ACL/EACL/EMNLP 2023 Tutorial Proposal
Everything you need to know about Multilingual LLMs: Towards fair, performant and reliable models for the languages of the world

Sunayana Sitaram
Microsoft Research India
sunayana,sitaram@microsoft.com

Monojit Choudhury
Microsoft Turing, India
monojitc@microsoft.com

Barun Patra
Microsoft Turing, USA
bapatra@microsoft.com

Vishrav Chaudhary
Microsoft Turing, USA
vchaudhary@microsoft.com

Kabir Ahuja
Microsoft Research India
t-kabirahuja@microsoft.com

Kalika Bali
Microsoft Research India
kalikab@microsoft.com

1 Tutorial content

This tutorial will describe various aspects of scaling up language technologies to many of the world's languages by presenting the latest research in Massively Multilingual Language Models (MMLMs). We will cover topics such as data collection, training and fine-tuning of models, Responsible AI issues such as fairness, bias and toxicity, linguistic diversity and evaluation in the context of MMLMs, specifically focusing on issues in non-English and low-resource languages. Further, we will also talk about some of the real-world challenges in deploying these models in language communities in the field. With the performance of MMLMs improving in the zero-shot setting for many languages, it is now becoming feasible to use them for building language technologies in many languages of the world, and this tutorial will provide the computational linguistics community with unique insights from the latest research in multilingual models. Although past tutorials have covered some of these topics (such as linguistic diversity, data and training of models), there has been a lot of interesting research in the recent past that the CL community will benefit from knowing about. Further, this will be the first tutorial (as per our knowledge) that will discuss issues of deployment in language communities and Responsible AI in the context of multilingual models.

This tutorial will present a broad survey covering work done by several research groups (as indicated in the references), including work done by the authors.

Type of the tutorial: cutting-edge

Target audience and pre-requisites: The target audience for this tutorial are researchers from industry and academia who work on Large Language Models, and are interested in learning about the latest research in multilingual models to build systems for non-English languages, low-resource languages and multilingual speakers. We will not be covering the basics of LLMs, so we expect that the audience will be familiar with (at least the English versions of) models such as BERT.

1.1 Outline of the tutorial

We plan to have five talks of 30/40 minutes each, along with a 10 minute introduction, with 10 minutes for general discussion/spillover.

Introduction: We will start with a short introduction on MMLMs, describing the models that are available today and present the SOTA in model performance on various tasks across different languages.

Data and pre-training: The main goal of this section would be to outline the techniques leveraged for creating a high quality corpus for pre-training strong MMLMs. We will cover the challenges encountered in creating such a corpus as highlighted in CC100 (Conneau et al., 2020), mC4 (Xue et al., 2021), OSCAR (Ortiz Suárez et al., 2020), ROOTS (Laurençon et al., 2022) etc., and provide an overview of the various stages of such a dataset creation pipeline. Ensuring the quality of the training corpus is highly important as it is directly correlated to the performance of MMLMs (Kaplan et al., 2020). In addition to this, we will also discuss the pre-training strategies and possible extensions for extending the recipe to multiple languages (Conneau and Lample, 2019; Artetxe and Schwenk, 2019) describing how scaling (both on the data and model axis) can substantially help improve model performance (Conneau et al., 2020;

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 6: Tutorial Abstracts, pages 21–26

Xue et al., 2021), aiding in bridging the gap between the English performance of a multilingual and an English only model, thereby reducing the curse of Multilinguality.

Training paradigms and fine-tuning: We will describe different training paradigms (Eg: an Electra based approach (Chi et al., 2022; He et al., 2021)) and how to leverage bitext data, discussing results of using contrastive learning approaches (Chi et al., 2021) or extensions to Electra based approaches (Chi et al., 2022), as well as showing the benefits of going beyond English centric bitexts (Patra et al., 2022). We will also discuss some orthogonal approaches of training encoder-decoder multilingual representation models (Liu et al., 2020; Ma et al., 2021; **?**), as well as complimentary techniques to build better encoder models (Eg: Adapter based approaches (Pfeiffer et al., 2022)). We will also focus on different strategies for improving the fine-tuning performance of these models. This includes techniques encouraging models to have more consistent predictions across languages (Zheng et al., 2021), leveraging weight perturbations to avoid overfitting (Wu et al., 2022) or techniques to reduce the sharpness of loss minima for better generalization (Foret et al., 2021; Bahri et al., 2022).

Performance evaluation and reliability: While the state-of-the-art multilingual models support around 100 languages of the world, most existing multilingual benchmarks contain evaluation data in a handful of languages (Ahuja et al., 2022b). We will discuss some potential approaches to scale up multilingual evaluation like performance prediction (Lin et al., 2019; Xia et al., 2020; Ahuja et al., 2022c) and structure probing (Müller-Eberstein et al., 2022; Clouâtre et al., 2022). We will also focus on measuring the cost-performance trade-offs and sample efficiencies of fine-tuning MMLMs with different sources of data (translation vs manual collection)(Ahuja et al., 2022a). Further, we will cover how to measure reliability in the confidence predictions of multilingual models under a zero-shot and few-shot setup by studying their calibration (Ahuja et al., 2022d).

FATE issues: LLMs are known to pick up the biases present in the datasets that are trained on. In case of multilingual LLMs, apart from bias and fairness issues at group and individual level, one also need to address the issue of disparity of zero-shot transfer accuracies across languages and varieties (Choudhury and Deshpande, 2021; Lauscher et al., 2020). Furthermore, there is little work done on the interaction among the biases in corpora from different languages, influence of grammatical gender (Cao and Daumé, 2021) and other syntactic and semantic factors on measurement and mitigation of biases, and socio-cultural aspects of biases (Sambasivan et al., 2021). In this section of the tutorial, we will survey the work done so far in non-English FATE issues and present challenges that remain to be addressed.

Deploying to language communities: LLMs today are trained using billions of parameters, making them infeasible to be used in low-memory footprint devices. Language communities (particularly those that speak under-resourced languages) that may benefit the most from Speech and NLP technologies may not have good enough connectivity to be able to use models hosted on the cloud. This necessitates the development or distillation of lightweight models for low-resource languages, and in this section, we will present research in this direction (Diddee et al., 2022). We will study the state of current LT to serve communities speaking different languages for critical situations such as healthcare bots (Mondal et al., 2022). Further, there are many social and cultural factors to be taken into account while deploying MMLMs to language communities, which we will also discuss in this section.

1.2 Diversity considerations

The topic of the tutorial inherently encourages linguistic diversity. In terms of gender diversity, two of the tutorial presenters are female, while four are male. In this tutorial, we will cover issues related to Responsible AI (fairness, toxicity) and deploying to under-resourced language communities which will improve diversity considerations while building LLMs. The instructors are a mix of senior, mid-career and junior researchers.

1.3 Reading list

Please check the references section for the reading list.

2 Instructor bios

Sunayana Sitaram is a Senior Researcher at Microsoft Research India, where she works on multilingual speech and NLP. Her current research interests include training and evaluation of Mas-

sively Multilingual Language Models and Responsible AI for NLP. Prior to coming to MSRI as a Post Doc, Sunayana completed her MS and PhD at the Language Technologies Institute, Carnegie Mellon University in 2015. Sunayana's research has been published in top NLP and Speech conferences including ACL, NAACL, EMNLP, Interspeech, ICASSP. She has organized special sessions and workshops on under-resourced languages, code-switching, multilingual evaluation and speech for social good. She has also led the creation of several benchmarks and datasets in code-switching, ASR, NLI and TTS that have been used by research groups all over the world.

Monojit Choudhury is a Principal Applied Scientist at Microsoft Turing, prior to which he was a Principal Researcher at Microsoft Research India. He is also a Professor of Practice at Plaksha University, and had held adjunct faculty positions at Ashoka University, IIIT Hyderabad and IIT Kharagpur. Over the past 15 years, Monojit has worked on several impactful projects on processing of code-mixed text, evaluation and linguistic fairness of large language models, and social impact through participatory design of technology for under-resourced languages like Gondi, Mundari, Idu Mishmi and Swahili. Monojit has served as Senior Area Chair and Area chair in leading NLP and AI conferences including EMNLP, ACL, NAACL, IJCNLP and AAAI. He has organized several successful workshops in *ACL conferences (SUMEval 2022, CALCS series, TextGraph series, etc.) and has delivered a tutorial on Code-mixed text processing at EMNLP 2019. He is the general chair of the Panini Linguistics Olympiad and the founding co-chair of Asia Pacific Linguistics Olympiad – programs to introduce bright young students to linguistics and computational linguistics through puzzles. Dr. Choudhury holds PhD and B.Tech degrees in Computer Science and Engineering from IIT Kharagpur.

Vishrav Chaudhary is a Principal Researcher at Microsoft Turing where he works on scaling and building efficient Multilingual and Multimodal representation and generation models. Prior to Microsoft, Vishrav was a Lead Researcher at FAIR and focused on several aspects of Machine Translation, Quality Estimation and Cross-lingual understanding. Over the past 10 years, Vishrav's research work has been published in several leading NLP and AI conferences and journals including ACL, EMNLP, NAACL, EACL, AACL, TACL, JMLR and AMTA. He has also organized several workshops successfully including SUMEval 2022, AmericasNLP 2021, WMT 2021 etc. He has also served as an Area Chair for EMNLP 2022. Vishrav has also led creation of benchmarks and datasets targeting 100+ languages which have been used to train state-of-the-art Cross Lingual Representation and Machine Translation models.

Barun Patra is an Applied Scientist at Microsoft Turing. His research interest revolves around building better foundational models that can help support numerous NLP tasks across different languages. Barun's research work focuses on improving the quality and efficiency of training these large multilingual foundational models, helping achieve state-of-the-art performance on cross-lingual NLP tasks.

Kabir Ahuja is a Research Fellow at Microsoft Research India, where he works on building linguistically fair multilingual models covering different aspects around their performance, calibration, evaluation, interpretation, and data collection. He is also interested in the analysis and interpretability of the computation mechanisms utilized by neural sequence models for solving different tasks.

Kalika Bali is a Principal Researcher at Microsoft Research India working in the areas of Machine Learning, Natural Language Systems and Applications, as well as Technology for Emerging Markets. Her research interests lie broadly in the area of Speech and Language Technology especially in the use of linguistic models for building technology that offers a more natural Human-Computer as well as Computer-Mediated interactions.

3 Other

Estimate of audience size: 50

Venues: We would prefer ACL 2023 to be the venue for the tutorial, but EMNLP and EACL are also acceptable. We do not forsee any special requirements for technical equipment.

3.1 Ethics statement

This tutorial will present current research on Multilingual model training, evaluation, Responsible AI issues and deploying models in the field. Although we aim to promote linguistic diversity by discussing issues pertaining to multilingual models trained on around 100 languages, many languages

of the world are not supported by these models. Further, the techniques that we will discuss mainly apply to written languages, while unwritten languages will be excluded from the tutorial.

References

Kabir Ahuja, Monojit Choudhury, and Sandipan Dandapat. 2022a. On the economics of multilingual few-shot learning: Modeling the cost-performance trade-offs of machine translated and manual data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1369–1384, Seattle, United States. Association for Computational Linguistics.

Kabir Ahuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2022b. Beyond static models and test sets: Benchmarking the potential of pre-trained models across tasks and languages. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 64–74, Dublin, Ireland. Association for Computational Linguistics.

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022c. Multi task learning for zero shot performance prediction of multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat, and Monojit Choudhury. 2022d. On the calibration of massively multilingual language models.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Dara Bahri, Hossein Mobahi, and Yi Tay. 2022. Sharpness-aware minimization improves language model generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7360–7371, Dublin, Ireland. Association for Computational Linguistics.

Yang Trista Cao and III Daumé, Hal. 2021. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*. *Computational Linguistics*, 47(3):615–661.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: Cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6170–6182, Dublin, Ireland. Association for Computational Linguistics.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12710–12718.

Louis Clouâtre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. 2022. Detecting languages unintelligible to multilingual models through local structure probes.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Harshita Diddee, Sandipan Dandapat, Monojit Choudhury, Tanuja Ganu, and Kalika Bali. 2022. Too brittle to touch: Comparing the stability of quantization and distillation towards developing lightweight low-resource mt models.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg

Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. 2022. The bigscience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv preprint arXiv:2106.13736*.

Ishani Mondal, Kabir Ahuja, Mohit Jain, Jacki O'Neill, Kalika Bali, and Monojit Choudhury. 2022. Global readiness of language technology for healthcare: What would it take to combat the next pandemic? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4320–4335, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2022. Sort by structure: Language model ranking as dependency probing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1296–1307, Seattle, United States. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Barun Patra, Saksham Singhal, Shaohan Huang, Zewen Chi, Li Dong, Furu Wei, Vishrav Chaudhary, and Xia Song. 2022. Beyond english-centric bitexts for better multilingual language representation learning. *arXiv preprint arXiv:2210.14867*.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 315–328, New York, NY, USA. Association for Computing Machinery.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. NoisyTune: A little noise can help you finetune pretrained language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 680–685, Dublin, Ireland. Association for Computational Linguistics.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. 2020. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8625–8646, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting

Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3403–3417, Online. Association for Computational Linguistics.