Thesis Proposal: Efficient Methods for Natural Language Generation/Understanding Systems

Nalin Kumar
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czechia
nkumar@ufal.mff.cuni.cz

Abstract

While Large Language Models (LLMs) have shown remarkable performance in various Natural Language Processing (NLP) tasks, their effectiveness seems to be heavily biased toward high-resource languages. This proposal aims to address this gap by developing efficient training strategies for low-resource languages. We propose various techniques for efficient learning in simulated low-resource settings for English. We then plan to adapt these methods for low-resource languages. We plan to experiment with both natural language generation and understanding models. We evaluate the models on similar benchmarks as the BabyLM challenge for English. For other languages, we plan to use treebanks and translation techniques to create our own silver test set to evaluate the low-resource LMs.

1 Introduction

General-purpose Large Language Models (LLMs) have shown exceptional performance in various Natural Language Processing (NLP) tasks (Achiam et al., 2023; Team et al., 2023; Dubey et al., 2024; Team et al., 2024). This is made possible using an extensive amount of data and computational resources to train the model, and then further finetuning or prompt tuning on the specific task. However, many such models have huge numbers of parameters and are closed-source (FitzGerald et al., 2022; Li et al., 2024). To counter this, many open-source LLMs have been released with comparable performance. However, the performance of current LLMs has largely been restricted to high-resource languages, even more so only for English, as they are predominantly trained on English and other high-resource languages (Li et al., 2024).

The availability of an adequate pretraining dataset plays the most important role in developing any LLM. Cleaning and processing web-crawled data is a common way of getting monolingual and parallel datasets (Conneau et al., 2020; De Gibert et al., 2024; Tiedemann, 2009). However, getting such data can be quite challenging for languages with minimal web presence, especially for a specific domain or task. Recent works alleviate the issue by creating synthetic data using zero-shot NMT systems. These works mainly involve using English as a pivot language and transferring the knowledge to the target language. Although there tends to be a performance improvement using such noisy data in contrast to a zero-shot setting, the models' applicability is still debatable (Maheshwari et al., 2024) simply due to the lack of ground truth. To counter this, various challenges have been organized (Cripwell et al., 2023). There has also been an effort to create linguistically rich datasets (Nivre et al., 2016). However, creating such corpora is too costly, which limits the amount of available data instances. Consequently, challenges such as BabyLM (Warstadt et al., 2023) focus on efficient training with the least training instances but are English-only.

Thesis objectives The performance of the current LLMs is mainly limited to high- and moderate-resource languages. The primary objective is to develop new methods for training models for low-resource languages. To achieve this, we will develop general approaches to training LLMs in a low-resource setting, which will first be tested on English for ease of evaluation. We will then work on exploring ways to transfer the data knowledge and tuning strategies to any low-resource language. The thesis will cover both theoretical and experimental aspects of the problem while keeping the solutions linguistically oriented. The secondary goal of this thesis is to release data and produce models for several languages. Using the thesis output, we can work on various NLP tasks for non-English languages. The contribution of this thesis will be three-fold: (1) we will develop efficient pretraining

Proceedings of the IJCNLP-AACL 2025 Student Research Workshop, pages 209–217

strategies with limited data, (2) we will release the intermediate synthetic silver data, and (3) we will release the created models.

Thesis Structure The thesis is structured into two main halves. The first half is focused on experiments with English in a low-resource setting (Section 3.1). We propose various approaches suitable for low-resource language modeling. We will evaluate these approaches based on the evaluation metrics used by the BabyLM challenge. These approaches will then be adapted to actual low-resource languages, which constitute the other half. One major challenge is finding ways to evaluate such LMs. We use state-of-the-art NMT systems and existing dataset resources to tackle it. We discuss more about the datasets and evaluation in Section 4. We finally conclude the proposal in Section 5.

Research Questions To summarize we aim to answer our following primary research questions:

- How can we design efficient pretraining strategies that maximize performance with minimal data for low-resource languages?
- Can modular approaches be shown to work better than end-to-end training? How significant a role do the embeddings play?
- Does introducing semantics and syntax knowledge separately help with model training?
- Does delexicalized pretraining improve robustness to sparsity in named entities and rare words?
- How effective are Reinforcement Learning from Human Feedback (RLHF) methods in aligning outputs with human preferences when training data is scarce?

2 Background

2.1 Token Representation

The efficiency of token-level representation plays a significant role in model's performance. Since languages have different scripts, converting them to a common script can make the representation more efficient. There have been various works to study the effectiveness of transliteration in the context of low-resource languages. While transliteration can lead to loss of phonological and morphological accuracy along with other ambiguities, romanization of languages has been shown to improve cross-lingual alignment (Amrhein and Sennrich, 2020; Purkayastha et al., 2023; Liu et al., 2024), as the base models usually are primarily trained on Roman script. However, the performance of such methods is mainly dependent on the tasks, model size, and target languages (Ma et al., 2024).

2.2 Multilingual LLMs

Multilingual LLMs (MLLMs) are trained on almost all available data in various languages with the hypothesis that a deprived language would benefit from the cross-lingual transfer with the higher-resourced ones (Lin et al., 2024; Üstün et al., 2024). However, Wang et al. (2020) show a negative interference for both high and low-resource languages because of the presence of language-specific parameters. The sub-par performance of lower-resourced languages can mainly be attributed to the huge training data imbalance and inefficient vocabulary and tokenization. Consequently, monolingual models, or models trained on better-sampled data, often capture richer linguistic features, especially for lower-resourced languages (Feijo and Moreira, 2020; Xue et al., 2021; Armengol-Estapé et al., 2021; Huang et al., 2023). Furthermore, multilingual models may lack cultural awareness for the under-represented languages (Hämmerl et al., 2022; Zhang et al., 2024).

2.3 Vocabulary Extension

Another way to extrapolate the performance of higher-resourced languages is through vocabulary extension and further pretraining on specific languages. Zhao et al. (2024) show that further pretraining, or pre-finetuning, on merely 1% of the pretraining data for non-English significantly improves the performance. However, tuning the model parameters entirely on new data often leads to catastrophic forgetting (Luo et al., 2023). To alleviate the issue, Marchisio et al. (2023) considered extending the vocabulary and proposed data mixing strategies. Kim et al. (2024) shows that expanding vocabulary along with several steps of training strategies to tune the model parameters can efficiently improve the model performance on non-English languages. However, the improvement is often limited to closely related languages. As most of the current works on low-resource languages focus on cross-lingual transfer instead of efficient training strategies, we try to bridge this gap with our work by focusing more on the latter.

2.4 Instruction Tuning and RLHF

There have been numerous works that include instruction tuning and training on human feedback to generate outputs better aligned with human preference (Ouyang et al., 2022; Achiam et al., 2023; Touvron et al., 2023), the current multilingual setups are typically not instruction-tuned due to data scarcity, which limits their performance. Direct Preference Optimization (DPO) (Rafailov et al., 2024) is among the recent frameworks that optimize directly on user preference data without the need for a separate reward model. It has proved to be effective for high-resource languages, but its applicability to low-resource ones is still unknown.

3 Proposed Approaches

Following state-of-the-art approaches for LLM training, we will use the standard transformer architecture for our experiments while focusing primarily on data and training improvements. Specifically, we try to use the data more efficiently by leveraging linguistic annotation. We design our experiments in two steps: (1) we will first benchmark several methods on English, (2) we will transfer those strategies to low-resource languages. Additionally, we will also experiment with several other methods.

3.1 Experiments in English

For simpler evaluation, we will begin with working with the English language in a simulated low-resource setting. The primary goal is to optimize the amount of pretraining tokens used. Specifically, we will experiment with the following strategies for efficient training of English LMs:

1. **Curriculum Learning:** We will use various linguistic features to measure the complexity of the training instances and consequently feed the model simpler instances first, then gradually increase complexity (i.e. build the curriculum). This approach has widely been used in the submitted works at the BabyLM challenge (Chobey et al., 2023; Nguyen et al., 2024; Salhan et al., 2024; Saha et al., 2024). However, the majority of them only categorize the complexity on the dataset-level, due to which potential outliers can get overlooked, whereas in the thesis proposal, we plan to step up to a more fine-grained instance-level curriculum. Specifically, we calculate the complexity for each training instance on various linguistic levels, e.g., height/number of edges of the dependency tree, etc.

2. **Lexical learning using WordNet:** WordNet provides a hierarchical, lexically rich database of words and synonyms, enabling embedding training focused on word relationships. To boost the initial training stage of our models without using large-scale plain text data, we will first initialize the subword embeddings of the model using the WordNet embeddings as ground truth (Saedi et al., 2018). We then employ different strategies using the WordNet dataset to tune the embeddings further. For example, given a sentence, we replace one of the words using WordNet and further train the model to predict if the two sentences are similar or not.

3. **Syntactic learning using UD treebanks:** We plan to train the encoder on syntactic tasks like Parts-Of-Speech (POS) tagging, using a dataset like the UD treebanks (de Marneffe et al., 2021), which supports syntactically rich and structured text. We will explore syntactically relevant pretraining objectives, such as part-of-speech tagging or masked prediction. Two such examples are given below:
   - Predict POS tags from text, building a foundation in syntactic structure.
   - Predict masked POS tags (sequence-to-sequence of POS tags), focusing on syntactic dependencies.

4. **Delexicalized Pretraining:** Named entities in the training data can lead to sparsity problems in the input data. Consequently, lower-resource language models often struggle with named entities and numbers. To mitigate this issue, we delexicalise the named entities by replacing them with placeholders, focusing instead on the syntactic structure and grammatical relationships.

3.2 Experiments in low-resource languages

We adapt the tuning strategies from English to the low-resource languages. We also propose additional methods to train the models more efficiently. We plan to experiment with the following strategies:

1. **Shuffling:** We plan to experiment with more sophisticated sentence-level shuffling as our pretraining technique. We will propose a self-supervised method that focuses on reconstructing a shuffled input without altering the subject-verb-object order, akin to BART's objective (Lewis et al., 2019) but adapted for linguistic nuances. Additionally, we experiment with instruction-tuning as well.

2. **Transliteration:** Romanized transliteration has shown better transfer between related languages (Amrhein and Sennrich, 2020). However, it might lead to a loss of information on the morphological level. (Micallef et al., 2023) demonstrated that transliterating to the original script might improve the performance for that language. Thus, we will also experiment on the effect of transliteration for the selected low-resource languages.

3. **Lexical and Syntactic Learning:** If WordNet-enhanced embeddings and syntactic learning prove effective in English, we plan to extend the approach to other languages. Training data for syntactic learning (UD treebank) already exists for the considered languages. For lexical learning, we plan to use NMT systems to generate the candidates for each lexicon in the training data.

4. **Using encoder as assistant for efficient finetuning:** The current LLMs perform significantly well on English language. Using this to our advantage, we plan to use a multi-encoder for faster finetuning on a downstream task. Specifically, we use an additional English encoder to assist the model in finetuning on downstream tasks. We use NMT system for generating the input for the English encoder. Additionally, during the tuning process, we plan to gradually decrease the dependence on the assistant encoder.

5. **Multilingual LMs with language-specific word embeddings:** We also plan to train the embeddings agnostic of other model parameters and vice versa. We aim to get language-specific embeddings while the model parameters serving as a universal grammatical representation. To check the effectiveness, we plan to experiment with different number and combinations of languages, e.g., languages from the same family. Previous works have shown that the embeddings generated from similar techniques are isomorphic across languages (Vulić et al., 2020). Consequently, we plan to swap embeddings along with further small finetuning to build a low-resource LM.

6. **Direct Preference Optimization (DPO):** DPO has emerged as an alternative to RLHF. It aims to align the outputs to the human-preferred generations. This method can be applied to various sequence-to-sequence tasks, such as summarization, question answering, paraphrasing, and machine translation. We will create substandard samples using back-translation with English as the pivot language. We plan to apply this method for finetuning and instruction-tuning on downstream tasks. We investigate its applicability by integrating it with previously discussed methods.

7. **Curriculum Learning and Delexicalised Pretraining:** We will adopt similar strategies from the English language for the other low-resource languages.

We will consider Aya (Üstün et al., 2024) and mT5 (Xue et al., 2021) as our baseline models, both of which contain the considered languages in their pretraining data. Aya, with 13B parameters, serves as a strong baseline performing well on a wide range of language understanding and generation tasks. We will also train a vanilla language model for each considered language using BART-inspired self-supervised pretraining techniques.

4 Training Dataset, Evaluation and Early Experiments

We will work with English in a limited data setting and 5 other diverse low-resource languages. We consider 2 European languages Irish (ga) and Scottish Gaelic (gd), a Semitic language, Maltese (mt), an Indic language, Urdu (ur), and an African language, Swahili (sw), for our experiments. The choice of languages is motivated by the existence of appropriate evaluation datasets. We will use *CC-100*[1] (Wenzek et al., 2020) corpus for getting the monolingual data. To get parallel data for our experiments using English as the pivot language, we will be using the *OPUS*[2] corpus. Additionally,

[1] https://data.statmt.org/cc-100/
[2] https://opus.nlpl.eu/

	BERT				mBERT			
Training →	full		non-emb		full		non-emb	
emb ↓ vocab →	model	custom	model	custom	model	custom	model	custom
model	0.1520	0.3642	0.2180	0.5220	0.2446	0.4392	0.3160	0.5454
fasttext	-	0.4356	-	0.5288	-	0.3570	-	**0.5588**
random	0.1976	0.4000	0.2094	0.5004	0.2011	0.3430	0.2047	0.5341

Table 1: Evaluation results of BERT and mBERT trained for the Scottish Gaelic language with different training settings (Training), embedding initializations (emb.) and vocabularies (vocab.).

we will use the *UD treebanks* (available for all the considered languages) (Nivre et al., 2020) and *WordNet* (Miller, 1995) for English.

4.1 Evaluation

We plan to test our English models on the BLiMP benchmark to evaluate grammatical competence, especially in minimal token usage, which stresses the model's syntactic and semantic efficiency.

Evaluating low-resource LMs gets tricky due to the nonavailability of appropriate evaluation sets. We use zero-shot NMT systems to address this challenge. For most of our evaluation in low-resource, we use English as our pivot language to generate test sets from the available monolingual corpora. Previous work (Kumar et al., 2023) has shown that generating via English has better performance than direct generation. Thus, to evaluate the applicability of our general-purpose LMs in low-resource languages, we will perform evaluation on three types of tasks:

Generation tasks 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. Specifically, for a given data instance y_l in language l, we first translate y_l to English y_{en}. We use current English-centric LLMs to generate corresponding synthetic input (for summarization - longer sentence) in English (x_{en}). We translate it back to the target language l (x_l) to get a silver parallel data, while preserving the naturalness of the task outputs. Additionally, we will test our methods on the WebNLG dataset for *data-to-text* generation for the Irish language.

Single-input Understanding Tasks We will use UD treebanks for training and testing on the *POS tagging* task for all the languages. We will also create silver test data for *NER*. We follow a similar approach as the previous paragraph. We translate the sentences into English, classify the named entities, and transfer the labels back to the target language using cross-attention scores.

Input-pair tasks We will use *XNLI* to evaluate Swahili and Urdu models. For the other three languages, we evaluate them again on the synthetic test data using English as a pivot language.

4.2 Early Experiments and Results

To start off, 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 perform 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 the overall training time. We further evaluate the natural language understanding (NLU) models on the mask-filling task. We present the accuracy scores in Table 1. Training only the non-embedding parameters consistently yields better results, while using a custom tokenizer provides a significant performance boost. Additionally, mBERT performs slightly better than BERT, and FastText embeddings offer only minimal improvement.

We also experiment with 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. 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 performance comparable to classic full-model pretraining.

5 Conclusion

The thesis proposal outlines various approaches to tune the models efficiently. We discuss related literature and current challenges specific to language modeling for low-resource languages.

We propose several techniques for efficient tuning in a simulated low-resource setting for English. Specifically, we plan to use curriculum learning at both the instance and dataset levels. We also plan to evaluate the role of grammar-rich datasets in model training. Furthermore, we also propose a delexicalised pretraining method to address the challenge of data sparsity in low-resource scenarios. We plan to train and evaluate the models for both generation and understanding tasks.

We further extend these approaches to actual low-resource languages. Additionally, we also try modular approaches to train the model separately on different linguistic levels. We also propose an encoder-assisted finetuning method for faster convergence and better knowledge transfer from higher-resource languages. We also plan to use DPO for generating better-aligned outputs to humans for low-resource languages. We evaluate our proposed approaches on various tasks, depending on the availability of test sets. We also plan to generate silver test sets using NMT systems on evaluation sets from higher-resource languages.

Challenges

We identify the following challenges and possible alternatives for the proposed approaches:

- Failing to adapt WordNet dataset for low-resource languages: Since this method depends on the chosen NMT system (NLLB, in this case), the quality of the generated data can be inadequate. We mitigate this issue by checking with several other NMT systems (Üstün et al., 2024; Fan et al., 2021; Zhang et al., 2020); if nothing works, we plan to use the Wiki dataset for lexical training.
- Curriculum learning on data instance level could prove ineffective: While this is a low-level risk, curriculum learning has proven to be effective on the dataset level for English (Mi, 2023). Thus, we can alleviate the issue by applying similar techniques to non-English languages.
- Delexicalised Pretraining may prove ineffective: In case this doesn't work out, we plan to delexicalise only during the inference, as this has been proven beneficial for end-to-end task-oriented dialogue systems (Kulhánek et al., 2021).
- Failure of language-specific embeddings for multilingual LMs: We permanently integrate the additional encoder into the model instead of relying on its assistance only during fine-tuning.

Acknowledgments

This work was supported by the European Research Council (Grant agreement No. 101039303, NG-NLG) and Grant Agency of Charles University (Grant No. 302425), and used resources of the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chantal Amrhein and Rico Sennrich. 2020. On Romanization for model transfer between scripts in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2461–2469, Online. Association for Computational Linguistics.

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online. Association for Computational Linguistics.

Cheng-Han Chiang and Hung yi Lee. 2022. On the transferability of pre-trained language models: A study from artificial datasets. *Preprint*, arXiv:2109.03537.

Aryaman Chobey, Oliver Smith, Anzi Wang, and Grusha Prasad. 2023. Can training neural language models on a curriculum with developmentally plausible data improve alignment with human reading behavior? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 98–111, Singapore. Association for Computational Linguistics.

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.

Liam Cripwell, Anya Belz, Claire Gardent, Albert Gatt, Claudia Borg, Marthese Borg, John Judge, Michela Lorandi, Anna Nikiforovskaya, and William Soto Martinez. 2023. The 2023 WebNLG shared task on low resource languages. overview and evaluation results (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 55–66, Prague, Czech Republic. Association for Computational Linguistics.

Ona De Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer Van Der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, and 1 others. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Diego de Vargas Feijo and Viviane Pereira Moreira. 2020. Mono vs multilingual transformer-based models: a comparison across several language tasks. *arXiv preprint arXiv:2007.09757*.

Jack FitzGerald, Shankar Ananthakrishnan, Konstantine Arkoudas, Davide Bernardi, Abhishek Bhagia, Claudio Delli Bovi, Jin Cao, Rakesh Chada, Amit Chauhan, Luoxin Chen, and 1 others. 2022. Alexa teacher model: Pretraining and distilling multi-billion-parameter encoders for natural language understanding systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2893–2902.

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovickỳ, Alexander Fraser, and Kristian Kersting. 2022. Do multilingual language models capture differing moral norms? *arXiv preprint arXiv:2203.09904*.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1048–1056.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. 2024. Efficient and effective vocabulary expansion towards multilingual large language models. *arXiv preprint arXiv:2402.14714*.

Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. 2021. AuGPT: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 198–210, Online. Association for Computational Linguistics.

Nalin Kumar, Saad Obaid Ul Islam, and Ondřej Dušek. 2023. Better translation+ split and generate for multilingual rdf-to-text (webnlg 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 73–79.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages. *arXiv preprint arXiv:2404.11553*.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Ayyoob Imani, Orgest Xhelili, Haotian Ye, Chunlan Ma, François Yvon, and Hinrich Schütze. 2024. How transliterations improve crosslingual alignment. *arXiv preprint arXiv:2409.17326*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Chunlan Ma, Yihong Liu, Haotian Ye, and Hinrich Schütze. 2024. Exploring the role of transliteration in in-context learning for low-resource languages written in non-latin scripts. *arXiv preprint arXiv:2407.02320*.

Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. 2024. Efficacy of synthetic data as a benchmark. *arXiv preprint arXiv:2409.11968*.

Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. 2023. Mini-model adaptation: Efficiently extending pretrained models to new languages via aligned shallow training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5474–5490, Toronto, Canada. Association for Computational Linguistics.

Maggie Mi. 2023. Mmi01 at the BabyLM challenge: Linguistically motivated curriculum learning for pretraining in low-resource settings. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 269–278, Singapore. Association for Computational Linguistics.

Kurt Micallef, Fadhl Eryani, Nizar Habash, Houda Bouamor, and Claudia Borg. 2023. Exploring the impact of transliteration on NLP performance: Treating Maltese as an Arabic dialect. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 22–32, Toronto, Canada. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Hiep Nguyen, Lynn Yip, and Justin DeBenedetto. 2024. Automatic quality estimation for data selection and curriculum learning. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 212–220, Miami, FL, USA. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

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 *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online. Association for Computational Linguistics.

Sukannya Purkayastha, Sebastian Ruder, Jonas Pfeiffer, Iryna Gurevych, and Ivan Vulić. 2023. Romanization-based large-scale adaptation of multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7996–8005.

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.

Chakaveh Saedi, António Branco, João António Rodrigues, and João Silva. 2018. WordNet embeddings. In *Proceedings of the Third Workshop on Representation Learning for NLP*, pages 122–131, Melbourne, Australia. Association for Computational Linguistics.

Rohan Saha, Abrar Fahim, Alona Fyshe, and Alex Murphy. 2024. Exploring curriculum learning for vision-language tasks: A study on small-scale multimodal training. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 65–81, Miami, FL, USA. Association for Computational Linguistics.

Suchir Salhan, Richard Diehl Martinez, Zébulon Goriely, and Paula Buttery. 2024. Less is more: Pretraining cross-lingual small-scale language models with cognitively-plausible curriculum learning strategies. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 174–188, Miami, FL, USA. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 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, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors. 2023. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Singapore.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024. MC^2: Towards transparent and culturally-aware NLP for minority languages in China. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8832–8850, Bangkok, Thailand. Association for Computational Linguistics.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

A Example Appendix

This is an appendix.