Research Statement

Xinyu Zhang (Crystina) `xinyucrystina.zhang@uwaterloo.ca`

Information access is a fundamental human right. The Universal Declaration of Human Rights by the United Nations articulates that "everyone has the right to freedom of opinion and expression", which includes the right "to seek, receive, and impart information and ideas through any media and regardless of frontiers" (Article 19). Information access systems are evolving at an unprecedented pace, driven by advancements in search algorithms, and recently, by generative AI models with sophisticated reasoning and instruction-following capabilities. These innovations have transformed how users speaking high-resource languages (especially English) access, understand, and interact with vast stores of knowledge. However, for low-resource or even medium-resource languages, the development of these systems has not kept pace, with challenges including a lack of high-quality datasets and limited understanding of the internals of multilingual models. This disparity not only hinders individuals' benefit from recent advancements, but also exacerbates language inequalities in knowledge access.

Driven by the mission of equal information access, as well as my interest in linguistics and psychology, my research focuses on **enabling and enhancing search systems on diverse multilingual data**. Search is an indispensable component of information access systems: Not only has it served as the primary tool for information access over the past decades, but also forms the foundation for generative models to produce factual answers via retrieval-augmented generation (RAG).[*] My work advances multilingual search systems from three aspects: constructing high-quality multilingual *data*, enhancing the model *training*, and *understanding* the model internals:

(1) *Data.* High-quality training data are fundamental for building multilingual models, and evaluation data are fundamental for understanding model capacity. I constructed the *first two large-scale human-annotated multilingual datasets for neural retrieval models* (**Mr. TyDi** and **MIRACL**). To foster community use, I hosted the WSDM Cup 2023 Challenge on MIRACL, receiving over *400 submissions from 46 unique teams.*

(2) *Training.* The involvement of multiple languages results in a much larger design space for model training. Leveraging high-quality multilingual datasets, I performed the first comprehensive study of training practices for multilingual *embedding models*, which encode queries and documents into vectors and thus convert the text search problem into the problem of nearest neighbor search in vector space (Figure 1). The paper provides concrete guidelines for each training stage from the perspective of data augmentation, knowledge distillation, model architecture, and so on, which *doubles zero-shot effectiveness* and *improves in-domain effectiveness by 15%.*

(3) *Understanding.* Higher-quality training data and optimized training practices bring higher scores on downstream tasks, yet they do not provide insights about model internals. My recent work [18] revealed that language models' understanding capabilities can largely be attributed to word-level semantic information, which can be shared across languages that are even in different scripts.[†]

My work has been widely used by the community for building and evaluating multilingual models, including OpenAI,[‡] Microsoft [2, 11, 12], Google DeepMind [5, 4], Cohere,[§] Alibaba [14, 13], HuggingFace [7], to list a few in the industry, and my work has also been recognized by researchers from diverse language backgrounds, including countries such as France, Brazil, China, Japan, Korea, Canada, United States, etc. As a faculty member, I am committed to continue my research with the ultimate goal of **building high-quality information access systems that are inexpensive and practical to develop for each language, enabling widespread accessibility.**

High-quality Multilingual Training Data

Since 2020, embedding models [3] have brought new opportunities to information retrieval, but they also pose new challenges for model training since they require large amounts of high-quality data and computational resources. These challenges are especially severe in the multilingual scenario: Traditionally, multilingual data in retrieval are designed for evaluation, which only requires dozens of questions and is thus far from sufficient for training embedding models.

To enable studies in this direction, I led two projects on constructing high-quality multilingual datasets for embedding models, namely **Mr. TyDi** [15] and **MIRACL** [17]. Mr. TyDi is the *first* monolingual retrieval dataset with sufficient data for training PLM-based retrievers models, covering 11 typologically diverse languages. Up to the date of writing, **Mr. TyDi has received over *84k* downloads on HuggingFace, and the associated models have received over *1M* downloads in total.** While Mr. TyDi took the first step towards building large-scale training and evaluation datasets for multilingual retrieval, we acknowledged it was limited by sparse labeling, which results in potential false negative in the evaluation. This motivated us to build MIRACL, a larger dataset with more comprehensive labeling and language coverage.

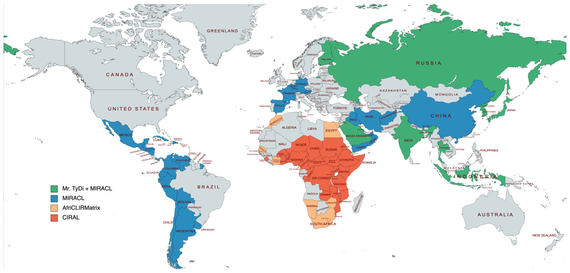

Figure 1. Geographical coverage of languages included in Mr. TYDI (green), MIRACL (green and blue), AfriCLIRMatrix (yellow), and CIRAL (orange).

[*] OpenAI enabled online search in ChatGPT: https://x.com/sama/status/1852041075793522911.
[†] For example, "tomato" and "tomate" refers to the same object in English and Spanish.
[‡] https://openai.com/index/new-embedding-models-and-api-updates/
[§] https://cohere.com/blog/introducing-embed-v3

In MIRACL, questions are re-annotated by native speakers and increased by 10 times compared to Mr. TyDi. This large label set provides a much more reliable training and evaluation basis for the study of multilingual retrieval. Additionally, we expand the language coverage to 18 languages, so that MIRACL not only covers typologically diverse languages but also similar languages (Figure 1). This allows the study of language transfer by language family, linguistic feature, geographic distance, and so on. It results in *600 thousand* annotations pairs in total, all annotated and verified by native speakers of the corresponding language. **All together, MIRACL took five person-years of annotation time, making it the largest human-annotated multilingual retrieval dataset with diverse language coverage and a widely-used training resource and evaluation benchmark.**

In addition to leading dataset construction, I mentored and assisted relevant work focusing on slightly different aspects: AfriCLIRMatrix [8] and CIRAL [1] on retrieval datasets specifically for African languages, NoMIRACL [10] for evaluating the hallucination of multilingual retrieval-augmented generation, and FoodieQA [6] for evaluating the foodie-culture knowledge of multi-modal models. For all datasets above, we provide reproducible robust baselines and resources including releasing the code, checkpoints, and pre-built indexes to support the community in replicating our results at any level.

Training Multilingual Embedding Models

While constructing Mr. TyDi, we found that the vanilla multilingual embedding models only perform similarly with the classic lexical matching algorithm BM25 [9], which was developed decades ago. How could we build more effective multilingual embedding models? High-quality multilingual datasets allow us to better approach this question.

In Zhang et al. [16], **I perform the first comprehensive study of training practices for multilingual embedding models**, covering a wide range of the design space (Figure 2). The study covers all the stages in the embedding training pipeline: language model pretraining (PT), continual pretraining (Cont. PT), pre-fine-tuning (Pre-FT), and fine-tuning. For each stage, we compared different model options (e.g., at a high-level, multilingual models versus monolingual models), data options (e.g., synthetic data designed for task-specific or language-specific augmentation, data in all languages or only the target language), as well as training strategies (e.g., knowledge distillation). **Overall, we** *doubled* **zero-shot effectiveness over the vanilla multilingual embedding models, and improved in-domain effectiveness by** *15%*.

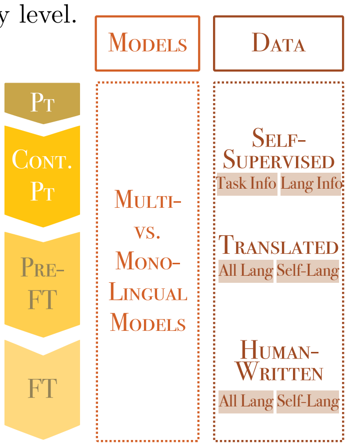

*Figure 2. Multilingual Embedding Models Training Pipeline. (**PT**: pre-training, **FT**: fine-tuning)*

Zhang et al. [16] was also the first to reveal the surprising cross-lingual transfer effect of embedding models that naturally emerges from multilingual language models: transfer learning can be achieved even between languages written in different scripts that share no natural subword units. For example, transferring from Finnish to Telugu achieves 72% of the effectiveness of an in-domain Telugu model. This indicates that transfer may substantially rely on semantic understanding of the input sentences, which drives further research on the understanding of multilingual models.

Understanding Multilingual Models

If cross-lingual transfer relies on deeper semantic understanding, at which level would semantic understanding and transferring happen? Empirically, human perception of a sentence is robust regarding interchanging words under the same or similar semantic concept (Figure 3). Motivated by this observation and intuition, my recent work [18] investigates how much shared semantics at the word-level enable language understanding and transferability.

In language models, natural-language words are mapped into a continuous vector space by the embedding layer, where each word has its own vector representation (embedding). As embedding vectors ideally capture the semantic meaning of each word, we group the vocabulary based on their word embedding similarity, where the grouped words form so-called "semantic tokens" and share the same "semantic embedding vector". That is, instead of perceiving the input sentence with word-level details, the language models now perceive the inputs at the semantic level. Results on diverse benchmarks and languages show that **language understanding and transfer could largely be achieved at the subword level**: With multiple multilingual language models, model variants with semantic tokens in 5%–20% of the vocabulary size could achieve over 85% effectiveness on downstream multilingual tasks. This finding suggests that the stage of "word recognition" (i.e., mapping the text into their semantic meanings) can be largely achieved at the subword level, and be separated from the stage of "understanding" based on lexical semantics.

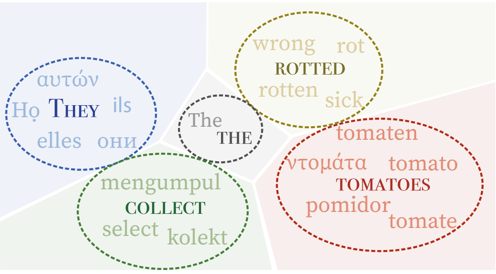

Figure 3. Words that fall under similar semantic concepts (as indicated by the colors), surrounding the keywords from sentence "They collected the rotted tomatoes."

Seeking a better understanding of multilingual models and training is a common theme in my research work. For example, one of the most important considerations when designing MIRACL is to include languages in a wide similarity spectrum in terms of the language family and linguistic diversity, which helps us discover their impact on the transfer effect. [17]

[§] https://openai.com/index/new-embedding-models-and-api-updates/

[§] https://cohere.com/blog/introducing-embed-v3

[§] "tomate": "tomato" in Spanish

Future Research

My past research experience convince me that better multilingual information access system eventually requires better multilingual pretrained language models. Looking forward, I am interested in studying alternative design of building multilingual models with the aim of *easier adaptation to new languages with minimal data requirement*. The current paradigm for building multilingual models is heavily data-centric, relying on scaling up from English to include a broader range of languages. This poses several challenges, for example, how can we effectively incorporate languages with less resources, especially when the model scales up and the data requirement increases? Even if such data is available, re-training the model for all the languages or specific languages is prohibitively expensive, creating significant barriers for institutions and researchers aiming to extend models to more languages that are not supported by English-centric data. To address these challenges, my future research will focus on the following directions and strategies.

Language-specific versus Language-agnostic Modelling. The multilingual model are often built by scaling the data from English-centric data to higher coverage of languages in order to achieve better multilingual capabilities. This is analogous to how human learn the native languages, where we learn our native language alongside cognitive and logical development. However, acquiring a second language typically follows a different pathway: we learn how to *perceive* (read, listen) and *articulate* (write, talk) the new language, but we do not re-learn how to *understand* or *reason* in the new language. Could similar intuition be applied to the modeling of multilingual models? For example, the models requires language-specific knowledge on perceiving the words in the new language (mapping the real text to their semantic meanings) and articulating the "thought" using the new language, yet the understanding based on the semantic meanings and the reasoning based on the understanding are more language-agnostic. The findings in Zhang et al. [18] suggest a separation between the word perception and the semantic understanding with encoder models. Could this finding be extended to generative models and their reasoning processes? If so, how might this insight reshape the design of multilingual models?

On-demand data requirement. If the hypothesis about language-specific and language-agnostic modeling holds true, it could open doors to more efficient multilingual data utilization and cross-lingual transfer. Intuitively, language-specific knowledge (e.g., word perception and thought articulation) would require additional adaptation and learning, while language-agnostic capabilities (e.g., understanding and decision-making) might leverage shared representations, with less requirement on data and computational resources. Is this hypothesis valid? If so, how can we design multilingual models that explicitly separate data and computation between these two aspects? Exploring this question potentially leads to a more scalable and efficient framework for building and expanding multilingual systems.

Multi-perceptional Modelling. Taking a step back and focus only on the "perception" stage: the current design of the language models primarily focus on perceiving written text, sometimes augmented with visual signals. However, how humans learn and transfer knowledge across languages highlights a more nuanced interplay of perceptual channels. When an English speaker learns French, the knowledge transfer may be largely based on the common spellings of words, while the listening and pronunciation require re-learning. This pattern is reflected in current tokenization and modeling approaches. However, the dynamics differ for some of the other language pairs. For example, when a Chinese speaker learns Korean, the knowledge transfer may primarily base on phonetic similarities while the word spellings and writing systems need to be completely re-learned given that the two languages are in distinct scripts. This observation motivates the concept of *multi-perceptual modeling*, where the models are designed to perceive the languages diverse perceptual channels beyond text.[¶] Such design may enable models to better capture the nuanced ways in which languages relate, enhancing cross-lingual learning and adaptation, yet how can we effectively represent the different perceptual similarities and leverage them across languages dynamically?

On-demand perceptual channel interaction and data requirement. Similarly, if the hypothesis on the multi-perceptional modeling holds true, it may reshape the multilingual data utilization and cross-lingual transfer. For example, language pairs with strong phonetic similarities but different scripts might benefit significantly from phonetic-based modeling. It also opens up multiple intriguing directions: Ideally, a model could dynamically determine which perceptual channels to rely on based on the source and target languages, optimizing cross-lingual transfer. But does the dynamic interaction work as expected in practice? Could the channels not only interact with but also reinforce from each other? Falling back to the initial motivation, could it lead to more efficient training process with less data and computation requirement, with better scalability across a diverse range of languages?

Low-resource Languages. The ultimate goal of the above directions is to deploy these methodologies in real-world applications for low-resource languages, where the challenges exist beyond the technological development, but also the deployment and accessibility within local communities, as well as the unanticipated obstacles that arise in real-world usage. Looking ahead, I aim to establish long-term collaborations with indigenous and marginalized communities to co-develop the necessary resources and infrastructure for their languages. Through such partnerships, I hope to not only address the known technical barriers but also to explore the unknown challenges of real-world adoption, which ultimately empowers language speakers with the choice to access information in their native languages with the solutions aligned with their cultural and linguistic needs.

In embarking on an academic career, I believe these research directions will have a long-term impact on future decisions about how to incorporate diverse multilingual data and knowledge into the language models, which is of paramount importance for the future of easy-to-build multilingual information access systems.

References

[1] Mofetoluwa Adeyemi, Akintunde Oladipo, Xinyu Zhang, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Boxing Chen, Abdul-Hakeem Omotayo, Idris Abdulmumin, Naome A. Etori, Toyib Babatunde Musa, Samuel Fanijo, Oluwabusayo Olufunke Awoyomi, Saheed Abdullahi

[¶] Importantly, these perceptual channels do not necessarily equate to modalities. For example, phonetic signals could be represented using the International Phonetic Alphabet (IPA) instead of relying on raw audio data.

Salahudeen, Labaran Adamu Mohammed, Daud Olamide Abolade, Falalu Ibrahim Lawan, Maryam Sabo Abubakar, Ruqayya Nasir Iro, Amina Imam Abubakar, Shafie Abdi Mohamed, Hanad Mohamud Mohamed, Tunde Oluwaseyi Ajayi, and Jimmy Lin. Ciral: A test collection for clir evaluations in african languages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 293–302, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. . URL https://doi.org/10.1145/3626772.3657884.

[2] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. . URL https://aclanthology.org/2024.findings-acl.137.

[3] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[4] Jinhyuk Lee, Zhuyun Dai, Sai Meher Karthik Duddu, Tao Lei, Iftekhar Naim, Ming-Wei Chang, and Vincent Zhao. Rethinking the role of token retrieval in multi-vector retrieval. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 15384–15405. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/31d997278ee9069d6721bc194174bb4c-Paper-Conference.pdf.

[5] Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*, 2024.

[6] Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA, November 2024. Association for Computational Linguistics. . URL https://aclanthology.org/2024.emnlp-main.1063.

[7] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. . URL https://aclanthology.org/2023.eacl-main.148.

[8] Odunayo Ogundepo, Xinyu Zhang, Shuo Sun, Kevin Duh, and Jimmy Lin. AfriCLIRMatrix: Enabling cross-lingual information retrieval for African languages. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8721–8728, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. . URL https://aclanthology.org/2022.emnlp-main.597.

[9] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669. . URL https://doi.org/10.1561/1500000019.

[10] Nandan Thakur, Luiz Bonifacio, Crystina Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, and Jimmy Lin. "knowing when you don't know": A multilingual relevance assessment dataset for robust retrieval-augmented generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12508–12526, Miami, Florida, USA, November 2024. Association for Computational Linguistics. . URL https://aclanthology.org/2024.findings-emnlp.730.

[11] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023.

[12] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

[13] Xin Zhang, Zehan Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. Language models are universal embedders. *arXiv preprint arXiv:2310.08232*, 2023.

[14] Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*, 2024.

[15] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin, editors, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. . URL https://aclanthology.org/2021.mrl-1.12.

[16] Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. Toward best practices for training multilingual dense retrieval models. *ACM Trans. Inf. Syst.*, (2), sep 2023. ISSN 1046-8188. . URL https://doi.org/10.1145/3613447.

[17] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, pages 1114–1131, 09 2023. ISSN 2307-387X. . URL https://doi.org/10.1162/tacl_a_00595.

[18] Xinyu Zhang, Jing Lu, Vinh Q. Tran, Tal Schuster, Donald Metzler, and Jimmy Lin. Tomato, tomahto, tomate: Measuring the role of shared semantics among subwords in multilingual language models, 2024. URL https://arxiv.org/abs/2411.04530.