The world's rapidly expanding web knowledge is being represented in diverse languages, modalities, and styles. I build AIs that interact with broad swaths of internet users to answer their questions, giving everyone equal access to the valuable information they need. Beneath the impressive progress in natural language processing (NLP), central challenges remain: (1) models are often exclusively built for English speakers, leaving other globally spoken languages behind; (2) models require massive computational resources that few groups can afford; and (3) they lack complex reasoning skills, exploiting spurious annotation artifacts. I tackle these challenges to bridge the gap between NLP research and real-world applications. My research has led to the following publications at premiere NLP and ML conferences.

(1) Multilingual NLP. Prior multilingual NLP research often focused on tasks where all information is known a priori, while real-world systems need to retrieve evidence. I laid the groundwork for developing universal question answering (QA) systems in many languages: XOR-TyDi QA [5] is the first large-scale annotated dataset for this challenge, covering 7 diverse languages. I further proposed the first unified multilingual retriever-generator framework, which effectively locates relevant Wikipedia passages beyond language boundaries and generates final answers conditioned on them with no translations, enabling answering of questions posed in 28 languages [7].

(2) Efficient NLP. Training large pre-trained models for each task achieves best performance at the cost of computational and storage inefficiency, making NLP inaccessible to many groups. I introduced BPR, a binary neural passage retriever [8] that reduces storage requirements 30 times, and ATTEMPT, a new learning paradigm that trains 2,000 times fewer task-specific parameters for each target task [6] relative to the fully trained models that require updating and storing millions of parameters.

(3) Neuro-Symbolic NLP. I led several influential research projects on neuro-symbolic approaches for complex QA. PathRetriever [4] finds reasoning paths on a large-scale document graph to sequentially find evidence, which was adapted to a COVID-19-related publications search [1]. I integrated logic tuples into neural models for consistency [3] and introduced pre-trained entity representations that leverage structured entity relationships [9, 10], which are now used in a range of NLP tasks beyond QA.

In addition to my research outputs, I am an active organizing member of research communities for these areas (e.g., the MIA 2022 Workshop) and advise students at the UW and internationally from underrepresented backgrounds. I am excited to continue pursuing my research interests as a professor, fostering the next generations in AI.

Future Plans. I plan to integrate my current efforts into a *general-purpose, lightweight retriever and neuro-symbolic generator* that can conduct complex reasoning over diverse inputs [2] and flexibly generate accessible outputs and to *improve pre-training* of them using structured data like language links to overcome data scarcity and acquire rich knowledge. My long-term vision includes: (1) advancing NLP in the languages and domains that have scant digital resources (e.g., indigenous languages), and (2) broadening the scope of my work to make an interdisciplinary impact in critical application domains, e.g., medicine and misinformation.

References

[1] Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. Covid-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *NPJ digital medicine*, 2021.

[2] Akari Asai and Eunsol Choi. Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval. In *ACL*, 2021.

[3] Akari Asai and Hannaneh Hajishirzi. Logic-guided data augmentation and regularization for consistent question answering. In *ACL*, 2020.

[4] Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *ICLR*, 2020.

[5] Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. XOR QA: Cross-lingual open-retrieval question answering. In *NAACL-HLT*, 2021.

[6] Akari Asai, Mohammadreza Salehi, Matthew E Peters, and Hannaneh Hajishirzi. Attentional mixtures of soft prompt tuning for parameter-efficient multi-task knowledge sharing. *arXiv preprint arXiv:2205.11961*, 2022.

[7] Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. One question answering model for many languages with cross-lingual dense passage retrieval. In *NeurIPS*, 2021.

[8] Ikuya Yamada, Akari Asai, and Hannaneh Hajishirzi. Efficient passage retrieval with hashing for open-domain question answering. In *ACL*, 2021.

[9] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *EMNLP: System Demonstrations*, 2020.

[10] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *EMNLP*, 2020.