Bachelor Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis author	Jan Provazník
Thesis title	Textual Ciphers as a Tool for Better Understanding the Transformers
Year submitted	2024
Study program	Computer Science
Specialization	Artificial Intelligence

Review author	Zdeněk Kasner	Reviewer
Department	Institute of Formal and Applied Linguistics	

Overall

	good	OK	poor	insufficient
Assignment difficulty	X	X		
Assignment fulfilled	X	X		
Total size *… text and code, overall workload*		X		

The thesis uses the task of deciphering textual ciphers for studying the Transformer model, a powerful neural network architecture underpinning recent neural language models. While the Transformer is not an appropriate tool for this task (as the ciphers can be easily cracked with existing cryptographic tools), the task helps to understand the model behavior in a controlled setting.

The student focuses on the ByT5 model, which uses UTF-8 bytes as text units. This aspect is important, as it helps to remove the influence of the specific subword vocabulary. The goal of the thesis is to understand the influence of the *properties of the input text* on the *performance of the model*. The properties are various measurable aspects of the input text – for example, the text length in characters, the depth of the dependency tree, or the text perplexity.

For the experiments, the student selects two substitution ciphers: Vigenère and Enigma. To create the data for finetuning the Transformer model, the ciphers are applied to English, German, and Czech news data. Separately from the fine-tuning process, the texts are also automatically annotated with the selected properties using available tools. The properties are subsequently used as input features for traditional machine learning algorithms, which are trained to predict the character error rate of the finetuned Transformer model.

The results from the experiments are negative: the linguistic properties of text (at least under the specified assumptions) do not seem to be predictive of the model performance. Despite that, the results still provide several insights. The model seems to be capable of learning to decrypt the Vigenère cipher but not the Enigma cipher, which gives us some insights into the intrinsic difficulty of the task. Fine-tuning the model for decrypting texts in Czech and German was also more difficult than for English, showing that the disproportionately large portion of English pre-training data helps the model even for this very specific task.

Overall, the goals of the thesis are understandable, the experimental setup is clearly presented, and the experiments are well executed. I appreciate the specific assumptions at the end of Section 2.2.1, which help to denote the scope of the thesis and the applicability of its results. Under these assumptions, the linguistic properties do not help to predict the model performance – which is, however, a valuable research outcome in itself.

(continues on the next page)

The results hint at several directions for future work, such as exploring the internal mechanisms the model uses to decrypt the Vigenère cipher, examining the learnability of decrypting the Enigma cipher, or analyzing the performance of larger models.

Questions:

- You emphasize multiple times that the task was selected to "remove interference with cultural aspects of meaning". Yet, you use a model pre-trained on a large-scale web corpus and look for the links between model performance and linguistic properties of natural languages (including, for example, the GPT-2 perplexity, which undeniably encodes a degree of text semantics). If you wanted to remove cultural inference, why have you not used a vanilla (non-pretrained) Transformer model and synthetically generated data, looking only for more generic properties related to the character distribution?

  (Note: I like the goal of the thesis, but I would suggest doing the opposite: emphasize that you were looking for language-specific aspects of the text, such as its syntax and semantics, that can help the model decipher the text – that would be a non-intuitive and interesting result.)

- How do you explain the predictive power of text length on the model performance, especially at the early training stages? Is the model just better at guessing the character distribution for specific sentences, or is there a deeper link?

Thesis Text

	good	OK	poor	insufficient
Form *… language, typography, references*	X			
Structure *… context, goals, analysis, design, evaluation, level of detail*	X			
Problem analysis	X			
Developer documentation	X			
User Documentation	X			

The thesis is written in English and has 34 pages of content. The thesis follows the logical structure appropriate for a research work. The text is fluent and generally gramatically correct, the figures and tables support the results, and the experiments are presented at appropriate level of detail.

I have only a minor objection towards Section 1.4.2 (Related work on decipherment modeling), which is crucial to understand relevance of this work. I think it would deserve more attention: for example, I am missing any links between the related work and this thesis.

The instructions for running the experimental code and replicating the experiments are included in Appendix A.

Thesis Code

	good	OK	poor	insufficient
Design *… architecture, algorithms, data structures, used technologies*	X			
Implementation *… naming conventions, formatting, comments, testing*	X			
Stability	X			

The code is provided in the Github repository linked in the thesis. The code consists of Python scripts, Jupyter notebooks, and a set of scripts for running the code on a computational cluster. I was able to run the code by following the instructions in the README file. All the code is well documented with docstrings, in-line comments, and type hints.

Overall grade	Výborně
Award level thesis	Ne

Date

Signature