Tutorial Proposal
Aligning Large Language Models to Low-Resource Languages

Nazar S. Beknazarov [†*] , Ahmet Üstün[‡ *] , Marzieh Fadaee[‡ *], Natalia Fedorova[†], Sergey Koshelev[†], Alisa Smirnova[†]

[*] Presenting speaker
[†] Toloka Belgrade, Toloka Lucerne, ‡Cohere for AI
ortofasfat@toloka.ai, {ahmet, marzieh}@cohere.com

Abstract

This tutorial unfolds as a comprehensive guide on how to collect data to align large language models (LLMs) to low-resource languages (LRL), a crucial yet under-explored area in computational linguistics. As our world becomes more and more connected, the need for scalable language technologies that can support all languages, particularly those that are less resourced, has grown. However, the scarcity of available data makes it challenging to pretrain and align language models for these languages, considering that machine translation is not capable of creating datasets of the required quality for LRL. This tutorial introduces a range of techniques combined into a pipeline that can produce high-quality data for a number of languages using Swahili as an example, but not limited to it. We offer strategies for gathering datasets for those languages, as well as whole guidance on establishing high-quality obtained data. Furthermore, we propose guidelines for aligning LLMs for this data. These methods allow to collect high-quality data, which can be used to align LLMs for LRL.

Keywords Large Language Models, Natural Language Processing , Low Resource Languages, African Languages, Deep Learning, Swahili

Introduction

The last few years have seen an extraordinary evolution in the sphere of Natural Language Processing (NLP). This has largely been propelled by the Transformer architecture (Vaswani et al. 2017), which has spurred a swift uptake and development of LLMs. The recent successes in fine-tuning these models through instruction and employing reinforcement learning from human feedback (RLHF) have sped up the deployment of these technologies from the research phase to real-world applications like dialogue and search engines. Notable examples of such applications are ChatGPT and GPT-4 (Schulman et al. 2022), universal chat assistants. It came into existence less than a year ago and is a direct offshoot of InstructGPT (Ouyang et al. 2022). The rise of ChatGPT as one of the most rapidly expanding technology products in contemporary history is quite notable (Hu et al. 2023). Along with it, an array of proprietary models have also drawn a great deal of interest due to their remarkable performance. These include Vicuna, Falcon, Palm, etc. Despite these advancements, a significant gap remains. Most of these models are designed for English or the top 20 languages globally, leaving out a plethora of African and Asian languages that are crucial in regions most in need of AI and LLM technologies. Even the models that claim to support 100+ languages often fall short when applied to our target LRLs. This is where conventional fine-tuning methods hit a roadblock, proving ineffectual for generating LLMs for these languages. Our tutorial aims to fill this void by guiding audience through the process of crowd-sourced data collection and demonstrating how to effectively align LLMs for a range of LRLs available in Toloka.

Description and outline

Our tutorial is divided into two parts: gathering data for LRL, presented by Toloka, and fine-tuning an LLM with this data, presented by Cohere For AI.

Outline:

Introduction	10 min
Gathering data for LRLs	35 min
Instruction Fine-tuning on LLM for LRLs	35 min
Hands-on Demonstration	25 min
Total	105 min
	Quarter day

Introduction (10 min, Nazar)

In this section, we will discuss the latest trends in LLMs, and how it's affecting job markets worldwide. We will also talk about how LLMs are not evenly spread across the globe. We will especially look at the comparison of how many different languages are represented in the data sets that are open to the public. Here we are going to note that the huge difference in the amount of data available for different languages means that LLMs function very differently depending on the language. We will demonstrate how the current state-of-the-art LLMs are limited when it comes to the language we have opted to use as our example - Swahili.

Gathering data for Low-Resource Languages (35 min, Nazar)

Here we will start by covering basic concepts of crowdsourcing such as decomposition, quality control, and inter-

face design. We will introduce the task of gathering data for LRL and suggest a crowd-sourcing pipeline that we will implement during the tutorial and give the participants the guidelines on how to replicate it themselves. The pipeline will consist of two parts: collecting answers for the given prompts and validating the answers.

Part I: Collecting Answers We will set up a project where crowd annotators will be writing answers for a previously translated set of prompts in Swahili. We will say that, given the constraints of language translation, it's generally not advisable to rely on translators for translating answers. However, since the quality of the prompt isn't a critical factor, using translators for translating questions is acceptable. We will say that, we have chosen the Swahili language as a modeled one, but our pipeline is not limited by it and can perform on a wider range of LRL. We will show the practical implementation of the crowd-sourcing concepts introduced in the beginning as well as a text generation-specific quality control technique of post-acceptance. This technique allows us to accept relevant answers and reject irrelevant ones based on the results of crowd validation in the next part.

Part II: Answer Validation In this part, we will show how to validate the answers written in Part I with the help of crowd annotators. This is an essential step for LRL. Although various metrics such as Harmlessness, Helpfulness, and Truthfulness are available for evaluation, we will focus solely on assessing Helpfulness for this real-time demonstration, given time constraints. Like in Part I, we will launch a project in real-time and show the practical application of quality control techniques for collecting such kind of data. Based on these results we will accept or reject the answers written in Part I and get a dataset of prompts and answers in Swahili.

Instruction Fine-tuning on LLM for Low Resource Languages (35 min, Ahmet)

Here we will talk about how to align existing LLM on the assessed dataset in LRL by using an instruction fine-tuning pipeline. This session will start with a brief overview of previous work on multilingual instruction tuning, including the existing data, models, and results.

Following the overview, we will fine-tune an LLM using the collected dataset. To fine-tune a multilingual LLM through instruction, the process involves several stages. These stages encompass selecting the most suitable base model and establishing a fine-tuning workflow. The objective is to outline the best practices for instruction fine-tuning an LLM, particularly for LRL.

Firstly, we will select a language model or an instruction-tuned model. This selection is based on factors such as language coverage, the parameter size, and the compute budget available for the fine-tuning process. Options include open-source models such as Aya (if released by the time of the conference) or in-house models such as Cohere. Next, we use previously collected dataset. The dataset's items are selected based on two key considerations: the linguistic diversity, especially if the fine-tuning process involves multiple languages, and the prompt completion rankings, which serve to pinpoint higher-quality examples that are conducive for fine-tuning the Language Learning Model (LLM)..

Finally, the fine-tuning workflow will be set up by establishing the necessary infrastructure and configuring the parameters for the fine-tuning and evaluation process. When evaluating an instruction fine-tuned LLM on multilingual benchmarks, various approaches can be employed. These approaches include utilizing academic multilingual benchmarks from zero-shot datasets, using held-out prompt/completion pairs from collected datasets, and human evaluation by ranking different models. Here, we aim to assess the performance and effectiveness of the instruction-tuned LLMs under different distributions of data. In this tutorial, we will evaluate the resulting fine-tuned LLM by using multilingual benchmarks that include LRL.

Hands-on Demonstration (25 min, Marzieh)

The second part of the tutorial which focuses on instruction fine-tuning and model evaluation will be accompanied by hands-on material that includes a python notebook demonstrating a simple fine-tuning and model evaluation pipeline.

During this part, we will offer participants to familiarize themselves with fine-tuning data by themselves. After this, the speaker will perform model fine-tuning and evaluation. Based on our language selection –Swahili–, we will fine-tune a publicly available multilingual T5 (Xue et al. 2020). This language model is a pretrained LLM that is trained on 101 languages and shows strong performance upon fine-tuning. Our hands-on tutorial will allow participants to fine-tune different sizes of mT5. We will also provide an evaluation suite consisting of XNLI, XCOPA, and XStory Close that includes the Swahili language. This hands-on demonstration will give all participants practical knowledge of how to align an LLM by themselves.

Goals

Given the limited availability of datasets for LRLs, it is anticipated that merely fine-tuning a pre-trained LLM in English will be insufficient for delivering satisfactory chat-assistent. Therefore, additional fine-tuning datasets in the target LRL are essential. The primary objective of this tutorial is to equip attendees with the skills necessary to assess and curate high-quality data for LRLs, such as Swahili. We aim to teach attendees with skills to align LLMs to these LRLs. The tutorial provides attendees with knowledge on filtering and training annotators, crafting intuitive tasks, establishing a fair pricing model, and whole aligning process. The expected learning outcomes encompass acquiring knowledge of techniques for creating the whole crowd-sourced data-gathering pipeline and how to align LLM on obtained data.

Who is the target audience?

Our target audience comprises researchers and practitioners in the field of NLP. These individuals are keen on both enhancing the functionalities of existing Large Language Models and exploring the potential of information gathering via crowd-sourcing. **What will the audience walk away with?** Participants will acquire practical knowledge on how

to curate datasets specifically tailored for LLM through the use of crowd-sourcing techniques. They will develop a comprehensive understanding of each aspect of this process, from generating prompts to evaluating responses for subsequent applications, whether it is for fine-tuning LLMs or utilizing them in RLHF approaches. Additionally, attendees will gain insights into the ethical implications of data biases and discover strategies to effectively mitigate them. **What makes the topic innovative?** What sets this workshop apart is its dual focus on technological alignability and linguistic inclusivity on previously unapproachable languages. Additionally, the approach to directly source data from communities with unique cultural and linguistic backgrounds helps in not only preserving these languages but also in ensuring a more unbiased representation. **What core concepts, methods, and modeling frameworks will be conveyed?** The workshop will explore key principles of data gathering if your model language is not well resourced. Especially bias-mitigation techniques, prompt gathering pipelines and LLM pretraining and fine-tuning. We will primarily utilize HuggingFace Transformers and PyTorch as our frameworks.

Tutorials History

This tutorial is partly based on joint research on the evaluation of machine translation of African languages that also included work with LRL (Adelani et al. 2022). Toloka Research has offered a series of related tutorials at multiple related conferences: ICML (Lambert et al. 2023) (apr. 300 participants), ECIR (apr. 20 participants), RecSys (Ustalov, Fedorova, and Pavlichenko 2022) (apr. 70 participants), ICWE (Ustalov et al. 2022) (apr. 20 participants), NAACL- HLT, TheWebConf, SIGMOD/PODS (Drutsa et al. 2020a) (apr. 70 participants), WSDM (Drutsa et al. 2020b) (apr. 50 participants), and KDD (apr. 20 participants). Every time we adapted our content and practice session to the conference, e.g., intent classification for ECIR '23, end-to-end recommendations for RecSys '22, e-commerce for ICWE '22, WWW '21, and WSDM '20, audio transcription for NAACL-HLT '21, and image segmentation for SIG- MOD-/PODS '20 and KDD '19.

Estimated number of participants

We are expecting a moderate to high level of engagement for this tutorial, with an estimated attendance ranging between 50 and 100 listeners.

Prerequisite knowledge

We expect from our potential participant to have basic understanding of NLP and LLM area, familiarity with frameworks like HuggingFace transformers & PyTorch. Some prior experience with data collection and annotation could be beneficial but is not mandatory.

Supplementary materials

Our previous tutorials on related or connected themes:

- Slides from our previous tutorial at ICML, which was particularly crowded: https://zenodo.org/record/8186168
- Slides from ECIR: https://toloka.ai/events/ecir-tutorial-2023/
- Slides from RecSys: https://toloka.ai/events/recsys-2022/

References

Adelani, D.; Alam, M. M. I.; Anastasopoulos, A.; Bhagia, A.; Costa-jussà, M. R.; Dodge, J.; Faisal, F.; Federmann, C.; Fedorova, N.; Guzmán, F.; et al. 2022. Findings of the WMT'22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 773–800.

Drutsa, A.; Fedorova, V.; Ustalov, D.; Megorskaya, O.; Zerminova, E.; and Baidakova, D. 2020a. Crowdsourcing practice for efficient data labeling: Aggregation, incremental relabeling, and pricing. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2623–2627.

Drutsa, A.; Fedorova, V.; Ustalov, D.; Megorskaya, O.; Zerminova, E.; and Baidakova, D. 2020b. Practice of efficient data collection via crowdsourcing: Aggregation, incremental relabelling, and pricing. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, 873–876.

Hu, K.; et al. 2023. ChatGPT sets record for fastest-growing user base-analyst note. *Reuters*, 12: 2023.

Lambert, N.; Dmitry, U.; Pavlichenko, N.; Ryabinin, M.; Rajani, N.; Tunstall, L.; and Koshelev, S. 2023. Reinforcement Learning from Human Feedback: A Tutorial. In *Proceedings of the 2023 ICML.*

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Schulman, J.; Zoph, B.; Kim, C.; Hilton, J.; Menick, J.; Weng, J.; Uribe, J. F. C.; Fedus, L.; Metz, L.; Pokorny, M.; et al. 2022. ChatGPT: Optimizing language models for dialogue. *OpenAI blog*.

Ustalov, D.; Fedorova, N.; and Pavlichenko, N. 2022. Improving Recommender Systems with Human-in-the-Loop. In *Proceedings of the 16th ACM Conference on Recommender Systems*, 708–709.

Ustalov, D.; Pavlichenko, N.; Tseytlin, B.; Baidakova, D.; and Drutsa, A. 2022. Web Engineering with Human-in-the-Loop. In *International Conference on Web Engineering*, 505–508. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; and Raffel, C. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.