The Amazing World of Neural Language Generation

Yangfeng Ji [1] **Antoine Bosselut** [2,3] **Thomas Wolf** [4] **Asli Celikyilmaz** [5]
[1] University of Virginia
[2] Allen Institute for Artificial Intelligence, Seattle, WA, USA
[3] Paul G. Allen School of Computer Science & Engineering, Seattle, WA, USA
[4] Huggingface Inc.
[5] Microsoft Research, Redmond, WA, USA
yangfeng@virginia.edu, antoineb@cs.washington.edu
thomwolf@gmail.com, aslicel@microsoft.com

Abstract

Neural Language Generation (NLG) – using neural network models to generate coherent text – is among the most promising methods for automated text creation. Recent years have seen a paradigm shift in neural text generation, caused by the advances in deep contextual language modeling (e.g., LSTMs, GPT, GPT2) and transfer learning (e.g., ELMo, BERT). While these tools have dramatically improved the state of NLG, particularly for low resources tasks, state-of-the-art NLG models still face many challenges: a lack of diversity in generated text, commonsense violations in depicted situations, difficulties in making use of factual information, and difficulties in designing reliable evaluation metrics. In this tutorial, we will present an overview of the current state-of-the-art in neural network architectures, and how they shaped recent research directions in text generation. We will discuss how and why these models succeed/fail at generating coherent text, and provide insights on several applications.

Type. Cutting-edge.

1 Introduction

Natural Language Generation (NLG) forms the basis of many Natural Language Processing (NLP) tasks such as document summarization, machine translation, image captioning, conversational dialogue, and creative writing, making it an essential component in human-machine communication tasks. With recent progress in training deep neural networks, there has been a paradigm shift from template based approaches to neural methods as the predominant building blocks for text generation systems. Specifically, the rich representation learning capabilities of neural networks have allowed NLG models to be trained directly from large amounts of training data, significantly reducing the need for manual feature engineering.

Many benefits have emerged from this new research direction. First, the prototypical framework for training neural networks in an end-to-end fashion has allowed for a diverse array of contextual information to be incorporable into text generation systems (Vaswani et al., 2017; Radford et al., 2019; Ziegler et al., 2019; Keskar et al., 2019), allowing for a richer range of stylistic variability in generated text. Simultaneously, the combination of deep neural networks, large-scale text data and cheap computational power has accelerated new developments in neural network language models.

However, NLG models still raise many challenges which are the focus of a growing body of work. Examples of such limitations are the lack of diversity in generated texts, difficulty in controlling the discourse coherence of the generated text, the lack of commonsense in generated outputs, an uncertain reliance on provided factual information, and more general open questions on architecture design and optimization settings.

In this tutorial, we will start with an introduction to neural language generation, presenting neural language models and encoder-decoder models. We will then discuss the capabilities and limitations of recent text generation models, the suitable architectures for text generation in various specific applications, and then provide insights into why and how these generation models can be adapted for a particular task (Wiseman et al., 2017; Li et al., 2017; See et al., 2017; Xie, 2017). The discussion on evaluation metrics will start from *n*-gram matching up to the recent progress on text generation evaluation metrics. In the end, this tutorial will be concluded by presenting and discussing major current research directions in the field of neural language generation. All materials (including slides, code, and demos) will be publicly available online on the day of the tutorial. We do not assume any particular prior knowl-

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 37–42
Online, November 19 - 20, 2020.
https://doi.org/10.18653/v1/P17

edge in text generation or language modeling. Familiarity with standard neural network modules (LSTM/CNN/Transformer) is a plus but not required. The intended length of the tutorial is 3 hours, including a coffee break.

2 Tutorial Goal and Description

2.1 Overview

This tutorial will mainly focus on the recent advances in neural networks for language generation and will have minimal coverage on traditional methods. We will provide an overview on the recent progress of neural language generation for those working in this research area, and will also introduce this exciting research area to the NLP researchers who are not familiar with newest advancements in neural text generation. This tutorial is designed for anyone who has basic knowledge background of NLP or deep learning, which makes it accessible to any attendee of an NLP conference.

2.2 Tutorial Organization

Fundamentals and Progression of Neural Text Generation. Interest in neural text generation was recently catalyzed by the renaissance of neural network research in natural language processing, particularly with the development of neural language models and encoder-decoder models. Requiring minimal templates and hand-designed rules, unlike classical language generation methods, neural language generation models massively reduce the time needed to design and build new text generation system.

In particular, language models and encoder-decoder models conveniently allows to incorporate contexts such as previous or parallel sentences, as exemplified in machine translation models. However the spectrum of applications of NLG systems extends far beyond machine translation and can involve: (1) complex reasoning processes that go behind semantically preserving mapping from one language to another, for instance to model discourse, dialog flows or multi-hop reasoning; (2) a wide range of context information, from memory to multi-modalities like images or speech; and (3) challenging evaluation, as multiple generated outputs can be simultaneously valid for a given context (so called high-entropy tasks). The tutorial will highlight some these topics and provide a comprehensive overview of the advances of neural language generation.

Technical Details for Training and Optimization Neural Text Generation. Many of the recent progresses in neural language generation can be characterized as approaches to address some of the above mentioned issues. By investigating the difference between language generation and other sequential modeling problems, novel training methods (e.g., reinforcement learning or imitation learning) can be designed to capture long-term dependencies in generation. New decoding methods like top-k (Fan et al., 2018), nucleus sampling (Holtzman et al., 2019) or penalized sampling (Keskar et al., 2019) are invented to resolve the diversity issues.

Eventually, smarter ways to incorporate various contextual information in neural network models (Golovanov et al., 2019; Ziegler et al., 2019; Radford et al., 2019; Keskar et al., 2019) provide more flexibility as well as a better reliance of the model on the conditioning inputs.

Evaluation of Text Generation. Finally, there is a formidable challenge in getting better metrics to evaluate the quality of generated texts that stems from open-ended nature of these models output. Leveraging recent advances in representation learning, the field of neural language generation has been able to move beyond evaluation methods based on n-gram matching and incorporate promising approaches to design more reliable evaluation metrics. This tutorial will cover recent progress in this field as well as highlighting pressing issues with the current state of experimental reporting in NLG. Together with evaluation, we will overview several text generation benchmarks commonly used in the field.

Lessons Learned, Future Directions and Practical Advances of Neural Text Geneation. The last part of this tutorial will discuss practical issues when using cutting-edge language generation techniques. Most of the content covered in this part will have corresponding code or demo implemented in a standard deep learning framework like PyTorch or TensorFlow. The concluding part of the tutorial, we will provide a summary of current and future research direction as well as of some open questions to open the discussion.

3 Diversity and Inclusion

Diversity. The background of the instructors of this tutorial is evenly distributed among academia and industry. The instructors consist of a group of researchers ranging from an assistant professor at University of Virginia (Yangfeng Ji), a senior Ph.D. student at University of Washington with years of industry research experience (Antoine Bosselut) and two senior research scientists in industry (Thomas Wolf and Asli Celikyilmaz), who both have years of industry research experience. The tutorial instructors are also from different countries and continents (the Netherlands and USA).

4 Outline

4.1 Schedule

The tutorial will be 3 hours long.

1. **Introduction of Natural Language Generation** (15 minutes long): This section will introduce the tutorial by presenting the recent impact of neural network modeling approaches on the field. We will briefly overview the classical text generation pipeline, and introduce basic building blocks of neural text generation: language modeling and the encoder-decoder frameworks. We will also discuss the limitations of the simple encoder-decoder frameworks and motivate the rest of the tutorial.

2. **Building blocks of Neural Network Models for Language Generation** (60 minutes long): This section will comprise three closely related topics corresponding to three fundamental aspects of building a neural language generation system: (1) selecting the architecture of the model among a variety of choices such as pre-trained language models (Devlin et al., 2018; Radford et al., 2019), variational autoencoders (Bowman et al., 2016; Hu et al., 2017), generative adversarial networks (Fedus et al., 2018; Subramanian et al., 2018), or neural template based methods (Wiseman et al., 2018; Xu et al., 2018); (2) training the model using techniques which can range from simple maximum likelihood estimate up to more advanced training techniques like scheduled sampling (Bengio et al., 2015), unlikelihood training (Welleck et al., 2019) or reinforcement/imitation learning (Kreutzer et al., 2018; Tan et al., 2018; Huang et al., 2019; Du and Ji, 2019) which can help alleviate exposure bias (He et al., 2019) and repetition issues, and improve handling long-term rewards; (3) selecting a decoding strategy, from classical methods like greedy decoding, beam search and random sampling up to more recent techniques like top-k (Fan et al., 2018), nucleus sampling (Holtzman et al., 2019) or penalized sampling (Keskar et al., 2019). This section will cover the material on classical techniques (30% of time) and mainly focus the recent progress on the related topics (70% of time)

3. **Break** (20 minutes)

4. **Generation with Rich Context** (25 minutes long): This section will discuss recent works on incorporating various types of context information in neural language generation. Going beyond simple context information provided by single sentence contexts, we will overview the growing body of work exploring various strategies to incorporate different types of context information either textual, e.g., syntactic, topic, and discourse information (Wang et al., 2019; Shen et al., 2019; Clark et al., 2018; Bosselut et al., 2018), or beyond text, including knowledge graph, database and images (Parthasarathi and Pineau, 2018; Dinan et al., 2018).

5. **Benchmarks and Evaluation** (30 minutes long): Given the diversity of text generation tasks and domains, it can be challenging to design reliable benchmarks and evaluation metrics (Lowe et al., 2017; Reiter, 2018; Clark et al., 2019; See et al., 2019). In this section, we will summarize the current status on these topics.

6. **Building Neural Models for Generation** (20 minutes long): This section will provide hand-on exercise, using existing deep learning packages, to build a neural language generation model. This section will also demonstrates how different learning/decoding strategies can have a strong impact on the quality of generated texts.

7. **Open problems and directions** (10 minutes long): In this final section, we will summarize the topics covered in the tutorial and point to a selection of open problems and future research directions.

4.2 Breadth

We estimate that the 30% of the tutorial will cover the recent work by the tutorial presenters, and the rest will be on cutting-research work by other researchers.

5 Information about the Presenters

Yangfeng Ji is the William Wulf Assistant Professor in the Department of Computer Science at the University of Virginia, where he leads the Natural Language Processing group. His research interests include building machine learning models for text understanding and generation. His work on entity-driven story generation won an Outstanding Paper Award at NAACL 2018.website

Antoine Bosselut is a PhD student in the Paul G. Allen School of Computer Science at the University of Washington and a Student Researcher at the Allen Institute for Artificial Intelligence (AI2). His research interests are in integrating commonsense knowledge and reasoning into downstream applications for text generation, summarization, and conversational dialogue. He regularly publishes papers at ACL, NAACL, EMNLP, and ICLR. He organized the NeuralGen workshop at NAACL 2019, and West Coast NLP (WeCNLP) in 2018 and 2019. website

Thomas Wolf leads the Science Team at Huggingface Inc., a Brooklyn-based startup working on Natural Language Generation and Understanding. He previously co-organized the NeuralGen 2019 workshop and the tutorial on Transfer Learning in NLP at NAACL 2019. His team has open-sourced several widely used libraries for coreference resolution and transfer learning in NLP and regularly publish research papers in ML and CL conferences (ICLR, ACL, AAAI...). His primary research interest is Natural Language Generation and Transfer Learning. website

Asli Celikyilmaz is Principal Researcher at Microsoft Research in Redmond, Washington. She is also an Affiliate Professor at the University of Washington. Her research interests are mainly in deep learning and natural language, specifically on long text generation, multi-document summarization, conversational modeling, human-computer interaction, and knowledge representation. She has presented several tutorials at venues including CoLing'18, ACL'17, ICASSP'17, Interspeech'17 and organized workshops at ACL, NAACL, Neurips. She has published several papers in ACL, EMNLP, NAACL, CVPR, NeurIPS, ICLR, ICASSP, IEEE TASLP, among other venues. She received several 'best of' awards including best paper award at Semantic Computing 2009, CVPR 2019. She received her Ph.D. degree from University of Toronto, Canada. website

6 Additional details

Audience Size. Based on the increasing interest in natural language generation (larger growth rate in submissions compared to other areas of NLP[1]), we anticipate that between 150 and 200 attendees will be interested in this tutorial.

Special Requirements. The tutorial will require internet access for participants to be able to access the slides and, optionally, to access hands-on coding notebooks.

Preferred Venues. Our preferred venues are EMNLP 2020, ACL 2020, and CoLing 2020.

Open Access. We agree to allow the publication of our slides and a video recording of our tutorial in the ACL Anthology. All our materials will additionally be posted on our tutorial website.

Small Reading List.

1. (Gatt and Krahmer, 2018): traditional methods on natural language generation
2. (Radford et al., 2019): large-scale language models as unsupervised multitask learners with generative capabilities
3. (Khandelwal et al., 2019): example highlighting the rise of pretrained language models for neural text generation
4. (Holtzman et al., 2019): studying the dramatic effect of decoding strategies on the quality of machine text

[1] http://acl2019pcblog.fileli.unipi.it/?p=152

5. (Kusner et al., 2015): going beyond n-gram matching, using representation learning to evaluate generation

6. (Ranzato et al., 2015): introduction to exposure bias and training with sequence-level objective functions

7. (Bowman et al., 2016): variational autoencoders for language generation

8. (Holtzman et al., 2018): designing neural networks as scoring functions during decoding

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.

Antoine Bosselut, Asli elikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *Proceedings of the 16th Annual Meeting of the North American Association for Computational Linguistics (NAACL)*.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence movers similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760.

Elizabeth Clark, Yangfeng Ji, and Noah A Smith. 2018. Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Wanyu Du and Yangfeng Ji. 2019. An empirical comparison on imitation learning and reinforcement learning for paraphrase generation. *arXiv preprint arXiv:1908.10835*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

William Fedus, Ian Goodfellow, and Andrew M Dai. 2018. Maskgan: better text generation via filling in the_. *arXiv preprint arXiv:1801.07736*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Sergey Golovanov, Rauf Kurbanov, Sergey I. Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *ACL*.

Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. 2019. Quantifying exposure bias for neural language generation. *arXiv preprint arXiv:1905.10617*.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *ACL*.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596. JMLR. org.

Qiuyuan Huang, Zhe Gan, Asli Celikyilmaz, Dapeng Wu, Jianfeng Wang, and Xiaodong He. 2019. Hierarchically structured reinforcement learning for topically coherent visual story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8465–8472.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

Urvashi Khandelwal, Kevin Clark, Dan Jurafsky, and Lukasz Kaiser. 2019. Sample efficient text summarization using a single pre-trained transformer. *arXiv preprint arXiv:1905.08836*.

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. *arXiv preprint arXiv:1805.10627*.

Matt J. Kusner, Yongqiang Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *ICML*.

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *EMNLP*.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *CoRR*, abs/1511.06732.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

Dinghan Shen, Asli Celikyilmaz, Yizhe Zhang, Liqun Chen, Xin Wang, Jianfeng Gao, and Lawrence Carin. 2019. Towards generating long and coherent text with multi-level latent variable models. *arXiv preprint arXiv:1902.00154*.

Sandeep Subramanian, Sai Rajeswar Mudumba, Alessandro Sordoni, Adam Trischler, Aaron C Courville, and Chris Pal. 2018. Towards text generation with adversarially learned neural outlines. In *Advances in Neural Information Processing Systems*, pages 7551–7563.

Bowen Tan, Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric Xing. 2018. Connecting the dots between mle and rl for sequence generation. *arXiv preprint arXiv:1811.09740*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 166–177.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. In *EMNLP*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2018. Learning neural templates for text generation. *arXiv preprint arXiv:1808.10122*.

Ziang Xie. 2017. Neural text generation: A practical guide. *arXiv preprint arXiv:1711.09534*.

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. *arXiv preprint arXiv:1808.06945*.

Zachary M Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann, and Alexander M Rush. 2019. Encoder-agnostic adaptation for conditional language generation. *arXiv preprint arXiv:1908.06938*.