Can I start? Ok

Well. Let's start. My name is Dmitry Pogrebnoy and I would like to present my work titled "Development of an automatic spelling correction tool for analyzing clinical text in Russian".

Thanks to impressive advances in machine learning, it has become possible to apply various predictive and decision-making models in medicine. In healthcare, such models are often based on electronic texts of patients' medical records. The quality of such models strongly depends on the quality of the original medical records, which are usually plain text. Such records often contain a lot of spelling errors, which significantly reduce the quality of the final models. A high-quality tool for automatic correction of spelling errors will be able to fix this problem and increase the quality of the models without additional costs.

Therefore, the purpose of this work is to design a method and implement a tool for automatic correction of spelling errors for the analysis of medical texts in Russian. The tool should accept the raw medical text and return the corrected text with a minimum number of errors.

The following tasks were set for this semester. The first task is design the spelling correction process. Then the task is to design the architecture of a new spellchecking tool. After that, there is a task to implement a prototype of the tool. And the last task is conduct testing of the developed tool.

Let's take a look at the spelling correction process. The process diagram is shown on the screen. First, the medical text is splitted into tokens. Then, for each token, it is checked whether it is suitable for correction. If the token is not suitable, then it gets into the final text as is. Otherwise it is checked whether it is correct or not. If the token is correct, then it also gets into the final text without changes. If the token is incorrect, then a list of candidates is generated from the prepared index and the most suitable candidate for correction is selected using a language model based on a machine learning. The corrected result is included into the final text. After all tokens are processed, they are assembled back into a single text and the finished result is returned from the tool.

In order to implement this process, the following tool architecture was designed. The architecture diagram is shown on the screen. The architecture of the tool consists of seven components. The main component is the Spellchecker Manager, which is responsible for coordinating other components and implementing high-level logic. The Preprocessor and PostProcessor components are responsible for splitting the incoming text and assembling the result. The Dictionary component contains a dictionary of correct words, which allows to check the correct word or not. Edit Distance Index component allows to optimize and speed up the calculation of the editing distance required to generate candicates for fixing an incorrect word. Error model is responsible for generating candidates for fixing incorrect words. And the last component Language Model, based on the fine-tuned RuRoberta model, ranks candidates for fixing and selects the most suitable word for correction. This architecture makes it possible to independently develop each component and flexibly implement the correction process.

Let's take a look at the test results. This table shows the results of the test for correcting a single incorrect word by various popular open source tools. The results of the developed prototype are presented at the bottom of the table. It’s named as New Tool for CPU and GPU modes. As you can see, the prototype shows an average result in precision and rather low performance. Such prototype metrics are explained by the fact that the heavy-weight RuRoberta model is used for ranking, which works well on words with context, but loses in tests on single words.

However, in another test, when an incorrect word has context, the precision and performance metrics look much better. Nevertheless, I believe that these metrics can be significantly improved in the future.

So let 's move on to the conclusion

As a result, over the past semester, the new spelling correction process and the architecture of the new spellchecker tool are designed. Also the prototype of the spellchecker tool is implemented and testing of the developed tool is conducted. In addition, this project was presented at the eleventh congress of young scientists.

In the future it is planned to improve and optimize the spelling correction process. It is also planned to test other machine learning models as a language model and conduct wider testing. In addition, it is also planned to measure the effect of using the developed tool on other medical models.

Well, that's all plans for now and that's all I have. So I am ready to answer your questions.