Ubiquitous Long-Form Text Generation

Automating the production of well-formed documents will have a profound impact on the way we interact with information. In the future, I envision a world where long-form text generation is ubiquitous, and anyone can produce a document on any topic with just a few clicks. For example, a physician can ask a summarization system to generate an up-to-date case review of a disease, so that they are more likely to come up with optimal treatment plans for their patients. A student can use a book generation system to write a textbook customized to their interest, knowledge background, and preferences, so that they can learn most efficiently at their own pace. A grandmother can use a story generation system to assist her in writing a memoir of her life that can be passed down to future generations.

To achieve this vision, my research focuses on the following three directions, each utilizing recent advances in deep learning combined with probabilistic modeling:
(1) **Transparent Generation** to enable users to understand, debug, and control document generation. When text generation systems are applied in high-stakes domains, such as healthcare, finance, and law, it is essential to interpret and control how these models come to their decisions.
(2) **Long-Form Coherence** to generate coherent documents with self-consistency and clear transitions. When systems are used for narrative generation, it is essential to generate texts that flow smoothly, preserving character motivation and plot development over a long span of text.
(3) **Efficient Systems** to handle the scale, complexity, and real-time requirements of long-form text generation. When systems are used in resource-limited settings such as mobile devices by billions of users, it is essential to generate text quickly and with low resource-usage.

Transparent Generation

For users to understand and control document generation, it is critical that models expose their internal decisions. For example, in a medical summarization system, a doctor needs to know why a machine-generated case review recommends a certain treatment plan.

I work on providing model transparency to support user intervention during the generation process. In *Latent Alignment and Variational Attention* [1], I explicitly apply probabilistic modeling to "attention", a mechanism aligning source elements with target elements which is traditionally treated as a deterministic black box. These *latent* attention variables can serve as control knobs by allowing the user to choose which parts in the source to focus on during generation. In a follow-up work [2], I use a source content selector to control the latent attention, which allows easy domain transfer by changing which source words are selected and established a new state-of-the-art on a document summarization benchmark. These papers have been widely cited (over **700 citations** in total) and used in various applications, such as controlling the underlying knowledge during dialogue generation [3] and the underlying language in synthesizing multilingual speech (code-mixing) [4].

Transparency is critical for debugging text generation, as it supports users in investigating why a system makes certain decisions and whether those decisions are aligned with their goals [5]. In *Rationales for Sequential Predictions* [6], I proposed a method for exposing the internal decisions of a sequence generation model by finding rationales, i.e., attributing the prediction of output words to subsets of input elements. The method can be applied to any model class. During training, the system randomly masks out a subset of the inputs such that the model learns to behave properly given

incomplete subsets of inputs. During inference, words are iteratively added back in to maximally increase the likelihood of the given output word, until that word is predicted by the model. This work was the first to find rationales for text generation models, and the rationales found by the model are similar to those by human experts. This work has also been extended to other domains such as debugging career trajectory prediction models [7].

Long-Form Coherence

Document generation requires coherence, e.g., self-consistency and clear section transitions. For example, a generated textbook should ensure that prerequisite concepts are introduced before advanced concepts and that there is a logical flow between sections. While there have been many studies on local fluency, less work has been done on evaluating and improving global, high-level coherence.

In *Model Criticism for Long-Form Text Generation* [8], I propose a statistical framework for evaluating the "coherence" of machine-generated texts. This approach first projects data to a latent space based on an assumptive critic generative process, and then compares the implied latent distributions between real data and generated texts. A mismatch of these latent distributions exposes high-level modeling errors. Using this tool, I show that it is possible for a text generation model to exhibit very strong local fluency, but fail to capture high-level dynamics in a controlled environment. By plugging in different critic generative processes, this method can critique specific failure modes of text generation. In particular, I find that while modern large language models can capture topical structures well, they are not currently strong at modeling discourse coherence or coreference, critical properties for document structure. This work is one of the first papers to quantitatively study these problems, and our findings suggest that there is much room for improvement in current language models, motivating future work.

A critic can also be used to improve model coherence [9]. In *Residual Energy-Based Models for Text Generation* [10], I propose to improve long-term modeling by incorporating a critic that identifies global errors in text. This model adds a penalty to the likelihood of text that is proportional to the critic's energy. The global critic is trained to distinguish real text from machine-generated text, such that it learns to penalize incoherent generations such as repeated sentences, as they appear more frequently under the model than in real data. Experiments on large-scale datasets show that generations using this approach are considered more coherent by humans, have better statistical matches to real data, and are harder to distinguish from real data under learned critics. Recently, this approach has been scaled by other researchers to even larger settings of tens of billions of parameters and is still shown to improve coherence compared to baselines [11].

Efficient Systems

Efficient document generation algorithms are necessary to make this technology accessible to those who it would benefit the most, such as a grandmother who wants to write a memoir using a life story generation system on her mobile phone. Current language models require vast computational resources to deploy. For example, the hardware cost of running GPT-3 is estimated to be between $100k and $150k a year without factoring in other costs (electricity, cooling, backup, etc.) [12].

To improve algorithmic efficiency, in *Cascaded Text Generation with Markov Transformers* [13], I

developed a text generation algorithm codesigned to run fast on parallel hardware. In this work, I modified a neural model to parameterize a cascade of conditional random fields (CRFs) [14]. This formulation enables using a coarse-to-fine decoding algorithm [15], which narrows down the search space of generation candidates using gradually more powerful models. Since running probabilistic inference on CRFs can be parallelized, this approach achieved the lowest inference time on a machine translation benchmark while maintaining a competitive generation quality [16].

To enable efficiently running text generation models, I also collaborated with hardware researchers aiming to deploy text generation models. We developed an efficient number encoding algorithm AdaptivFloat [17]. Based on the insight that different layers in models with layer normalization (such as Transformers) exhibit very different dynamic ranges, AdaptivFloat maximizes its available dynamic range by customizing its exponent range per layer before applying quantization. This number encoding algorithm is easy to implement using basic hardware logic, yet it enables high model compression ratios (up to 5X) without losing accuracy on a machine translation task and a speech recognition task. This work won a **DAC 2020 Best Paper Award**.

Open-Source Projects

To make text generation technologies more accessible, I also work on open-source tools that make it easy for anyone to use text generation systems. I was a major contributor to OpenNMT [18], an open-source neural machine translation framework available in multiple programming languages and deployable to various hardware platforms. This library has been widely used in academia and industry and has won an **ACL 2017 Best Demo Paper Award Runner-Up**.

In *Image-to-Markup Generation with Coarse-to-Fine Attention* [19], I proposed to use a cascade of attention mechanisms to gradually zoom into the input image. On a math image recognition task, this work only uses a tenth of the original input image per output word after amortization without sacrificing accuracy. The open-source implementation Im2LaTeX [19] beat the best commercial system for math image recognition by 20% accuracy, and has been used by companies such as MathPix for real-world PDF-to-LaTeX conversion.

Future Directions

Multimodal Long-Form Summarization In domains such as education, healthcare, news, and research, documents are highly multimodal, consisting of text, images, tables, and databases. I plan to develop models that can generate multimodal long-form summaries of documents that take into account the different modalities and their relationships. By attacking the challenges of coherent modeling—how to generate a summary that is coherent across multiple modalities, and efficient modeling—how to make training and inference on long inputs and long outputs possible [20], this technology would bring my vision of ubiquitous long-form text generation one step closer to reality, and would ultimately enable accessible information production such as personalized education, quality healthcare, cutting-edge news, and affordable legal services. My past experience in image-to-text generation [19] and text-to-image generation [21] will be helpful for this direction.

Creative Text Generation In many applications such as assistive content creation, we want the text generation system to be "creative", i.e., to generate *new* text that is not just a regurgitation

of existing information. Existing language models fail to generate creative text because they cannot model high-level dynamics such as the story progression in a fiction book [22]. Besides, it is hard to control these models to generate text with desired properties such as a "happy ending". To address these challenges, I plan to pursue three directions. First, I will investigate why language models have a hard time capturing high-level dynamics, utilizing the findings of my work on model criticism [8]. Second, I will develop long-form coherent models by guiding the generation process with a high-level content plan, and/or by more efficiently using training data such as by focusing on learning high-level dynamics rather than low-level details late during training. Third, I will work on exposing interpretable control knobs of the model to enable human control. These advancements could eventually enable many exciting applications such as creating new fiction books according to user preferences, writing a sequel or prequel of a book, and even generating a new research proposal based on existing publications.

Non-Language Sequences Advances in general-purpose long-form text generation can be applied to other domains where the outputs take the form of long discrete sequences, such as genome sequences. In an ongoing project, I collaborate with Nvidia and Argonne National Lab on applying long-form text generation technologies to genome modeling. We have trained the largest foundation model on genome sequences (25 billion parameters) [23], and our work has won an **ACM Gordon Bell Covid Prize**. Since genome sequences are extremely long (Covid sequences have 10k codons on average), my research on efficient and coherent modeling can help train models that are computationally tractable yet consider long-form context. This research direction presents huge potential impacts on life science and healthcare. For example, a prediction of how the genome sequence of a certain virus evolves will assist in vaccine development.

References

[1] **Yuntian Deng**, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. 2018. Latent alignment and variational attention. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[2] Sebastian Gehrmann, **Yuntian Deng**, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

[3] Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.

[4] SaiKrishna Rallabandi and Alan W. Black. 2019. Variational attention using articulatory priors for generating code mixed speech using monolingual corpora. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, pages 3735–3739, Graz, Austria. International Speech Communication Association.

[5] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

[6] Keyon Vafa, **Yuntian Deng**, David Blei, and Alexander Rush. 2021. Rationales for sequential predictions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pages 10314–10332, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[7] Keyon Vafa, Emil Palikot, Tianyu Du, Ayush Kanodia, Susan Athey, and David Blei. 2022. CAREER: Economic prediction of labor sequence data under distribution shift. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*.

[8] **Yuntian Deng**, Volodymyr Kuleshov, and Alexander Rush. 2022. Model criticism for long-form text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, UAE. Association for Computational Linguistics.

[9] Anton Bakhtin*, **Yuntian Deng***, Sam Gross, Myle Ott, Marc'Aurelio Ranzato, and Arthur Szlam. 2021. Residual energy-based models for text. *Journal of Machine Learning Research*, 22(40):1–41. *Equal Contribution.

[10] **Yuntian Deng**, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. 2020. Residual energy-based models for text generation. In *International Conference on Learning Representations*.

[11] Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. Rankgen: Improving text generation with large ranking models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, UAE. Association for Computational Linguistics.

[12] Ben Dickson. 2020. The GPT-3 economy. `https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/`. [Online; accessed 30-Oct-2022].

[13] **Yuntian Deng** and Alexander Rush. 2020. Cascaded text generation with markov transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 170–181. Curran Associates, Inc.

[14] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, San Francisco, USA. Morgan Kaufmann Publishers Inc.

[15] David Weiss and Benjamin Taskar. 2010. Structured prediction cascades. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 916–923, Chia Laguna Resort, Sardinia, Italy. PMLR.

[16] **Yuntian Deng** and Alexander Rush. 2021. Sequence-to-lattice models for fast translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3765–3772, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[17] Thierry Tambe, En-Yu Yang, Zishen Wan, **Yuntian Deng**, Vijay Janapa Reddi, Alexander Rush, David Brooks, and Gu-Yeon Wei. 2020. Algorithm-hardware co-design of adaptive floating-point encodings for resilient deep learning inference. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6.

[18] Guillaume Klein, Yoon Kim, **Yuntian Deng**, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics. **Best Demo Paper Award Runner-Up**.

[19] **Yuntian Deng**, Anssi Kanervisto, Jeffrey Ling, and Alexander M. Rush. 2017. Image-to-markup generation with coarse-to-fine attention. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 980–989. PMLR.

[20] Justin Chiu*, **Yuntian Deng***, and Alexander Rush. 2021. Low-rank constraints for fast inference in structured models. In *Advances in Neural Information Processing Systems*, volume 34, pages 2887–2898. Curran Associates, Inc. *Equal Contribution.

[21] **Yuntian Deng**, Noriyuki Kojima, and Alexander M. Rush. 2022. Markup-to-image diffusion models with scheduled sampling. *arXiv preprint*, arXiv:1710.03740.

[22] Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision.

[23] Maxim Zvyagin*, Alexander Brace*, Kyle Hippe*, **Yuntian Deng***, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, Carla M. Mann, Michael Irvin, J. Gregory Pauloski, Logan Ward, Valerie Hayot, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, Chaowei Xiao, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, and Arvind Ramanathan. 2022. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *bioRxiv*. *Equal Contribution. **ACM Gordon Bell Covid Prize**.