# Importance of Data and Controllability in Neural Text Simplification

Wei Xu

*Georgia Institute of Technology, Atlanta, GA, U.S.A.*

## Abstract

Natural language generation has become one of the fastest-growing areas in NLP and a popular playground for studying deep learning techniques. Many variants of sequence-to-sequence models with complicated components have been developed. Yet, as I will demonstrate in this talk, creating high-quality training data and injecting linguistic knowledge can lead to significant performance improvements that overshadow gains from many of these model variants. I will present two recent works from my group on text simplification, a task that requires both lexical and syntactic paraphrasing to improve text accessibility: 1) a neural conditional random field (CRF) based semantic model [1, 2] to create parallel training data [3]; 2) a controllable text generation approach [4] that incorporates syntax through pairwise ranking and data argumentation.

In the first work, we show that the success of a text simplification system heavily depends on the quality and quantity of complex-simple sentence pairs in the training corpus, which are extracted by aligning sentences between parallel articles. To evaluate and improve sentence alignment quality, we create two manually annotated sentence-aligned datasets from two commonly used text simplification corpora, Newsela and Wikipedia. We propose a novel neural CRF alignment model which not only leverages the sequential nature of sentences in parallel documents but also utilizes a neural sentence pair model to capture semantic similarity. Experiments demonstrate that our proposed approach outperforms all the previous work on monolingual sentence alignment tasks by more than 5 points in F1. We apply our CRF aligner to construct two new text simplification datasets, Newsela-Auto and Wiki-Auto, which are much larger and of better quality compared to the existing datasets. A Transformer-based seq2seq model trained on our datasets outperforms other state-of-the-art approaches for text simplification.

In the second work, we explore how text simplification improves the readability of sentences through several rewriting transformations, such as lexical paraphrasing, deletion, and splitting. Current simplification systems are predominantly sequence-to-sequence models that are trained end-to-end to perform all these operations simultaneously. However, such systems limit themselves to mostly deleting words and cannot easily adapt to the requirements of different target audiences. In this paper, we propose a novel hybrid approach that leverages linguistically-motivated rules for splitting and deletion, and couples them with a neural paraphrasing model to produce varied rewriting styles. We introduce a new data augmentation method to improve the paraphrasing capability of our model. Through automatic and manual evaluations, we show that our proposed model establishes a new state-of-the art for the task, paraphrasing more often than the existing systems, and can control the degree of each simplification operation applied to the input texts.

Throughout the talk, I will also briefly cover some of our works on evaluation metrics [5], lexical simplification [6], and document-level simplification [7]. To conclude, I will discuss a few questions from the CLEF reviewers: “Whilst text simplification has a long history, recent advances have significantly increased the quality and this may have opened up novel real-world applications: Is the quality sufficient for operational systems? what are the applications that are currently within our grasp? what is the main

*CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania*

🖂 wei.xu@cc.gatech.edu (W. Xu)
🌐 https://cocoxu.github.io/ (W. Xu)
iD 0000-0002-7044-3232 (W. Xu)

Code and data are available at: https://github.com/chaojiang06/wiki-auto

barrier for wide-scale deployement (comparable to MT)? Can we formulate a challenge for the obvious next step in the evolution?”

**Keywords**
text simplification, paraphrasing, sentence alignment, word alignment, readability

## References

[1] C. Jiang, M. Maddela, W. Lan, Y. Zhong, W. Xu, Neural CRF model for sentence alignment in text simplification, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 7943–7960. URL: https://www.aclweb.org/anthology/2020.acl-main.709.

[2] W. Lan, C. Jiang, W. Xu, Neural semi-Markov CRF for monolingual word alignment, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), 2021.

[3] W. Xu, C. Callison-Burch, C. Napoles, Problems in current text simplification research: New data can help, Transactions of the Association for Computational Linguistics (TACL) 3 (2015) 283–297. URL: https://www.aclweb.org/anthology/Q15-1021.

[4] M. Maddela, F. Alva-Manchego, W. Xu, Controllable text simplification with explicit paraphrasing, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2021. URL: https://arxiv.org/abs/2010.11004.

[5] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics (TACL) 4 (2016) 401–415. URL: https://www.aclweb.org/anthology/Q16-1029.

[6] M. Maddela, W. Xu, A word-complexity lexicon and a neural readability ranking model for lexical simplification, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (ENNLP), 2018, pp. 3749–3760. URL: https://www.aclweb.org/anthology/D18-1410.

[7] Y. Zhong, C. Jiang, W. Xu, J. J. Li, Discourse level factors for sentence deletion in text simplification, Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) 34 (2020) 9709–9716. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6520.