1	68.6	**73.5**
umb	50.4	51.7	40.3	50.5	54.9	61.0	56.8	54.3	61.8	**63.3**
Ave.	51.8	53.7	50.7	56.1	57.3	63.3	62.8	61.4	64.6	**68.4**

Table 2: **Benchmark results**: comparing the effectiveness of OFA to random initialization before multilingual adaptive fine-tuning (MAFT)

5. AfroXLMR-base-76L (Adelani et al., 2023): developed using the MAFT approach, it covers languages with at least 10MB of data on the web. It expands coverage to include more languages, notably those listed in the NLLB-200 MT model. Synthetic data was also generated for approximately 30 languages with limited data, including all five Angolan languages. In total, it covers 76 languages.

6. OFA (Liu et al., 2023a): integrates OFA embedding initialization alongside MAFT using Glot500-c (ImaniGooghari et al., 2023), thus including all languages addressed in this work.

6 RESULTS AND DISCUSSION

Table2 shows the performance of our baseline models using the **weighted F1 metric**. We discuss our key findings below:

Region-specific PLMs are better than those pre-trained from scratch with many languages Our results shows that ANGXLM-R created with MAFT performed better than XLM-R, AfroXLMR, Serengeti and Glot-500 with $+5.5$, $+1.2$, $+3.6$, $+6.6$ points respectively. The last two PLMs have been pre-trained on 500+ languages with few Angolan languages but performed worse than AfroXLMR (adapted through MAFT to 20 languages), and ANGXLM-R (adapted to five Angolan languages). This shows that region-specific PLMs covering related languages within the same language family can be more effective.

MAFT results can be boosted by leveraging synthetic monolingual data By incorporating additional synthetic data, ANGXLM-R (+SYN data) performance improved by $+5.5$ over the ANGXLM-R without synthetic data. However, it failed to beat the performance of AfroXLMR-base-76L that has been trained on 76 African languages including all Angolan languages except for Luba-Kasai with the largest data. Our experiment showed that the adapted PLM to 76 languages performed better than Serengeti pre-trained on 500 languages, which further shows that we can create better PLMs to cover more languages through adaptation without the expensive process of pre-training from scratch.

OFA embedding initialization with larger data is more effective Models initialized with OFA demonstrated a consistent improvement compared with other baselines. This indicates that OFA, which explicitly leverages information encoded in the embeddings of the source model and external multilingual embeddings, is superior to random initialization. Notably, ANGOFA's advantage over OFA is accentuated by its access to a significantly larger corpus of data for the respective languages through the use of synthetic data. Without the additional synthetic data ANGOFA performed worse than OFA pre-trained on 500 languages with a drop of -3.2. However, when we trained on the synthetic data, ANGOFA achieved the best overall performance with $+16.6$ over XLM-R, $+12.3$ over AfroXLMR, and $+5.6$ over ANGXLM-R (with synthetic data).

7 CONCLUSION AND FUTURE WORK

This paper introduces four multilingual PLMs models tailored for Angolan languages. Our experimental findings illustrate that employing informed embedding initialization significantly enhances the performance of a MAFT model in downstream tasks. While models initialized with OFA exhibit superior results compared to their counterparts, even in the case where ANGXLM-R finetuned on a larger corpus of data for the respective languages performs poorly as compared to OFA finetuned on a smaller corpus. Nevertheless, the specific factors contributing to ANGXLM-R 's superiority over OFA, especially in the context of Luba-Kassai, raise intriguing questions about the primary determinants influencing the performance of models in downstream tasks, including considerations like dataset size versus informed embedding initialization. These questions are left for future investigation. Furthermore, we aim to expand the application of OFA to more African languages for further exploration.

ACKNOWLEDGMENTS

This work was supported in part by Oracle Cloud credits and related resources provided by Oracle. David Adelani acknowledges the support of DeepMind Academic Fellowship programme.

REFERENCES

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. SERENGETI: Massively multilingual language models for Africa. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1498–1537, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.97. URL https://aclanthology.org/2023.findings-acl.97.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4488–4508, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.298. URL https://aclanthology.org/2022.emnlp-main.298.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9: 1116–1131, 2021. doi: 10.1162/tacl_a_00416. URL https://aclanthology.org/2021.tacl-1.66.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects, 2023. URL https://arxiv.org/abs/2309.07445.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.382.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Konstantin Dobler and Gerard de Melo. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13440–13454, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.829. URL https://aclanthology.org/2023.emnlp-main.829.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In Angela Fan, Iryna Gurevych, Yufang Hou, Zornitsa Kozareva, Sasha Luccioni, Nafise Sadat Moosavi, Sujith Ravi, Gyuwan Kim, Roy Schwartz, and Andreas Rücklé (eds.), *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pp. 52–64, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.sustainlp-1.11. URL https://aclanthology.org/2022.sustainlp-1.11.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. Continual pre-training of large language models: How to re-warm your model? In *Workshop on Efficient Systems for Foundation Models @ ICML2023*, 2023. URL https://openreview.net/forum?id=pg7PUJe0Tl.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. Glot500: Scaling multilingual corpora and language models to 500 languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1082–1117, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.61. URL https://aclanthology.org/2023.acl-long.61.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. Findings of the 2022 conference on machine translation (WMT22). In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 1–45, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.1.

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schütze. Ofa: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining, 2023a. URL https://arxiv.org/abs/2311.08849.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Renhao Pei, and Hinrich Schuetze. Crosslingual transfer learning for low-resource languages based on multilingual colexification graphs. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b. URL https://openreview.net/forum?id=Tn5hALAaA4.

Chunlan Ma, Ayyoob ImaniGooghari, Haotian Ye, Ehsaneddin Asgari, and Hinrich Schütze. Taxi1500: A multilingual dataset for text classification in 1500 languages, 2023. URL https://arxiv.org/abs/2305.08487.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3992–4006, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.293. URL https://aclanthology.org/2022.naacl-main.293.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif Mohammad, Meriem Beloucif, and Sebastian Ruder. SemEval-2023 task 12: Sentiment analysis for African languages (AfriSenti-SemEval). In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori (eds.), *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 2319–2337, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.315. URL https://aclanthology.org/2023.semeval-1.315.

NLLB-Team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672, 2022. URL https://api.semanticscholar.org/CorpusID:250425961.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin (eds.), *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.11. URL https://aclanthology.org/2021.mrl-1.11.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. UNKs everywhere: Adapting multilingual language models to new scripts. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10186–10203, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.800. URL https://aclanthology.org/2021.emnlp-main.800.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1306–1320, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.99. URL https://aclanthology.org/2021.emnlp-main.99.

Edwin W. Smith. The classification of the bantu languages. by malcolm guthrie, ph.d. published for the international african institute by the oxford university press, 1948. pp. 91. map. 8s. 6d. net. *Africa*, 19(1):73–74, 1949. doi: 10.2307/1156267.

Amane Sugiyama and Naoki Yoshinaga. Data augmentation using back-translation for context-aware neural machine translation. In Andrei Popescu-Belis, Sharid Loáiciga, Christian Hardmeier, and Deyi Xiong (eds.), *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pp. 35–44, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6504. URL https://aclanthology.org/D19-6504.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pretraining and finetuning transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=f2OYVDyfIB.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, and Dong Yu. Improving pre-trained multilingual model with vocabulary expansion. In Mohit Bansal and Aline Villavicencio (eds.), *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 316–327, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1030. URL https://aclanthology.org/K19-1030.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 863–877, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.61. URL https://aclanthology.org/2022.acl-long.61.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo

Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL `https://arxiv.org/abs/2211.05100`.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. Generalized data augmentation for low-resource translation. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5786–5796, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.

18653/v1/P19-1579. URL `https://aclanthology.org/P19-1579`.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL `https://aclanthology.org/2021.naacl-main.41`.