Can I start? Ok

Well. Let's start. My name is Dmitry Pogrebnoy and I would like to present my work titled "Machine learning technology for correcting electronic medical texts in Russian".

There are many predictive and decision-making models in health care. 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 unstructured plain text. So such records usually 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 medical texts in Russian.

The following tasks were set for this semester. The first task is to collect and prepare dataset for training language models. Then the task is to select and fine-tune BERT models for ranking edit candidate task. After that, there is a task to conduct extensive testing of the developed tool. And the last task is to assemble the tool into a package and publish it.

First of all, let's look at the architecture of the tool. The tool consists of seven components. So the most important components here are the Error Model and Language Model. The Error Model generates candidates for editing, and the Language Model ranks them and chooses the most suitable one. The precision of the entire spellechecker mostly depends on how well the Language Model ranks. Therefore, it was important to experiment with such models.

In order to train a language model, I needed to collect a dataset. So I assembled and prepared a dataset with medical texts for training language models. I used two public datasets and two private ones. And all four datasets were pre-processed and combined into one final dataset.

So resulted dataset was used to fine-tune the language models. I selected three basic BERT models of different sizes and fine-tuned them to rank candidates. Well, the fill mask task was used for fine-tuning. And in addition, before fine-tuning, the DistilBert model was converted from multilingual to monolingual model for Russian language. So as a result, all three fine-tuned models were published on the Hugging Face service.

Let's look at the tests. So I made two tests. One test with single incorrect words, and another test with contexts around the incorrect words. And the incorrect words are the same in both tests.

Let's take a look at the test results. This table shows the results of the test for correcting a single word by various popular open source tools. The results of the developed tool are presented at the bottom of the table in green section. Well, as you can see, the new tool shows an average result in error precision and rather low performance, but the new tool achieve the highest lexical precision. However, the new tool uses a language model that takes into account the context around the incorrect word, so the tool can work much better with words with context. So let’s check it.

And next table shows the results of the test for word with context. In this test, the new tool reveals itself and outperforms other tools in terms of precision metrics. Despite the high precision score, the new tool has average performance and doesn't fall much in this metric compared to competitors~~.~~

So let’s go next. I also assembled the developed tool into a pip package and published a Beta version of the tool. This package contains the source code and necessary classes, as well as a dictionary of correct words. Also the package does not contain models and they are downloaded automatically when needed.

So let 's move on to the conclusion

As a result, over the past semester, the dataset for training language models is collected. Also three different BERT models are fine-tuned for ranking task. In addition extensive testing of the developed tool is conducted. And the pip package with the new tool is assembled.

In the future it is planned to improve and optimize the spelling correction process. It is also planned to try to fine-tune smaller language models and test the tool with them. In addition, it is also planned to evaluate the effect of the developed tool on medical models.

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