Founding a Framework for Accessible and Reliable Information Retrieval Across the World

Duygu Ataman
New York University

David Adelani
McGill University

Abstract

Access to reliable information is crucial for empowering individuals and creating knowledgeable and equitable societies. Online platforms like Wikipedia play a pivotal role in this regard by allowing humans collaboratively to collect and refine diverse knowledge. However, this process is often challenging across languages and geographic locations, emphasizing the need for automatic information retrieval (IR) systems. Recent advances, particularly Large Language Models (LLMs), hold great potential in shaping applications in vital social domains. Despite their capabilities, the lack of verification in constructing these models raises concerns about the reliability of their outputs. This project addresses the absence of an evaluation methodology for testing LLMs' accuracy across languages. Building on recent efforts, we propose creating the first multilingual factual IR (MFIR) benchmark using Wikipedia, supported by the Wikimedia Foundation, ensuring the reliability of automatic IR systems based on LLMs. This initiative includes organizing an international competition within the Multilingual Representation Learning (MRL) Workshop to also help raise awareness on this critical topic.

Introduction

Wikipedia, with its vast global readership, is a prominent supporter of open and reliable information access. However, maintaining the balance and accessibility of content across languages poses a challenge. Translating diverse topics or trying to create new content can be costly, hindering efforts to mitigate this inequality. Recent advancements in LLMs have transformed IR, showing promise in natural language processing (NLP) tasks (Garcia, 2022, Wei, 2022, Mao, 2023). However, LLMs face challenges in applicability to underrepresented languages and factuality. Initiatives for standardized benchmarks exist, but many lack human-prepared data for information access (Ruder, 2023). As organizers of the MRL at EMNLP 2023, we recently initiated the 1st shared task on Multi-lingual Multi-task Information Retrieval (MMIR) (Tinner et al., 2023) where we annotated Wikipedia articles in seven languages with varying degrees of resources and linguistic typology: Azerbaijani, Turkish, Uzbek, Igbo, Yoruba, Indonesian, and Swiss German in two tasks crucial for IR: named entity recognition (NER) and reading comprehension (RC). With support from Wikimedia, we hope to extend this benchmark to increase the amount of examples and provide annotations in other tasks useful for IR. With more advancement in technology, we believe our benchmark will remain essential for supporting fairness and applicability in IR systems.

Dates*: June 1, 2024 - June 30, 2025.*

Related work

IR involves extracting relevant information from a diverse data collection with the objective of matching user queries with the most pertinent resources, related to tasks such as **NER**, **RC**, **entity linking (EL)** and **fact verification (FV)**. NER classifies phrases referring to predefined categories (e.g., persons or organizations) and is crucial for many applications such as knowledge verification or localization, whereas EL connects entity mentions to a knowledge base. Beyond structuring the information, querying these later to infer relevant and factual answers to a given question is very important for the final task of IR, modeled with the RC and FV tasks. Notably, a comprehensive benchmark with evaluation data in multiple languages was lacking until our efforts.

Methods

Our benchmark is based on the textual data provided on Wikimedia downloads. In the pilot study for constructing the evaluation sets we sampled around 200 articles in diverse topics and picked lengthy paragraphs to assess RC and NER. The annotations are divided to around 100 per validation and testing purposes. Due to the high quality of the current data sets and the verified level of the challenge the task presents, we intend to expand the dataset to include training data for RC and NER, and also add the tasks of EL and factual verification.

Expected output

Our expected outcome from this project is the new MFIR benchmark for accessible and reliable IR. We will organize a shared task in the 2024-2025 edition of the MRL Workshop and promote collaboration on this competition with a public campaign. We will collect our findings in a paper which will be presented at the workshop.

Risks

One risk we anticipate is the unavailability of plenty of Wikipedia articles in some of the selected languages. However, we believe that even the few paragraphs available for these languages can be a valuable contribution to assessing the performance of LLMs in these languages.

Community impact plan

As chairs of MRL Workshop 2024, we aim to collaborate with our organizing and program committees comprising researchers from academia and industry specializing in low-resourced NLP. Our evaluation campaign and findings will be disseminated within the community actively engaged in our venue, supporting their ongoing research endeavors. Significantly, we aspire to achieve the first state-of-the-art results on the applicability of leading NLP methods to languages and dialects previously excluded from benchmarks, offering valuable insights for their integration into future technologies.

Evaluation

The accuracy of all annotations will be evaluated with cross-validation among annotators and the overall data sets will be examined for comparability and representativeness. The main objective of the resulting data set will be its comprehensiveness and through evaluation of LLMs in extracting accurate information using only the defined context. Therefore, the results of the competition will directly allow us to assess the level of challenge the new benchmark presents.

Budget

We approximate for a sufficient number of articles to be annotated in each language, annotation costs to be at least 3,000 USD. Therefore we would appreciate **25,000 USD** ($21,000 USD for annotation, including tool cost and administrative overhead).

Prior contributions

Both PIs have significant contributions to improving diversity and inclusion in NLP.
Duygu Ataman is an Assistant Professor at New York University and has been the co-organizer of the MRL Workshop at EMNLP since 2021. She also founded the ACL Special Interest Group on Turkic Languages (SIGTURK).
David Adelani is an incoming Assistant Professor at McGill University, core academic member at Mila - Quebec AI Institute, and a DeepMind Academic Fellow at University College London. He is a co-organizer for MRL Workshop at EMNLP for 2023 and 2024, and a co-organizer of AfricaNLP @ICLR for 2022-2024.

References

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., ... & Osei, S. (2021). MasakhaNER: Named Entity Recognition for African Languages.

Doddapaneni, S., Aralikatte, R., Ramesh, G., Goyal, S., Khapra, M. M., Kunchukuttan, A., & Kumar, P. (2023). Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 12402-12426).

Garcia, X., Bansal, Y., Cherry, C., Foster, G., Krikun, M., Johnson, M., & Firat, O. (2023). The unreasonable effectiveness of few-shot learning for machine translation. In the International *Conference on Machine Learning* (pp. 10867-10878). PMLR.

Mao, J., Jiang, W., Liu, H., Wang, X., & Lyu, Y. (2023, June). Inferential knowledge-enhanced integrated reasoning for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 11, pp. 13380-13388).

Ruder, S., Clark, J. H., Gutkin, A., Kale, M., Ma, M., Nicosia, M., ... & Talukdar, P. (2023). XTREME-UP: A User-Centric Scarce-Data Benchmark for Under-Represented Languages. *arXiv e-prints*, arXiv-2305.

Tinner, F., Adelani, D. I., Emezue, C., Hajili, M., Goldman, O., Adilazuarda, M. F., ... & Ataman, D. (2023). Findings of the 1st Shared Task on Multi-lingual Multi-task Information Retrieval at MRL 2023. *In Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)* (pp. 310-323).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824-24837.