CS 7170 *Seminar in Artificial Intelligence*

Multilingual Language Modeling

Time: MR 11:45-1:25
Location: WVH 166
Instructor: Terra Blevins
Office Hours: Monday 1:45-2:45 (after class) in 177 Huntington, Room 2225, or by appointment
Canvas: https://northeastern.instructure.com/courses/235307/

Description

This seminar provides an overview of multilingual NLP in the era of generative modeling. We will discuss the fundamentals of multilingual language processing, as well as recent advances (and bottlenecks) in the field, due to LLMs. The course will involve lectures introducing topics in multilingual modeling, student-led paper discussions, and a final project proposal.

Prerequisites

There are no formal prerequisites for this course, but students should have some familiarity with machine learning methods and natural language processing.

While there are no language or package requirements for the computational portion of the project proposal, most research on language models is done in Python and with the Huggingface packages. Here is a tutorial to getting started working with LLMs in the Huggingface ecosystem.

Coursework and Grading

Participation + Presenting (40%): The primary focus of this course will be on paper discussions of seminal and recent works in multilingual NLP from the assigned reading list. Students will be expected to present papers (from the PDF) and lead discussions of the works throughout the semester, and to participate in these discussions when not presenting.

- **EMNLP Recap:** Students will create summary presentations of multilingual NLP @ EMNLP 2025 to discuss the most recent work in the field, and present their findings to the class on 11/10.

Project Proposal (60%): The final project for this course is a *proposal*, accompanied by *initial results* and a *literature review*, for a novel multilingual NLP project with LMs; the primary requirement for this project scope is for the research questions to address multiple languages rather than just one. The final version of the proposal should also include a description of, and initial results from, a proof-of-concept for the proposed research direction. The project consists of four (4) deliverables, outlined below:

- **Proposal (5%, due 09/25)**: A one-page summary outlining the motivation, hypothesis, and approach for the student's project. This portion is analogous to the introduction to a conference paper submission.
- **Literature Review (15%, due 10/30)**: A comprehensive review of related works to the student's project proposal. This proposal can be written similarly to a paper's related work section; however, the write-up should be *complete* and *thoroughly engage with all of the included works.* This submission should consist of (1) the completed literature review and corresponding

bibliography and (2) a revised proposal that addresses the relationship between the project as proposed and existing prior work.

- **Final Report (25%, due 12/11)**: This writeup should include your original proposal, a summary of your literature review, a description of your proof of concept implementation, and a discussion of the initial results and their implications for the research direction. The final write-up should be no more than four pages in length.
- **Final Presentation (15%, due 12/08-11):** Students will give a presentation based on their project proposal and initial findings to the class. These presentations should take the form of a research conference talk, and students should expect to answer questions from the instructor/class about their project.

Tentative Schedule

Topic	Date		Reading/ Assignments Due
Introduction	09/04	Overview of the syllabus and introduction to multilingual language modeling	Background Reading: chapters 10, 11, and 12 of *Jurafsky and Martin (2025)*, particularly if you have less familiarity with recent advances in NLP. Additional Slides on: Language Modeling (from Noah Smith) and Multilingual NLP (from Graham Neubig)
Architectures for Multilingual Models	09/08	Lecture introduction and paper discussion	Papers: Conneau and Lample (2019)
	09/11	Paper discussion on distributed architectures	Papers: Blevins et al. (2024), Xue et al. (2021), Jiang et al. (2024)
	Additional References: mBERT Github Documentation, Conneau et al. (2020a), Lin et al. (2022)		
Multilingual (Pre)training data	09/15	Lecture introduction and paper discussion	Papers: Blevins and Zettlemoyer (2023), Imani et al. (2023)
	09/18	Paper discussion	Papers: Kreutzer et al. (2022), van Esch et al. (2024)
	Additional References: Longre et al. (2023), Briakou et al. (2023), Laurençon et al. (2022), Suárez et al. (2019), MADLAD-400 (Kudugunta et al., 2023), Seto et al. (2025), Lau et al. (2025), World Atlas of Language Structures (WALS), lang2vec package		
Multilingual Tokenization	09/22	Lecture introduction and paper discussion	Papers: Rust et al. (2019)
	09/25	Paper discussion on subword tokenization	Papers: Arnett and Bergen (2025), Ahia et al. (2023)

		challenges	Due: Project Proposals
	09/29	Byte modeling paper discussion	Papers: Xue et al. (2022), Limisiewicz et al. (2024)
	Additional References: Sennich et al. (2016), Mielke et al. (2021), Reddy et al. (2025), Clark et al. (2022), Limisiewicz et al. (2023), Downey et al. (2023), Foroutan et al. (2025), Apidianaki (2023)		
Analyzing Multilingual Models	10/02	Lecture introduction and paper discussion	Papers: Choenni et al. (2023)
	10/06	Paper discussion on internal representations	Papers: Foroutan et al. (2022), Riemenschneider and Frank (2025)
	10/09	Paper discussion on English biases	Papers: Papadimitriou et al. (2023), Wendler et al. (2024)
	Additional References: de Varda and Marelli (2023), Blevins et al. (2022), Taktasheva et al. (2021), Brinkmann et al. (2025), Belinkov (2022), de Vries et al. 2020, Tang et al., 2024, Zhang et al. (2024), Gerz et al. (2018)		
	10/13	No Class (Indigenous Peoples Day)	
	10/16	Guest Lecture: **Kalika Bali**, Researcher @ MSR India	
Multilingual Evaluation	10/20	Lecture introduction and paper discussion	Papers: Üstün et al. (2024), Ahuja et al. (2023)
	10/23	Paper discussion	Papers: Liu et al. (2025), Lignos et al. (2022)
	Additional References: Asai et al. (2024), Ruder et al. (2023), Universal Dependencies, UniversalNER, MasakhaneNLP, Adelani et al. (2021), Faisal et al. (2024)		
Curse of Multilinguality	10/27	Lecture introduction and paper discussion	Papers: K et al. (2020), Chang et al. (2024)
	10/30	Paper discussion	Papers: Pfeiffer et al. (2022), Downey et al. (2024), Due: project literature review
	Additional References: Wu and Dredze (2020)		
Cross-lingual Transfer	11/03	Lecture introduction and paper discussion	Papers: Wu and Dredze (2019), Conneau et al. (2020b)
	11/06	Paper discussion	Papers: Malkin et al. (2022), Shaham et al. (2024)

	Additional references: Pires et al. (2020), Wang et al. (2020), Dober and de Melo (2023), Downey et al. (2023)		
EMNLP Recap	11/10	Multilingual NLP @ EMNLP 2025	Due: student presentations on EMNLP
Low-Resource Language Modeling	11/13	Lecture introduction and paper discussion	Papers: Ebrahimi and Kann (2021), Yong et al. (2023)
	Additional References: Ogueji et al. (2021), Chen et al. (2023), Ògúnrèmí et al. (2023), Muennighoff et al. (2023)		
Multiculturalism in Language Models	11/17	Lecture introduction and paper discussion	Papers: Pawar et al. (2025), AlKhamissi et al. (2024)
	11/20	Paper discussion	Papers: Keleg and Magdy (2023), Veselovsky et al. (2025)
	Additional References: Ersoy et al. (2023), Naous et al. (2024), Karamolegkou et al. (2024), Han et al. (2023),		
The Future of Multilingual Language Models	11/24	Paper discussion on models	Papers: Chitale et al. (2025), Dash et al. (2025)
	11/27	No Class (Thanksgiving)	
	12/01	Paper discussion on applications	Papers: Yang et al. (2025), Hanley and Durumeric (2025)
	Additional References: Dang et al. (2024), Shafique et al. (2025), Amirzadeh et al. (2025), Zhang et al. (2024), Upadhayay and Behzadan (2025)		
Project Presentations	12/04	Student presentations	Due: student presentations on final project
	12/08	Student presentations	Due: student presentations on final project Due 12/11: final project reports

Course Policies

Attendance and Tardiness: Classes will be held in-person each week -- any deviation in the delivery of the lecture will be announced through Canvas. Attendance and active engagement during lecture is highly recommended, and presenting and participation in class each week is a key part of your learning experience and grade. *If you have circumstances that can prevent you from being in class on time, please email the instructor in advance (if possible).*

Make-up Policy: The majority of course assignments are based on in-class participation. For submitted assignments throughout the semester, students can obtain late days for legitimate reasons such as illness

and family emergencies; *however, these cases need to be approved prior to the submission deadline by the instructor.* No late days can be used for the final project deliverables.

Regrade Considerations: If you are confused or concerned about feedback on course assignments, please directly email the instructor. You must submit any requests for grading reconsideration within 7 days after the feedback was released.

Academic Integrity: Please read the Northeastern Academic Integrity Policy. All students are required to adhere to this policy during the course. Note that while students are encouraged to discuss course materials, no plagiarism/copying is allowed.

This course also has a firm no AI usage policy for completing the following portions of the coursework: *reading* (do not have a model summarize or otherwise explain the paper for you), *presenting*, and *writing*. All work submitted to this course should be primarily completed by you, without the help of an AI. The following cases are exceptions to this policy:

- *Coding Support*: You may use an AI such as CoPilot to help you with the project implementation.
- *Grammar*: You may use AI support to help you edit your writing, once you have produced the first draft. The AI may help you with grammar and wording, but the content of the text should be wholly produced by you.

However, ***in any case where you use AI***, you must thoroughly check the output of the model. If you choose to use AI for these cases, you take ownership of the final product, regardless of any errors or artifacts introduced by the AI.

The first time you are found in violation of the policies in this syllabus on an assignment, you will receive a 0 for the associated work. A second violation, or a violation during the final project deliverables, will result in failing the course.

Classroom Environment: To create and preserve a classroom atmosphere that optimizes teaching and learning, all participants share a responsibility in creating a civil and non-disruptive forum for the discussion of ideas. Students are expected to conduct themselves at all times in a manner that is respectful towards all participants and does not disrupt teaching or learning. The instructor reserves the right to interrupt conversations that deviate from these expectations. Repeated unprofessional or disrespectful conduct may result in a lower grade or more severe consequences.

Title IX: Northeastern's Title IX Policy prohibits Prohibited Offenses, which are defined as sexual harassment, sexual assault, relationship or domestic violence, and stalking. The Title IX Policy applies to the entire community, including male, female, transgender students, faculty and staff.

If you or someone you know has been a survivor of a Prohibited Offense, *confidential* support and guidance can be found through **University Health and Counseling Services** staff and the **Center for Spiritual Dialogue and Service** clergy members. By law, those employees are not required to report allegations of sex or gender-based discrimination to the University.

Alleged violations can be reported non-confidentially to the Title IX Coordinator within **The Office for Gender Equity and Compliance** at titleix@northeastern.edu and/or through **NUPD** (Emergency

617.373.3333; Non-Emergency 617.373.2121). Reporting Prohibited Offenses to NUPD does **NOT** commit the victim/affected party to future legal action.

Faculty members are considered "mandatory reporters" at Northeastern University, meaning they are required to report all allegations of sex or gender-based discrimination to the Title IX Coordinator.

Please visit https://www.northeastern.edu/titleix for a complete list of reporting options and resources both on- and off-campus.

Students with Disabilities: Students who have disabilities who wish to receive academic services and/or accommodations should visit the Disability Resource Center at 20 Dodge Hall or call (617) 373-2675. If you have already done so, please provide your letter from the DRC to me early in the semester so that I can arrange those accommodations.

References:

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, et al. *MasakhaNER: Named Entity Recognition for African Languages*. Transactions of the Association for Computational Linguistics. 2021.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, Yulia Tsvetkov. D*o All Languages Cost the Same? Tokenization in the Era of Commercial Language Models*. In the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, Sunayana Sitaram. *MEGA: Multilingual Evaluation of Generative AI*. In the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, Mona Diab. *Investigating Cultural Alignment of Large Language Models*. In the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024.

Catherine Arnett and Benjamin K. Bergen. *Why do language models perform worse for morphologically complex languages?* In the Proceedings of the 31st International Conference on Computational Linguistics (COLING). 2025.

Terra Blevins and Luke Zettlemoyer. *Language Contamination Helps Explain the Cross-lingual Capabilities of English Pretrained Models*. In the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, Luke Zettlemoyer. *Breaking the Curse of Multilinguality with Cross-lingual Expert Language Models*. In the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. *When is multilinguality a curse? Language modeling for 250 high-and low-resource languages*. In the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024.

Pranjal A. Chitale, Varun Gumma, Sanchit Ahuja, Prashant Kodali, Manan Uppadhyay, Deepthi Sudharsan, and Sunayana Sitaram. *The role of synthetic data in Multilingual, Multi-cultural AI systems: Lessons from Indic Languages*. arXiv preprint arXiv:2509.21294. 2025.

Rochelle Choenni, Dan Garrette, Ekaterina Shutova. *How do languages influence each other? Studying cross-lingual data sharing during LM fine-tuning*. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Alexis Conneau and Guillaume Lample. *Cross-lingual language model pretraining*. Advances in neural information processing systems 32 (2019).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov. *Unsupervised Cross-lingual Representation Learning at Scale*. In the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020a.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. *Emerging Cross-lingual Structure in Pretrained Language Models*. In the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020b.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, et al. Aya Vision: Advancing the Frontier of Multilingual Multimodality. arXiv preprint arXiv:2505.08751. 2025.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, et al. *Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier*. arXiv preprint arXiv:2412.04261. 2024.

C. M. Downey, Terra Blevins, Dhwani Serai, Dwija Parikh, Shane Steinert-Threlkeld. *Targeted Multilingual Adaptation for Low-resource Language Families.* In Findings of the Association for Computational Linguistics: EMNLP 2024.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, Karl Aberer. *Discovering Language-neutral Sub-networks in Multilingual Language Models*. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022.

Hans William Alexander Hanley and Zakir Durumeric. *Hierarchical Level-Wise News Article Clustering via Multilingual Matryoshka Embeddings*. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, Hinrich Schütze. *Glot500: Scaling*

Multilingual Corpora and Language Models to 500 Languages. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, et al. *Mixtral of Experts*. arXiv preprint arXiv:2401.04088. 2024.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition*. Online manuscript released January 12, 2025.

Karthikeyan K, Zihan Wang, Stephen Mayhew, Dan Roth. *Cross-Lingual Ability of Multilingual BERT: An Empirical Study*. In Proceedings of the International Conference on Learning Representations, 2020.

Amr Keleg and Walid Magdy. *DLAMA: A Framework for Curating Culturally Diverse Facts for Probing the Knowledge of Pretrained Language Models*. In Findings of the Association for Computational Linguistics: ACL 2023.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, et al., *Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets*. Transactions of the Association for Computational Linguistics, 2022.

Constantine Lignos, Nolan Holley, Chester Palen-Michel, Jonne Sälevä. *Toward More Meaningful Resources for Lower-resourced Languages*. Findings of the Association for Computational Linguistics: ACL 2022.

Tomasz Limisiewicz, Jiří Balhar, David Mareček. *Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages*. In Findings of the Association for Computational Linguistics: ACL 2023.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, Luke Zettlemoyer. *MYTE: Morphology-Driven Byte Encoding for Better and Fairer Multilingual Language Modeling*. In the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, Lidong Bing. *Is Translation All You Need? A Study on Solving Multilingual Tasks with Large Language Models*. In the Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. 2025.

Dan Malkin, Tomasz Limisiewicz, Gabriel Stanovsky. *A Balanced Data Approach for Evaluating Cross-Lingual Transfer: Mapping the Linguistic Blood Bank*. In the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, Colin A Raffel. *Scaling Data-Constrained Language Models*. In Proceedings of Advances in Neural Information Processing Systems, 2023.

Tolulope Ogunremi, Dan Jurafsky, Christopher Manning. *Mini But Mighty: Efficient Multilingual Pretraining with Linguistically-Informed Data Selection*. In Findings of the Association for Computational Linguistics: EACL 2023.

Isabel Papadimitriou, Kezia Lopez, Dan Jurafsky. *Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models*. In Findings of the Association for Computational Linguistics: EACL 2023.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, Isabelle Augenstein. *Survey of Cultural Awareness in Language Models: Text and Beyond.* Computational Linguistics, 2023.

Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, Mikel Artetxe. *Lifting the Curse of Multilinguality by Pre-training Modular Transformers*. In the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022.

Riemenschneider and Frank. *Cross-Lingual Generalization and Compression: From Language-Specific to Shared Neurons*. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, Iryna Gurevych. *How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models*. In the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. 2021.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, Matan Eyal. *Multilingual Instruction Tuning With Just a Pinch of Multilinguality*. In Findings of the Association for Computational Linguistics: ACL 2024.

Daan van Esch, Sandy Ritchie, Sebastian Ruder, Julia Kreutzer, Clara Rivera, Ishank Saxena, Isaac Caswell. *Connecting Language Technologies with Rich, Diverse Data Sources Covering Thousands of Languages*. In the Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING). 2024.

Veselovsky, Veniamin, Berke Argin, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. *Localized Cultural Knowledge is Conserved and Controllable in Large Language Models*. arXiv preprint arXiv:2504.10191, 2025.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, Robert West. D*o Llamas Work in English? On the Latent Language of Multilingual Transformers*. In the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024.

Shijie Wu and Mark Dredze. *Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT*. In

Shijie Wu and Mark Dredze. *Are All Languages Created Equal in Multilingual BERT?* In the Proceedings of the 5th Workshop on Representation Learning for NLP. 2020.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, Colin Raffel. *mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer*. In the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, Colin Raffel. *ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models*. Transactions of the Association for Computational Linguistics, 10. 2022.

Ivory Yang, Weicheng Ma, and Soroush Vosoughi. *NüshuRescue: Reviving the Endangered Nüshu Language with AI*. In Proceedings of the 31st International Conference on Computational Linguistics. 2025.

Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, et al. *BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting*. In the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023.