Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Petr Pechman		
Název práce	Czech Grammar Error Correction		
Rok odevzdání	2024		
Studijní program	Informatika	**Studijní obor**	Umělá inteligence

Autor posudku	Milan Straka	**Role**	Vedoucí
Pracoviště	Institute of Formal and Applied Linguistics		

Text posudku:

The goal of the thesis was to develop a state-of-the-art system for grammar error correction specifically for Czech. The thesis starts by describing the used evaluation methods, existing dataset, and existing approaches in Chapter 1. The Chapter 2 describes the system proposed and implemented by the author, and the experimental results are presented in Chapter 3. The main contributions of the thesis are:

- The author implemented a pipeline for introducing synthetic errors and training and evaluation of a GEC model [Chapter 2]. The pipeline has several distinctive properties:
  - The synthetic errors can be generated on the fly using multiple processes.
  - The training and inference can run on multiple GPUs.
  - The training data are batched effectively according to their length.
  - The CPU-based part of evaluation can run in parallel to GPU utilization.
- The author proposed several approaches for synthetic error generation – using Czech morphological and derivational dictionaries [Section 2.2.1], and introducing typical Czech-specific errors categorized by the author [Sections 2.2.2, 2.3].
- Most importantly, the author performed an extensive evaluation of many components and hyperparameters of GEC training [Chapter 3]. This evaluation required considerable engineering effort and expertise, lot of computational resources of many thousands of GPU hours, and its extent significantly surpasses the level required for a Master thesis. The results are compared with existing systems [Table 3.13], and the main findings are summarized clearly in the conclusion. To summarize the interesting outcomes:
  - The length of the pretraining phase positively impacts the performance during finetuning, even if the metrics in the pretraining phase no longer improve.
  - The quality of the plain texts used for synthetic error generation is more important than their size.
  - Training a model from scratch is competitive to finetuning an existing one.
  - When multiple domains are available during finetuning, careful balancing is needed to obtain good results on all of them.

The thesis is written in very good English, making it relevant to wide audience. Overall, I consider the thesis to be of high quality, with the author demonstrating rigorous scientific work in an area of international significance. Finally, the source code used in all experiments has been released, allowing replicability and follow-up work.

I recommend the thesis to be defended.

Práci doporučuji **k obhajobě.**

Práci nenavrhuji **na zvláštní ocenění.**

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 31. května 2024 **Podpis**