Non-Autoregressive Models for Fast Sequence Generation

Yang Feng[1,2] **Chenze Shao**[1,2]
[1] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] University of Chinese Academy of Sciences, Beijing, China
{fengyang,shaochenze18z}@ict.ac.cn

1 Tutorial Introduction

Autoregressive (AR) models have achieved great success in various sequence generation tasks (Bahdanau et al., 2015; Vaswani et al., 2017). However, AR models can only generate the target sequence word-by-word due to the AR mechanism and hence suffer from slow inference. Recently, non-autoregressive (NAR) models, which generate all the tokens in parallel by removing the sequential dependencies within the target sequence, have received increasing attention in sequence generation tasks such as neural machine translation (NMT, Gu et al., 2018), automatic speech recognition (ASR, Salazar et al., 2019), and text to speech (TTS, Ren et al., 2019).

Recently, non-autoregressive (NAR) models have received much attention in various sequence generation tasks, which generate all tokens in parallel by ignoring the sequential dependency within the target sequence. Gu et al. (2018) proposed the first NAR translation model for the efficient inference of neural machine translation, and NAR generation has subsequently been applied to a wide range of sequence generation tasks, where the two most successful application scenarios are ASR and TTS. The major challenge faced by NAR generation is the multi-modality problem: there may exist multiple correct outputs for the same source input, but the naive NAR model is unable to capture the multi-modal data distribution. Therefore, the direct application of NAR generation will usually lead to significant performance degradation compared to the autoregressive counterpart.

In this tutorial, we will provide a comprehensive introduction to non-autoregressive sequence generation. First, we start with the background of sequence generation, giving the motivation of NAR generation and the challenge faced by NAR models. We will briefly introduce the autoregressive generation mechanism and autoregressive sequence models that evolve from recurrent neural networks (Schuster and Paliwal, 1997) to self-attention networks (Vaswani et al., 2017). We point out their problems caused by the autoregressive mechanism, including exposure bias (Ranzato et al., 2016), error propagation, fixed generation direction, causal attention, and most importantly, the high inference latency. We will then introduce the NAR model that solves the above-mentioned problems by generating all target tokens in parallel, and point out the multi-modality challenge faced by NAR models (Gu et al., 2018).

Second, we will introduce research work that aims to improve the performance of NAR generation, mainly focusing on non-autoregressive translation in this part. The involved work covers efforts over knowledge distillation (Kim and Rush, 2016; Zhou et al., 2020; Sun and Yang, 2020; Ding et al., 2021; Shao et al., 2022b), better training objectives (Shao et al., 2019, 2020; Ghazvininejad et al., 2020; Du et al., 2021, 2022; Tu et al., 2020; Shao et al., 2021; Shao and Feng, 2022; Li et al., 2022b; Anonymous, 2023), latent modeling (Gu et al., 2018; Kaiser et al., 2018; Ma et al., 2019; Ran et al., 2021; Song et al., 2021; Shu et al., 2020; Bao et al., 2021, 2022), more expressive NAR models (Wang et al., 2017; Libovický and Helcl, 2018; Sun et al., 2019; Huang et al., 2022), improved decoding approaches (Lee et al., 2018; Ghazvininejad et al., 2019; Gu et al., 2019; Ran et al., 2020; Saharia et al., 2020; Deng and Rush, 2020; Geng et al., 2021; Stern et al., 2018, 2019; Xia et al., 2022; Shao et al., 2022a), etc.

Third, we will introduce NAR models on other sequence generation tasks, where the two most successful application scenarios are ASR and TTS. The idea of NAR generation was first pervading in ASR, where Graves et al. (2006) proposed the CTC network which predicts outputs independently, but the recurrent network architecture prevents it from parallel decoding. With the emergence of paralleliz-

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 30 - 35
December 7-8, 2022

able self-attention network (Vaswani et al., 2017), CTC-based NAR models soon became a promising direction in ASR (Higuchi et al., 2020; Chen et al., 2020). In TTS, parallel generation is particularly necessary due to the extremely large length of output sequence. The first attempt is Parallel WaveNet (Oord et al., 2018) which keeps the autoregressive mechanism but enables parallel generation with inverse autoregressive flow (Kingma et al., 2016). NAR models are subsequently proposed for TTS (Ren et al., 2019, 2020a; Prenger et al., 2019), which caught up with AR models in a short time and soon became the mainstream method for TTS.

We will also introduce other applications of NAR models like language modeling (Huang et al., 2021; Li et al., 2022a), image/video captioning (Gao et al., 2019; Yang et al., 2021), dialogue generation (Wu et al., 2020; Le et al., 2020), and even object detection (Carion et al., 2020). It is observed that NAR models perform well on some tasks but suffer from performance degradation on other tasks. This phenomenon can be explained from the perspective of multi-modality (Gu et al., 2018) or target token dependency (Ren et al., 2020b).

Finally, we will conclude this tutorial by summarizing the strengths and challenges of NAR models and discussing current concerns and future directions of NAR generation.

2 Type of Tutorial

The type of tutorial is cutting-edge. Non-autoregressive generation is a newly emerging topic, which has attracted increasing attention from researchers and achieved remarkable advancement in the past several years. This is the second tutorial on this topic in the history of ACL, EMNLP, NAACL, EACL, COLING, and AACL (Gu and Tan, 2022).

3 Tutorial Outline

Part I: Introduction (20 min)

- Autoregressive sequence generation
- Problems of AR generation
  - High inference latency
  - Exposure bias
  - Error propagation
- Non-autoregressive generation
- Multi-modality challenge

Part II: Non-Autoregressive Machine Translation (80 min)

- Knowledge distillation
- Training objectives
  - Token-level
  - Ngram-level
  - Sequence-level
- Latent modeling
  - Variational autoencoder
  - Vector quantization
  - Word alignment
- Expressive NAR models
  - CTC
  - DA-Transformer
- Decoding approaches
  - Iterative decoding
  - Semi-autoregressive decoding
  - Speculative decoding

Part III: Non-Autoregressive Sequence Generation (60 min)

- Non-autoregressive ASR
- Non-autoregressive TTS
- Other generation tasks
  - language modeling
  - Image/video captioning
  - Dialogue generation
  - Object detection
- What kind of tasks are NAR models good at?
  - Multi-modality
  - Target token dependency

Part IV: Conclusion (20 min)

4 Breadth

This tutorial will provide a comprehensive introduction to non-autoregressive sequence generation. We anticipate that at least 90% of the tutorial will cover work by other researchers.

5 Diversity

In the past, NAR sequence generation usually involves one or two languages. Recently, some researchers have found that NAR models are good at multilingual translation (Song et al., 2022), which may stimulate the progress of NAR generation in multilingual scenarios.

Yang Feng is a senior instructor and Chenze Shao is a junior instructor.

6 Prerequisites

The attendees have to understand the basics of neural networks and the sequence-to-sequence framework, including word embeddings, encoder-decoder models, and the Transformer architecture.

7 Reading List

We recommend attendees to read the following papers before the tutorial:

- Vaswani et al. (2017): the parallelizable Transformer network based on attention mechanisms.
- Gu et al. (2018): first propose non-autoregressive generation for parallel decoding and point out the multi-modality problem.
- Kim and Rush (2016): train the student model with the teacher output, alleviating the multi-modality by reducing data complexity.
- Shao et al. (2021): train NAR models with sequence-level objectives, which evaluate model outputs as a whole and optimize the overall translation quality.
- Shu et al. (2020): use latent variables to model the non-determinism in the translation process.
- Ghazvininejad et al. (2019): iteratively refine model outputs by repeatedly masking out and regenerating partial target tokens.
- Graves et al. (2006): the early exploration of non-autoregressive generation, and the proposed CTC loss is widely used in recent NAR models.
- Ren et al. (2019): non-autoregressive text-to-speech model, which matches autoregressive models in terms of speech quality.
- Ren et al. (2020b): a study on NAR models that analyzes the difficulty of NAR generation on different generation tasks

8 Tutorial Presenters

Yang Feng is a professor in Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS). She got her PhD degree in ICT/CAS and then worked in University of Sheffield and Information Sciences Institute, University of Southern California, and now leads the natural language processing group in ICT/CAS. Her research interests are natural language process, mainly focusing on machine translation and dialogue. She was the recipient of the Best Long Paper Award of ACL 2019. She served as a senior area chair of EMNLP 2021 and area chairs of ACL, EMNLP, COLING etc., and she is serving as an Action Editor of ACL Roling Review and an editorial board member of the Northern European Journal of Language Technology. She has given a tutorial in the 10th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC2021) and has been invited to give talks in NLPCC, CCL(China National Conference on Computational Linguistics) etc.

Chenze Shao is a fifth-year PhD student in Institute of Computing Technology, Chinese Academy of Sciences. His research interests are natural language processing and neural machine translation. His recent research topic is non-autoregressive (NAR) sequence generation. He has published papers on NAR generation in CL, ACL, EMNLP, NAACL, AAAI and NeurIPS.

9 Other Information

Technical Requirements This tutorial does not have special requirements for technical equipment.

Ethics Statement The technique of non-autoregressive generation improves the efficiency of text generation and may reduce the cost of generating malicious text.

Open Access. All of our tutorial materials can be shared in the ACL Anthology.

References

Anonymous. 2023. Fuzzy alignments in directed acyclic graph for non-autoregressive machine translation. In *Submitted to The Eleventh International Conference on Learning Representations*. Under review.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yu Bao, Shujian Huang, Tong Xiao, Dongqi Wang, Xinyu Dai, and Jiajun Chen. 2021. Non-autoregressive translation by learning target categorical codes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5749–5759, Online. Association for Computational Linguistics.

Yu Bao, Hao Zhou, Shujian Huang, Dongqi Wang, Lihua Qian, Xinyu Dai, Jiajun Chen, and Lei Li. 2022. latent-GLAT: Glancing at latent variables for parallel text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8398–8409, Dublin, Ireland. Association for Computational Linguistics.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer.

Nanxin Chen, Shinji Watanabe, Jesús Villalba, Piotr Żelasko, and Najim Dehak. 2020. Non-autoregressive transformer for speech recognition. *IEEE Signal Processing Letters*, 28:121–125.

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.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021. Understanding and improving lexical choice in non-autoregressive translation. In *International Conference on Learning Representations*.

Cunxiao Du, Zhaopeng Tu, and Jing Jiang. 2021. Order-agnostic cross entropy for non-autoregressive machine translation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2849–2859. PMLR.

Cunxiao Du, Zhaopeng Tu, Longyue Wang, and Jing Jiang. 2022. ngram-oaxe: Phrase-based order-agnostic cross entropy for non-autoregressive machine translation. *arXiv preprint arXiv:2210.03999*.

Junlong Gao, Xi Meng, Shiqi Wang, Xia Li, Shanshe Wang, Siwei Ma, and Wen Gao. 2019. Masked non-autoregressive image captioning. *arXiv preprint arXiv:1906.00717*.

Xinwei Geng, Xiaocheng Feng, and Bing Qin. 2021. Learning to rewrite for non-autoregressive neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3297–3308, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marjan Ghazvininejad, Vladimir Karpukhin, Luke Zettlemoyer, and Omer Levy. 2020. Aligned cross entropy for non-autoregressive machine translation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3515–3523. PMLR.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Jiatao Gu and Xu Tan. 2022. Non-autoregressive sequence generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 21–27.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11179–11189.

Yosuke Higuchi, Shinji Watanabe, Nanxin Chen, Tetsuji Ogawa, and Tetsunori Kobayashi. 2020. Mask ctc: Non-autoregressive end-to-end asr with ctc and mask predict. *Proc. Interspeech 2020*, pages 3655–3659.

Fei Huang, Jian Guan, Pei Ke, Qihan Guo, Xiaoyan Zhu, and Minlie Huang. 2021. A text {gan} for language generation with non-autoregressive generator.

Fei Huang, Hao Zhou, Yang Liu, Hang Li, and Minlie Huang. 2022. Directed acyclic transformer for non-autoregressive machine translation. In *Proceedings of the 39th International Conference on Machine Learning, ICML 2022*.

Lukasz Kaiser, Samy Bengio, Aurko Roy, Ashish Vaswani, Niki Parmar, Jakob Uszkoreit, and Noam Shazeer. 2018. Fast decoding in sequence models using discrete latent variables. In *Proceedings of the*

35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pages 2390–2399. PMLR.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29:4743–4751.

Hung Le, Richard Socher, and Steven C.H. Hoi. 2020. Non-autoregressive dialog state tracking. In *International Conference on Learning Representations*.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Brussels, Belgium. Association for Computational Linguistics.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. 2022a. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*.

Yafu Li, Leyang Cui, Yongjing Yin, and Yue Zhang. 2022b. Multi-granularity optimization for non-autoregressive translation. In *EMNLP 2022*.

Jindřich Libovický and Jindřich Helcl. 2018. End-to-end non-autoregressive neural machine translation with connectionist temporal classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3016–3021, Brussels, Belgium. Association for Computational Linguistics.

Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. 2019. FlowSeq: Non-autoregressive conditional sequence generation with generative flow. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4282–4292, Hong Kong, China. Association for Computational Linguistics.

Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pages 3918–3926. PMLR.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE.

Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. 2020. Learning to recover from multi-modality errors for non-autoregressive neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3059–3069, Online. Association for Computational Linguistics.

Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. 2021. Guiding non-autoregressive neural machine translation decoding with reordering information. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Event, February 2-9, 2021*, pages 13727–13735. AAAI Press.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020a. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

Yi Ren, Jinglin Liu, Xu Tan, Zhou Zhao, Sheng Zhao, and Tie-Yan Liu. 2020b. A study of non-autoregressive model for sequence generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 149–159.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3165–3174.

Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. 2020. Non-autoregressive machine translation with latent alignments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1098–1108, Online. Association for Computational Linguistics.

Julian Salazar, Katrin Kirchhoff, and Zhiheng Huang. 2019. Self-attention networks for connectionist temporal classification in speech recognition. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Chenze Shao and Yang Feng. 2022. Non-monotonic latent alignments for ctc-based non-autoregressive machine translation. In *Proceedings of NeurIPS 2022*.

Chenze Shao, Yang Feng, Jinchao Zhang, Fandong Meng, Xilin Chen, and Jie Zhou. 2019. Retrieving sequential information for non-autoregressive neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3013–3024, Florence, Italy. Association for Computational Linguistics.

Chenze Shao, Yang Feng, Jinchao Zhang, Fandong Meng, and Jie Zhou. 2021. Sequence-Level Training for Non-Autoregressive Neural Machine Translation. *Computational Linguistics*, pages 1–35.

Chenze Shao, Zhengrui Ma, and Yang Feng. 2022a. Viterbi decoding of directed acyclic transformer for non-autoregressive machine translation. In *Findings of EMNLP 2022*.

Chenze Shao, Xuanfu Wu, and Yang Feng. 2022b. One reference is not enough: Diverse distillation with reference selection for non-autoregressive translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3779–3791, Seattle, United States. Association for Computational Linguistics.

Chenze Shao, Jinchao Zhang, Yang Feng, Fandong Meng, and Jie Zhou. 2020. Minimizing the bag-of-ngrams difference for non-autoregressive neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 198–205. AAAI Press.

Raphael Shu, Jason Lee, Hideki Nakayama, and Kyunghyun Cho. 2020. Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8846–8853. AAAI Press.

Jongyoon Song, Sungwon Kim, and Sungroh Yoon. 2021. AligNART: Non-autoregressive neural machine translation by jointly learning to estimate alignment and translate. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1–14, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhenqiao Song, Hao Zhou, Lihua Qian, Jingjing Xu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2022. switch-GLAT: Multilingual parallel machine translation via code-switch decoder. In *International Conference on Learning Representations*.

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. 2019. Insertion transformer: Flexible sequence generation via insertion operations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5976–5985. PMLR.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. 2018. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems*, 31.

Zhiqing Sun, Zhuohan Li, Haoqing Wang, Di He, Zi Lin, and Zhihong Deng. 2019. Fast structured decoding for sequence models. In *Advances in Neural Information Processing Systems 32*, pages 3016–3026.

Zhiqing Sun and Yiming Yang. 2020. An EM approach to non-autoregressive conditional sequence generation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9249–9258. PMLR.

Lifu Tu, Richard Yuanzhe Pang, Sam Wiseman, and Kevin Gimpel. 2020. ENGINE: Energy-based inference networks for non-autoregressive machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2826, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Chong Wang, Yining Wang, Po-Sen Huang, Abdelrahman Mohamed, Dengyong Zhou, and Li Deng. 2017. Sequence modeling via segmentations. In *International Conference on Machine Learning*, pages 3674–3683. PMLR.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020. SlotRefine: A fast non-autoregressive model for joint intent detection and slot filling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1932–1937, Online. Association for Computational Linguistics.

Heming Xia, Tao Ge, Furu Wei, and Zhifang Sui. 2022. Lossless speedup of autoregressive translation with generalized aggressive decoding. *arXiv preprint arXiv:2203.16487*.

Bang Yang, Yuexian Zou, Fenglin Liu, and Can Zhang. 2021. Non-autoregressive coarse-to-fine video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3119–3127.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.