[15] Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. arXiv preprint arXiv:1909.05858 (2019).

[16] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. Retrieved from http://arxiv.org/abs/1412.6980

[17] Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 66–75. Retrieved from http://aclweb.org/anthology/P18-1007

[18] Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 66–71.

[19] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. Advances in Neural Information Processing Systems (NeurIPS) (2019).

[20] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the Variance of the Adaptive Learning Rate and Beyond. In International Conference on Learning Representations. Retrieved from https://openreview.net/forum?id=rkgz2aEKDr

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692, (2019). Retrieved from http://arxiv.org/abs/1907.11692

[22] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and Optimizing LSTM Language Models. In International Conference on Learning Representations. Retrieved from https://openreview.net/forum?id=SyyGPP0TZ

[23] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. An Analysis of Neural Language Modeling at Multiple Scales. CoRR abs/1803.08240, (2018). Retrieved from http://arxiv.org/abs/1803.08240

[24] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In Advances in Neural Information Processing Systems, 4694–4703.

[25] Michael Noseworthy, Ryan Lowe, Iulian V Serban, Yoshua Bengio, Iulian Vlad Serban, Nicolas Angelard-Gontier, 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), Association for Computational Linguistics, Stroudsburg, PA, USA, 1116–1126. DOI:https://doi.org/10.18653/v1/P17-1103

[26] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 1532–1543. DOI:https://doi.org/10.3115/v1/D14-1162

[27] Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), Association for Computational Linguistics, Florence, Italy, 7–14. DOI:https://doi.org/10.18653/v1/W19-4302

[28] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 1715–1725. DOI:https://doi.org/10.18653/v1/P16-1162

[29] Leslie N. Smith. 2017. Cyclical Learning Rates for Training Neural Networks. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 464–472. DOI:https://doi.org/10.1109/WACV.2017.58

[30] Leslie N Smith. 2018. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. arXiv.org (2018). Retrieved from http://arxiv.org/abs/1803.09820

[31] Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Madrid, Spain, 271–280. DOI:https://doi.org/10.3115/976909.979652

[32] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. CoRR abs/1910.03771, (2019). Retrieved from http://arxiv.org/abs/1910.03771

[33] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. CoRR abs/1609.08144, (2016). Retrieved from http://arxiv.org/abs/1609.08144

[34] Ming Yan, Maofu Liu, and Junyi Xiang. 2019. WUST at the NTCIR-14 STC-3 Dialogue Quality and Nugget Detection Subtask. In Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies.

[35] Hongwei Yong, Jianqiang Huang, Xiansheng Hua, and Lei Zhang. 2020. Gradient Centralization: A New Optimization Technique for Deep Neural Networks. arXiv preprint arXiv:2004.01461 (2020).

[36] Yang You, Igor Gitman, and Boris Ginsburg. 2017. Large batch training of convolutional networks. arXiv preprint arXiv:1708.03888 (2017).

[37] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. In International Conference on Learning Representations. Retrieved from https://openreview.net/forum?id=Syx4wnEtvH

[38] Zhaohao Zeng, Sosuke Kato, and Tetsuya Sakai. 2019. Overview of the NTCIR-14 short text conversation task: Dialogue quality and nugget detection subtasks. In Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies.

[39] Zhaohao Zeng, Sosuke Kato, Tetsuya Sakai, and Inho Kang. 2020. Overview of the NTCIR-15 Dialogue Evaluation (DialEval-1) Task. In Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies.

[40] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. 2019. Lookahead optimizer: k steps forward, 1 step back. In Advances in Neural Information Processing Systems, 9597–9608.

[41] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Final Report of the NTCIR-14 FinNum Task: Challenges and Current Status of Fine-Grained Numeral Understanding in Financial Social Media Data. In NII Conference on Testbeds and Community for Information Access Research, Springer, 183–192.

[42] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2020. Overview of the NTCIR-15 FinNum-2 Task: Numeral Attachment in Financial Tweets. In Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies.