Fine-tuning LLMs: Gradual text simplification in English

1st Semester of 2023-2024, github.com/dragosconst/MT_FT

Dragos Tantaru
dragos.tantaru@s.unibuc.ro

Rares Patrascu
rares.patrascu@s.unibuc.ro

Bogdan Raia
bogdan.raia@s.unibuc.ro

Abstract

Abstract: We perform article-level simplification on the newstela english news corpus from 2016. We use Mistral to test different levels of simplification, from least simplified to most. We try different sets of parameters and end up obtaining decent results.

1 Introduction

Text simplification is a common topic in natural language processing to make understanding of a text easier while maintaining the meaning. This can be of great importance in regard to accessibility, making text more accessible to a larger audience: enabling people with cognitive, verbal or learning disabilities to understand a larger corpus of text (Carroll et al., 1999) (Evans et al., 2014) (Rello et al., 2013). It can also be useful in education, where it can provide reading assistance to children (Kajiwara et al., 2013), or online media platforms, where it's important to have easily digestible content.

Text simplification has also been shown to help increase the performance on natural language processing tasks, such as parsing (Chandrasekar et al., 1996), information extraction (Miwa et al., 2010) and machine translation (Chen et al., 2012). A large body of work has been done on the task, from statistical machine translation approaches (Xu et al., 2016), to seq2seq neural machine translation models (Nisioi et al., 2017), to transformer based models (Jiang et al., 2020).

Similar approaches have been tried before: by using the BERT architecture (Qiang et al., 2020) obtained state-of-the-art results on three different benchmarks, or by using control tokens combined with a T5 model (Sheang and Saggion, 2021) also performed over current state-of-the-art.

Dragos was responsible for writing the fine-tuning process, testing and debugging the generate method; Rares handled the evaluation script and metrics and supported Dragos if he had issues, while Bogdan wrote the paper and assisted the other authors.

2 Approach

The dataset we used is Newsela, a high-quality corpus of news written by professional editors to meet readability standards for children at multiple grade levels (Xu et al., 2015). It has a total of 10786 samples of varying lengths. It addresses many of the shortcomings found in text simplification done on Simple Wikipedia datasets, such as phrases that are either misaligned or lack real simplification.

Example of two such levels of simplification:

Grade 0: *Will the next egg you crack come from a chicken raised in a roomier barn?*
Foodies and farmers are in unusual agreement on the answer: If not now, then soon enough.
Both say McDonald's recent decision to transition to "cage-free" eggs for its McMuffins and other menu items was a tipping point in the $9 billion egg industry, which still produces 96 percent of its eggs in barns full of stacked wire cages.

Grade 2: *Will the next egg you eat come from a chicken raised in a roomier barn?*
Foodies and farmers say it probably will pretty soon.
Both say McDonald's recent decision to switch to "cage-free" eggs will change the $9 billion egg industry. Right now, egg companies still produce 96 percent of eggs in barns full of stacked wire cages.

Grade 4: *McDonald's has changed how it will buy eggs. It decided to only buy "cage-free" eggs. Those eggs come from hens that have more room. They can move around. The hens are not stuck in small wire cages.*

Fine-tuning

We used Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023) and flash

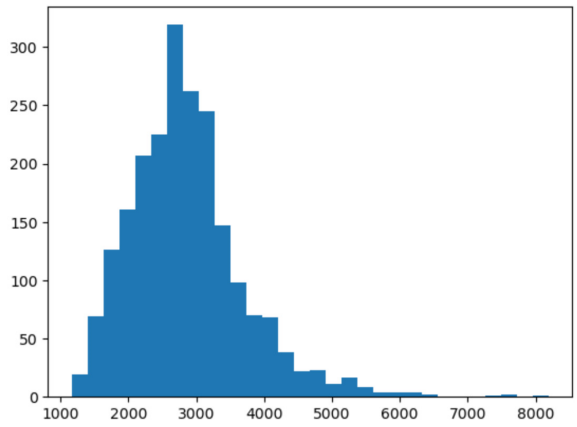

Figure 1: Context length of Newsela

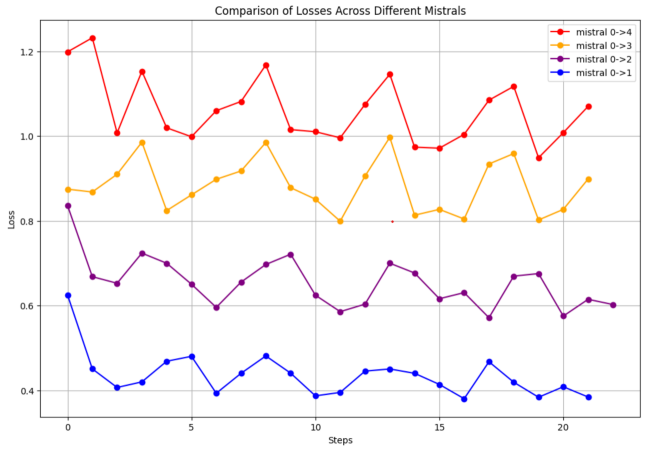

Figure 2: Mistral losses

attention in order to manage to fit the model on a 24GB RTX 3090 GPU.

We settled on using 4096 context length, as it covered most of our dataset, excluding only about $\approx 6\%$, as can be noted from Figure 1.

In terms of parameters, we used batch size of 1 with 8 gradient accumulation steps, fine tuned for a single epoch, with a learning rate of $3e-4$ using a cosine learning rate scheduler and an AdamW optimizer. For QLoRA we used rank 32, an alpha of 128 and dropout of 0.1.

Prompting

We used the following prompting scheme:

- For $0 \rightarrow 1$ simplification scheme, used *Slightley simplify the following text, as to make it more accesible*.
- For $0 \rightarrow 2$ simplification scheme, used *Moderately simplify the following text, as to make it accesible even for English speakers with just below decent language skills*.
- For $0 \rightarrow 3$,4 simplification schemes, used *Heavily simplify the following text, as to make it accesible even for English speakers with very limited language skills*.

Data types

For learning, we used int8 parameters, but for inference it took a really long time per sample (approx. 5 min). We had an idea and used bf16 for inference, which led to a speedup of about 80%.

Results

Loss-wise, we can see that the losses in Figure 2. follow a somewhat sinusoidal pattern, but that they have an overall declining trend. Metric-wise, we computed some classic metrics commonly found in text simplification literature, such as BLEU, SARI and ROUGE, but also the more modern COMET.

Mistral 7B	BLEU	ROUGE	COMET	SARI
$0 \rightarrow 1$	**0.546**	**0.674**	**0.888**	40.911
$0 \rightarrow 2$	0.349	0.492	0.874	**42.020**
$0 \rightarrow 3$	0.153	0.280	0.851	41.718
$0 \rightarrow 4$	0.074	0.209	0.832	41.163

Table 1: Mistral results on different levels of simplification

It can be seen, as expected, that the lowest level of simplification performs the best on almost all levels, while the highest performs the worst. Notably, COMET and SARI don't scale nearly as bad as BLEU and ROUGE do.

3 Limitations

We were heavily limited in terms of what models we can try out, due to a combination of limited GPU resources and very large models. This also implies we couldn't test the models with different numerical precision types, nor that we could try larger models.

Regarding similar sized models, we only tried Llama 2 7B as another model, which obtained very similar results - there could be additional work done regarding 1-10B parameter models.

Our work is also limited in terms of how we process the text - there could be other work to include

control tokens such as in (Sheang and Saggion, 2021) or the CRF aligner (Jiang et al., 2020).

4 Conclusions and Future Work

Overall we liked the project - Rares and Bogdan hadn't worked as much with huggingface, so it was an opportunity to learn something new. Main complaint was about timing, as it overlapped with other projects (and work) and couldn't properly focus on completing the project.
It was also a good team effort, as everyone took a substantial amount of the work and nobody ended up being swallowed and isolated in their tasks.
Regarding learning, for Bogdan it was helpful to understand more of the MT literature, especially regarding Text simplification. Further research can be done by using different, larger models, e.g. Mixtral 8x7 B, Llama 2 70B or Bloom, to test out how they perform and compare them to our results.

References

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270, Bergen, Norway. Association for Computational Linguistics.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Han-Bin Chen, Hen-Hsen Huang, Hsin-Hsi Chen, and Ching-Ting Tan. 2012. A simplification-translation-restoration framework for cross-domain SMT applications. In *Proceedings of COLING 2012*, pages 545–560, Mumbai, India. The COLING 2012 Organizing Committee.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Richard Evans, Constantin Orăsan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 131–140, Gothenburg, Sweden. Association for Computational Linguistics.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. 2010. Entity-focused sentence simplification for relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 788–796, Beijing, China. Coling 2010 Organizing Committee.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders.

Luz Rello, Clara Bayarri, Azuki Gòrriz, Ricardo Baeza-Yates, Saurabh Gupta, Gaurang Kanvinde, Horacio Saggion, Stefan Bott, Roberto Carlini, and Vasile Topac. 2013. Dyswebxia 2.0! more accessible text for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, New York, NY, USA. Association for Computing Machinery.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.