

We present a grammatical error correction system for correcting the Czech language. The system is based on the neural machine translation approach. We utilize the Transformer architecture, which depends on a large amount of annotated data. Given that for most languages, including Czech, there is not enough annotated data available, we opt to generate synthetic data with artificial errors. We generate not only using simple language-independent errors, but we also introduce typical Czech errors. To facilitate quick experimentation, we develop a flexible training pipeline capable of real-time data generation. Consequently, we evaluate the effect of several proposed improvements such as oversampling of language domains or a choice of data source for synthetic generation. Our best-performing model achieves state-of-the-art results in the Czech language for comparable model size. The implementation is released on GitHub at https://github.com/petrpechman/czech_gec/tree/MasterThesis_PechmanPetr_2024.