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".

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 unstructured plain text. So 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 medical texts in Russian. The tool should accept the raw medical text and return corrected text with a minimum number of errors.

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. The Preprocessor component splits the incoming text into tokens and transforms it into an internal representation. The Dictionary component then detects whether the word is correct or not. After that, candidates for correction are generated for the incorrect word using the Error Model component. The Edit Distance Index component speeds up this process greatly with the use of a pre-calculated index. The Language Model then ranks the candidates and chooses the most suitable one. And the corrected text is then assembled and returned as a result.

Well, one important component is the language model. The precision of the entire tool depends on how accurate the ranking is. 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. The public ones are RuMedNLI and RuMedPrimeData datasets. And the private ones are the dataset from Almazov National Medical Research Center and the dataset from the Research Institute of the Russian Academy of Sciences. 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 for editing to correct an incorrect word. Three models of different sizes were chosen to study how much the precision of spellchecker correction would depend on the size of the ranking model. 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. This operation reduced the size of the model by half. 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. In these tests, I evaluated error precision, that is, how many the incorrect word is correctly corrected, as well as lexical precision, that is, how many of the correct words remain untouched. In addition, I evaluated the performance of the tools, that is, how many words per second they process.

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. So error precision is very close to Aspell-python tool, and the other two precision metrics are noticeably higher. Despite the high precision score, the new tool has average performance and doesn't fall much in this metric compared to competitors. As a result, I can say that the new tool is outperform the existing tools in correcting words in full medical texts.

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. And it is important to note that the package does not contain models and they are downloaded automatically when needed from the open repository on Hugging Face. So in this way, I managed to keep the size of the package small and provide access to prepared models.

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.