I envision the democratization of natural language processing (NLP) and artificial intelligence (AI) knowledge, where people with diverse skill levels and research backgrounds can: build, use, analyze, and evaluate models; collaborate to solve research problems; and accelerate advances in NLP and AI. AI and NLP have made remarkable progress from recent, large-scale training on massive datasets. These technologies are being developed and used by many cross-disciplinary researchers and practitioners. People with scant computer science training—including physicians, translators, and historians—now rely on AI models for work problems that can be solved by using massive amounts of data.

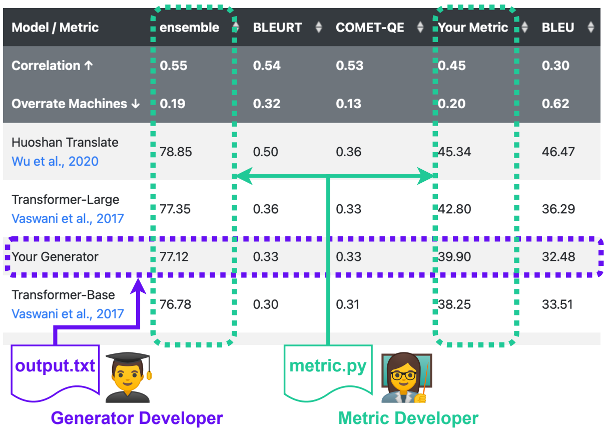
I focus on ways to make AI/NLP more accessible to researchers, practitioners, and users. How can we encourage model builders and practitioners to work as a community to broaden the appeal and utility of NLP and AI models across disciplines? How can we make it easier for them to formulate and answer complex real-world questions using these technologies and ensure these models are robustly evaluated? How can we guarantee that global citizens enjoy the benefits of NLP and AI advances in the way English speakers do today?

During my Ph.D., I made the following contributions towards realizing this long-term vision.

- **Transparent and Reliable Evaluation Methodologies for NLP (§1).** I designed methodologies and interfaces to make model evaluations more transparent, consistent, and reliable [1, 2, 3, 4, 5].
- **Flexible and Customizable Inference Algorithms (§2).** I built algorithms for flexible and customizable language generation in the areas of collaborative inference between diverse models [6]; controlled generation [7]; and fast, accurate inference [8, 9]. All of these methods avoid the computationally (and thus financially and environmentally) expensive training process of large models.
- **Efficient and Accessible NLP (§3).** I introduced and empirically demonstrated efficient architectures and learning paradigms for state-of-the-art NLP models [10, 11, 12, 13, 14]. More efficient methods lower the cost of developing and using these models, making them deployable to less-well-funded fields or institutions.

1 Transparent and Reliable Evaluation Methodologies for NLP

Evaluation is the tool that lets us measure progress, understand model behaviors or failures, and guide future research directions. Evaluating NLP systems presents a serious challenge, particularly for language generation tasks such as machine translation and summarization. Language generation is inherently open-ended, and generation quality cannot be measured using a simple metric like classification accuracy on ImageNet. A large body of NLP work therefore develops automatic evaluation methods that enable fast, inexpensive development cycles [15]. However, this progress in evaluation has been largely overlooked by researchers focused on model advancements [16, 17] mainly because they prioritize consistency of evaluation practices over time. **I believe that this separation between generation and evaluation offers an opportunity for each sub-community to more rapidly benefit from the advances of the other.**



Model / Metric	ensemble	BLEURT	COMET-QE	Your Metric	BLEU
Correlation ↑	0.55	0.54	0.53	0.45	0.30
Overrate Machines ↓	0.19	0.32	0.13	0.20	0.62
Huoshan Translate Wu et al., 2020	78.85	0.50	0.36	45.34	46.47
Transformer-Large Vaswani et al., 2017	77.35	0.36	0.33	42.80	36.29
Your Generator	77.12	0.33	0.33	39.90	32.48
Transformer-Base Vaswani et al., 2017	76.78	0.30	0.31	38.25	33.51

Figure 1: Billboard interface.

Bidimensional leaderboards: an interface that bridges modeling and evaluation research. My NAACL 2022 work [1] introduces an abstraction of leaderboards, called **bidimensional leaderboards** (Billboards), that simultaneously facilitates progress in natural language generation and its evaluation (Fig. 1).[1] A Billboard accepts two types of submissions related to a given task and dataset: **generators** and **metrics**. Unlike conventional leaderboards, Billboards do not tie model ranking to a predetermined set of metrics: generators are ranked based on the most reliable metric currently available. Metric submissions are ranked by their correlations to human judgments, and each is stored as an executable program, which is then used to evaluate future generation submissions. This interface facilitates communication between the modeling and evaluation sub-communities. I currently maintain four Billboards over machine translation, summarization, and image captioning tasks. Each has more than 15 metrics to date, promoting changes in how evaluation is performed and studied in these sub-areas. Billboard's built-in analysis shows that **most automatic evaluations overrate machine over human-written generation, demonstrating the importance of updating metrics as generation models become stronger (and perhaps more similar to humans) in the future.** Though this work was only published in July 2022, it has already impacted evaluation practices of other tasks in NLP and beyond [18, 19], and I anticipate it will continue to do so.

Transparent and scalable human evaluation. As AI technology becomes further enmeshed in our society, I believe it is vital to systematically collect feedback about models from human users. For example, image captioning technology is already used to improve information accessibility for people with visual impairments; my NAACL 2022 work develops a transparent evaluation rubric for state-of-the-art image captioning models [2]. Unlike evaluations conducted by previous work (e.g., evaluations without any interpretable rubric), our human-in-the-loop evaluation demonstrates that machine-generated captions continue to fall short of human-written ones. Our follow-up work [4], Genie, uses crowdsourcing to further scale up human-in-the-loop evaluation to more NLP tasks and models. Genie is the first leaderboard that supports human-in-the-loop evaluation for a broad set of natural language tasks.[2] It has received 63 submissions over six benchmarks, including summarization, machine translation, and commonsense reasoning tasks. We used Genie with WMT [20], an annual and long-standing benchmarking effort for machine translation and NLP, to evaluate system submissions from more than 20 institutions in academia and industry.

Future directions. As my work illustrates, **continual, community-wide efforts are needed to evaluate AI systems**. I am excited to extend my effort beyond NLP to speech and computer vision. Moreover, I would like to extend Billboards to **multidimensional leaderboards**. AI models have many more assessment dimensions than output quality (on a particular test set), such as training and inference efficiency, sample efficiency, environmental impacts, and robustness. These dimensions, often ignored in the current evaluation paradigm, are nonetheless critical to better serving practitioners' needs [21, 22, 23]. I plan to build evaluation platforms where users can explore trade-offs of these different aspects to find a model that best fits their needs and where researchers can contribute AI models that have different strengths. I believe that evaluation should be aligned with the needs of a wide range of stakeholders.

2 Flexible and Customizable Inference Algorithms

Modern AI and NLP models undergo two stages: training, where model parameters are learned based on large datasets, and **inference**, where the model is used to generate desired outputs. Inference is thus a core algorithmic component of NLP systems that has been actively

[1] https://nlp.cs.washington.edu/billboard/. [2] https://genie.apps.allenai.org/.

studied in NLP research. Generation of language handles complex structure and exponentially many possibilities; for a given output length N and vocabulary size V, there are V^N possibilities, making it impossible to enumerate and find the best one. Several inference algorithms (e.g., greedy/beam search) have proven effective on a variety of language generation tasks and continue to be used in recent large-scale models, such as GPT-3 [24] and Google Translate [25]. Overcoming the limitations of commonly used algorithms, I develop inference methods that enable flexible and customizable generation schemes: **collaborative** generation between two diverse generation models [6], **efficient** generation with parallel computation [9], and **controllable** generation [7].

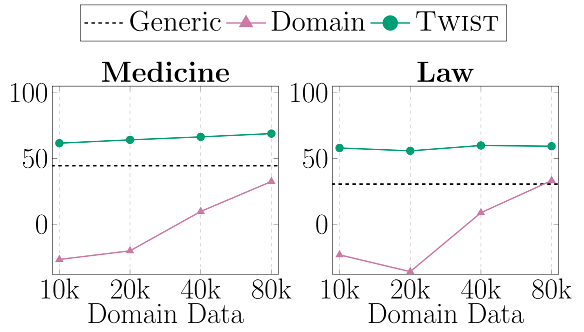


Figure 2: Twist inference between domain and generic models vs. inference in isolation. The y-axis measures the translation quality.

Best of both generators. Combining diverse models (possibly from different institutions) can lead to further progress by **leveraging their complementary strengths**. Conventional ensembling methods make strong assumptions about vocabulary, tokenization, and probability factorization that often do not hold in practice. For example, these assumptions make it challenging to combine models with different specializations (e.g., medical/legal domains). I introduced the Twist algorithm, a simple yet effective method that relaxes these assumptions while avoiding any additional training, which could be expensive. Twist performs standard beam search alternately between two diverse generators with **mutual guidance**. Figure 2 shows results from combining a domain-specific model (medicine/law) with a generic model for German-to-English translation. Quality (y-axis) is measured using a state-of-the-art evaluation method from my Billboard [1]. Here, the number of domain training examples (x-axis) is much smaller than the generic training data (more than 30 million) since the creation of domain data requires bilingual speakers with domain expertise. **Even though the domain model performs poorly by itself, it improves the generic model through use of the Twist algorithm, demonstrating that Twist can make use of complementary strengths.**

Fast, parallel inference algorithm. Inference speed is key to many AI applications, ranging from simultaneous machine interpretation to large image generation. Inference is conventionally done sequentially and word by word (or pixel by pixel), limiting the usage of fast parallel computations from modern hardware (e.g., GPUs and TPUs). I anticipate that a major driving force of computing improvement will be increased parallelism, given physical constraints on the clock speed (i.e., the *heat wall*, [26]). For this reason, I believe that developing a parallelizable inference algorithm is one promising approach to fast generation. In my ICML 2019 work [9], I introduced a parallel algorithm that achieves similar generation quality with a speedup of more than 3x on a modern GPU compared to standard sequential inference. **This work has received continuing interest from the NLP, machine learning, and even speech communities** (e.g., [27, 28, 29, 30]).

Future directions. AI model usability depends heavily on inference algorithms. Developing flexible, customizable, and efficient inference algorithms is therefore essential to my long-term goal of building accessible NLP. Looking forward, I believe that there is a critical limitation of current algorithms that should be addressed to this end: current inference methods lack control over generation outputs, occasionally resulting in hallucinations and non-inclusive or toxic language. These problems can have serious implications for downstream applications, as observed in my evaluation work on image captioning [2]. My collaborators and I developed an A* search-based algorithm that enables lexical constraints on language generation [7], which won the best paper award at NAACL 2022. Building upon this work, I aim for the even more challenging controllable generation that, for instance, avoids toxic language (and images). Specifically, I plan to integrate my work on evaluation methodologies with an inference algorithm so that, for example, an evaluation metric from BILLBOARDS can capture the inclusiveness of generated language, which sends a useful signal to inference.

3 Efficient and Accessible NLP

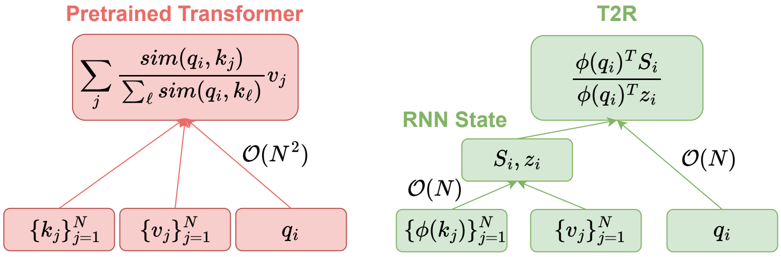


Figure 3: T2R converts pretrained transformers to efficient RNNs.

The growing computational (and thus financial and environmental [21]) cost of large-scale AI/NLP models makes it difficult for many researchers to develop or even just use them. Only researchers working at a handful of extremely well-funded industry labs are able to perform necessary operations on the current state-of-the-art models to make progress. Efficient and accessible methods are needed to promote the accessibility of our technology and to ensure that people from diverse backgrounds can contribute to its progress from their various perspectives. **I strongly believe that the inclusiveness of the AI community will be key to its success.**

Finetuning off-the-self transformers into efficient RNNs. The transformer architecture [31] is a backbone of recent advances in NLP, computer vision, speech, computational biology, and beyond (e.g., [24, 32, 33]). Transformers outperform recurrent neural networks (RNNs) at the expense of their increased computational cost: their time and memory complexity scales quadratically with sequence length, in contrast to the linear complexity of RNNs. This computational requirement limits the usability of many strong transformers in the AI community. My EMNLP 2021 work [10] introduced transformer-to-RNN (T2R), a method that converts any off-the-shelf transformer into an efficient RNN counterpart by performing a small amount of finetuning (Figure 3). Drawing inspiration from the classical kernel methods in machine learning, T2R learns to approximate transformers' quadratic computation during lightweight finetuning, resulting in a recurrent model with linear complexity. I empirically demonstrated that T2R generates long text with quality similar to the original transformer while achieving substantial speedup and memory savings (**e.g., 10x speedup when producing 2048 consecutive words**; Figure 4).

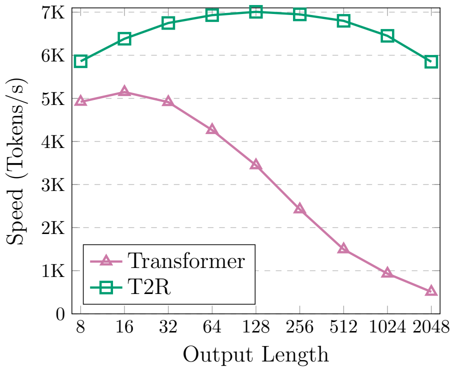


Figure 4: Text generation speed with varying lengths.

Future directions. Efficiency research is a subject of increasing interest to AI and NLP researchers. Developing efficient methods often requires a deep understanding of tasks and applications of interest. For example, I showed that a simple strategy of **deep encoder, shal-**

low decoder improves efficiency while retaining strong performance in sequence-to-sequence applications, such as machine translation [11]. I am passionate about developing efficient machine learning methods that are motivated by NLP applications. Conversely, I also find it exciting to draw mathematical connections among various approaches; such unified abstractions can help us better understand seemingly disparate research and inspire new advances, as illustrated by our ACL 2022 work [12]. Finally, efficiency research should not be limited to computational aspects. In particular, sample and annotation efficiency for data creation are also crucial to empower diverse AI applications. I had initial success in this attempt [34, 14], and I hope to expand such efforts.

4 Future Research Agenda

Dynamic and real-world evaluations. The AI train has already left the station: GPT-3 is used in more than 300 apps, and Google Translate translates more than 100 billion words per day. **At present, most AI and NLP models are evaluated statically and only once, when papers are written.** Since AI models have proven useful in many tasks, there has never been a more opportune time for the community to explore dynamic and real-world evaluations beyond fixed test data. My collaborators and I recently started **RealTime QA, a dynamic benchmarking effort that evaluates question answering systems in real time** [3].[3] **Such dynamic evaluations benefit from interdisciplinary collaborations**; real-time question answering systems can, for instance, facilitate emergency management of natural disasters and pandemics. I am excited to contribute to this effort from NLP and machine learning perspectives. I believe that real-world evaluations will reveal crucial challenges that guide our future research in many aspects. For example: How should a real-time system combat fake news or toxic content? How can we update an NLP system **quickly and efficiently** to fulfill real-time information needs? I am confident that my expertise in evaluation methodologies, efficient NLP, and language generation will add valuable insights to this research area.

Massively multilingual NLP. Current NLP data creation and model development focus heavily on the English language, leading to over-representation of English-centric problems (e.g., information needs about American politics). **This lack of multilingual NLP research and resources limits the diversity and accessibility of AI technology and heightens the barriers to the use of many AI applications in the world.** Indeed, English covers only one quarter of global web users.[4] I will pursue massively multilingual processing that benefits people around the world; I will contribute to creating linguistically diverse resources and advancing models that can be used for applications in many languages. During the early years of my Ph.D., my collaborators and I showed the possibility of effectively expanding AI models to diverse languages with smaller or even no labeled training data by using our multilingual vector representation [35, 36]. Multilingual vector representations continue to be studied in more recent work (e.g., [37]). I am particularly excited about exploring multilingual processing from the perspectives of **inference algorithms and efficiency**. Many global languages lack large knowledge sources, such as Wikipedia, but much current work assumes that user questions can be answered based solely on a knowledge source in the users' own language [38, 39, 40]. Can we find an inference algorithm that systematically combines knowledge sources from various languages and provides answers to users regardless of their language? Can we develop models that efficiently process many languages? With the rich body of tools and methods currently available, I believe that the time is ripe to tackle these challenging research problems and improve the inclusiveness of language technologies for the billions of speakers of the 7,000+ languages other than English.

[3] https://realtimeqa.github.io/. [4] https://www.internetworldstats.com/stats7.htm.

References

[1] **Jungo Kasai**, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander R. Fabbri, Yejin Choi, and Noah A. Smith. Bidimensional leaderboards: Generate and evaluate language hand in hand. In *Proc. of NAACL*, 2022.

[2] **Jungo Kasai**, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. Transparent human evaluation for image captioning. In *Proc. of NAACL*, 2022.

[3] **Jungo Kasai**, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. RealTime QA: What's the answer right now?, 2022. Under review.

[4] Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, **Jungo Kasai**, Yejin Choi, Noah A. Smith, and Daniel S. Weld. GENIE: Toward reproducible and standardized human evaluation for text generation. In *Proc. of EMNLP*, 2022.

[5] Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, **Jungo Kasai**, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic, 2022. Under review.

[6] **Jungo Kasai**, Keisuke Sakaguchi, Ronan Le Bras, Hao Peng, Ximing Lu, Dragomir Radev, Yejin Choi, and Noah A. Smith. Twist decoding: Diverse generators guide each other. In *Proc. of EMNLP*, 2022.

[7] Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, **Jungo Kasai**, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. Neurologic A*esque decoding: Constrained text generation with lookahead heuristics. In *Proc. of NAACL*, 2022. **Best Paper Award**.

[8] **Jungo Kasai**, Keisuke Sakaguchi, Ronan Le Bras, Dragomir Radev, Yejin Choi, and Noah A. Smith. Beam decoding with controlled patience, 2022. Under review.

[9] **Jungo Kasai**, James Cross, Marjan Ghazvininejad, and Jiatao Gu. Non-autoregressive machine translation with disentangled context transformer. In *Proc. of ICML*, 2020.

[10] **Jungo Kasai**, Hao Peng, Yizhe Zhang, Dani Yogatama, Gabriel Ilharco, Nikolaos Pappas, Yi Mao, Weizhu Chen, and Noah A. Smith. Finetuning pretrained transformers into RNNs. In *Proc. of EMNLP*, 2021.

[11] **Jungo Kasai**, Nikolaos Pappas, Hao Peng, James Cross, and Noah A. Smith. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *Proc. of ICLR*, 2021.

[12] Hao Peng, **Jungo Kasai**, Nikolaos Pappas, Dani Yogatama, Zhaofeng Wu, Lingpeng Kong, Roy Schwartz, and Noah A. Smith. ABC: Attention with bounded-memory control. In *Proc. of ACL*, 2022.

[13] Haoxin Li, Phillip Keung, Daniel Cheng, **Jungo Kasai**, and Noah A. Smith. NarrowBERT: Accelerating masked language model pretraining and inference. Under review.

[14] Hongjin Su, **Jungo Kasai**, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners. Under review.

[15] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey, 2020.

[16] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proc. of ACL*, 2020.

[17] Benjamin Marie, Atsushi Fujita, and Raphael Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proc. of ACL*, 2021.

[18] Afra Feyza Akyürek, Muhammed Yusuf Kocyigit, Sejin Paik, and Derry Tanti Wijaya. Challenges in measuring bias via open-ended language generation. In *Proc. of GeBNLP*, 2022.

[19] David Harbecke, Yuxuan Chen, Leonhard Hennig, and Christoph Alt. Why only micro-F1? class weighting of measures for relation classification. In *Proc. of the Workshop on Efficient Benchmarking in NLP*, 2022.

[20] Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, **Jungo Kasai**, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. Findings of the 2021 conference on machine translation (WMT21). In *Proc. of WMT*, 2021.

[21] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *CACM*, 2019.

[22] Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proc. of EMNLP*, November 2020.

[23] Swaroop Mishra and Anjana Arunkumar. How robust are model rankings : A leaderboard customization approach for equitable evaluation. In *Proc. of AAAI*, 2021.

[24] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proc. of NeurIPS*, 2020.

[25] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 2017.

[26] Daniel Etiemble. 45-year CPU evolution: one law and two equations. In *Proc. of* WP^3, 2018.

[27] Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. Non-autoregressive machine translation with latent alignments. In *Proc. of EMNLP*, 2020.

[28] Xiao Shi Huang, Felipe Perez, and Maksims Volkovs. Improving non-autoregressive translation models without distillation. In *Proc. of ICLR*, 2022.

[29] Fei Huang, Tianhua Tao, Hao Zhou, Lei Li, and Minlie Huang. On the learning of non-autoregressive transformers. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proc. of ICML*, 2022.

[30] Chuan-Fei Zhang, Yan Liu, Tian-Hao Zhang, Song-Lu Chen, Feng Chen, and Xu-Cheng Yin. Non-autoregressive transformer with unified bidirectional decoder for automatic speech recognition. In *Proc. of ICASSP*, 2022.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. of NeurIPS*, 2017.

[32] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *Proc. of ICML*, 2018.

[33] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021.

[34] **Jungo Kasai**, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. Low-resource deep entity resolution with transfer and active learning. In *Proc. of ACL*, 2019.

[35] Phoebe Mulcaire, **Jungo Kasai**, and Noah A. Smith. Polyglot contextual representations improve crosslingual transfer. In *Proc. of NAACL*, 2019.

[36] Phoebe Mulcaire*, **Jungo Kasai***, and Noah A. Smith. Low-resource parsing with crosslingual contextualized representations. In *Proc. of CoNLL*, 2019. * equal contribution.

[37] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. of NAACL*, 2021.

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

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

[40] Akari Asai, Shayne Longpre, **Jungo Kasai**, Chia-Hsuan Lee, Rui Zhang, Junjie Hu, Ikuya Yamada, Jonathan H. Clark, and Eunsol Choi. MIA 2022 shared task: Evaluating cross-lingual open-retrieval question answering for 16 diverse languages. In *Proc. of MIA*, 2022.