Nalin Kumar

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic

nkumar@ufal.mff.cuni.cz
https://ufal.mff.cuni.cz/nalin-kumar

1 Research interests

General-purpose Large Language Models (LLMs) have demonstrated exceptional performance in diverse NLP tasks by leveraging massive datasets and computational resources, followed by fine-tuning or prompt-tuning for specific tasks Team et al. (2023); Dubey et al. (2024); Team et al. (2024). While many leading models are closed-source and heavily biased toward high-resource languages, particularly English, open-source alternatives have emerged with comparable performance. However, low-resource languages continue to face challenges due to limited pretraining data, high costs of creating linguistically rich corpora, and inefficiencies in tokenization and vocabulary. My main goal is to explore and find **efficient approaches** to **training LLMs** for **low-resource languages**. My previous research includes work on data-to-text generation, abstractive summarization, and linguistic entrainment in dialogue generation.

1.1 Building LLMs for Low-Resource Languages

My current research focuses on developing novel methods for training language models in low-resource language settings. I am particularly interested in designing generalizable approaches for training LLMs with limited data. One of my current works explores a modular training strategy that adapts multilingual models into monolingual ones.

In this work, we hypothesize that full model tuning is often unnecessary and propose a more modular approach. Specifically, our method involves first training a language-specific tokenizer and creating corresponding embeddings, followed by tuning only the non-embedding parameters. We conduct a comprehensive analysis across multiple scenarios, including multilingual-to-monolingual transfer and adaptation from high-resource to low-resource monolingual models. When applied to multilingual models, our method significantly reduces the number of tunable parameters and overall training time. Evaluation on natural language understanding (NLU) tasks shows that our approach improves performance when adapting mBERT to low-resource languages. Specifically, we use POS-tagging Nivre et al. (2020), NER Rahimi et al. (2019), and mask-filling task to evaluate our approach. We notice an increase of 1.5x in the mask-filling scores for our modular approach over full-tuning of the model, in the case of Scottish Gaelic language. The scores stay almost stagnant on the other two evaluation metrics. Additionally, we also experiment with replacing the embeddings with custom FastText embeddings, however, the results stay similar to the default configuration with mBERT embeddings. We perform similar experiments with Irish as well, which, unlike Scottish Gaelic, is present in the pre-training languages of mBERT. Consequently, using a randomly initialised transformer, we plan to investigate if the inclusion of a language in pretraining dataset helps with better initialisation of non-embedding representations for our modular training. We plan to further validate our method on decoder-based as well. We choose paraphrasing and summarization tasks to evaluate the models on their language generation capability. Since there is no gold data available, we plan to create silver test data using the NMT system and the available monolingual corpora. Specifically, for each data instance in the monolingual corpus, we will create its corresponding synthetic input using NMT systems and state-of-the-art English LLMs, depending on the downstream task.

1.2 Parameter-efficient training with Artificial Language Initialisation

In parallel, I am also working on parameter-efficient training methods through artificial language-based pretraining strategies. Prior studies Papadimitriou and Jurafsky (2020); Chiang and yi Lee (2022) demonstrate that models pretrained on non-linguistic data can achieve performance comparable to those trained on English sentences. We adapt the best performing approach followed by a parameter-efficient *pretraining* for language acquisition from limited data.

Our method initializes the model using token embeddings trained with a shallow model, followed by tuning only the non-embedding parameters on non-linguistic data to introduce structural biases. To initialize the model with structural biases, we experiment with pretraining on two artificial languages proposed by Papadimitriou and

Proceedings of the 1st Workshop for Young Researchers in Natural Language Generation, pages 21–23

Jurafsky (2023): the nested parentheses language (NEST) and the crossing parentheses language (CROSS). Subsequently, the model is frozen and further pretrained on the 10M-token BabyLM corpus using LoRA adapters. Experiments on small-scale dataset show that this approach leads to faster convergence while maintaining performance comparable to classic full-model pretraining.

1.3 Future Research

For my future research, I plan to enhance these approaches with linguistically informed curriculum training. A slightly more ambitious goal of my research plan is to develop multilingual language generation models by swapping the embeddings using existing embedding alignment approaches. Specifically, the non-embedding model parameters are treated as a universal language representation, while injecting language-specific information using just the embeddings. Additionally, I plan to experiment with Direct Preference Optimization (DPO) Rafailov et al. (2024) and delexicalized pretraining to further improve training efficiency in low-resource scenarios.

2 Questions for Senior Researchers

- How to land an internship and how to prepare for the interviews
- How should one usually decide whether a research idea should be pursued further, or when it's better to move on?
- Where do you think NLG research is heading in the next few years, and what areas seem most promising, especially in the era of LLMs?

3 Suggested topics for discussion

- Multilingual Text Generation: How important is the role of tokenization and vocabulary extension? How can we efficiently adapt the large vocabulary from a multilingual model to a language-specific monolingual model? How do we address catastrophic forgetting while finetuning a multilingual model on a lower-resource language?
- Data-efficient training of NLG systems: How reliable is training using synthetic data? How is the performance comparable to the gold data? How do we address problems for low-resource languages that are linguistically distant from any higher resource language? How can we evaluate NLG models for languages with minimal resources?

3.1 Acknowledgments

The study was supported by the Charles University, project GA UK No. 302425.

References

Cheng-Han Chiang and Hung yi Lee. 2022. On the transferability of pre-trained language models: A study from artificial datasets. https://arxiv.org/abs/2109.03537.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* .

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 4034–4043. https://aclanthology.org/2020.lrec-1.497.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, pages 6829–6839. https://doi.org/10.18653/v1/2020.emnlp-main.554.

Isabel Papadimitriou and Dan Jurafsky. 2023. Injecting structural hints: Using language models to study inductive biases in language learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 8402–8413. https://doi.org/10.18653/v1/2023.findings-emnlp.563.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 151–164. https://www.aclweb.org/anthology/P19-1015.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* .

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* .

Biographical sketch

Nalin Kumar is a first-year PhD student at the Institute of Formal and Applied Linguistics (UFAL), Charles University in Prague. He is working on exploring efficient methods for Natural Language Generation (NLG) systems for his dissertation. He has published a couple of papers in NLG, one of which was presented at NAACL-Findings 2024. His other works include topics like Neural Machine Translation and Sentiment Analysis. He has also worked in text-to-data and data-to-text generation in low-resource languages for his Master's thesis at UFAL.