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. The first task is to perform an overview of the Russian medical texts correction. Then the task is to analyze existing solutions for correcting Russian texts. After that, there is a task to design a new method for correcting spelling errors. Then a task is to design the architecture and implement a new spelling correction tool. After that the task is to conduct approbation of the developed tool. And the last task is to compare results of the developed tool and existing ones.

First of all, this work only considers spelling errors of six types. Examples of these errors are shown on the screen. The first 4 types of errors are related to letters and the last two types of errors are related to spaces.

There are several well-known tools for correcting spelling errors in Russian. However, none of them is intended to correct errors in medical texts. And besides, none of them uses advanced language models to improve the quality of corrections. So this work closes this gap.

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 incorrect, then a list of candidates is generated from the prepared index and the most suitable candidate is selected. The corrected result is included into the final text. After all tokens are processed, they are assembled into a single text and the finished result is returned from the tool.

Let's look at the architecture. 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. Error model is responsible for generating candidates for fixing incorrect words. The Edit Distance Index component optimizes and speeds up the generation of editing candidates. And the last component is a Language Model that ranks candidates for correction and selects the most suitable word for correction.

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. Also, two existing BERT models were adapted for use in the new tool.

The developed tool is intended for use only for medical texts in Russian. Preferably for anamnesis and medical histories of patients. The tool should be used in the text preprocessing pipeline before any preprocessing steps. It is also advisable to make sure that everything is in order after applying the tool.

Let's look at the tests. So I made three tests. One test with single incorrect words, another test with contexts around the incorrect words, and the last test with real anamnesis.

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 the 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 achieves the highest lexical precision. However, the new tool uses a language model that takes into account the context around the incorrect word, so it can work much better with words with context.

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

In the test with real anamneses, the new tool also showed a best result in the ratio of correct to unnecessary corrections. Here, dashes mean that after applying the tool, the original text changed so much that it was impossible to count the metrics. It is also worth noting that LanguageTool made the greatest number of correct corrections, which is two more than the best result of the new tool. So it’s not so many.

So let’s go next. Also I 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 doesn’t contain models and they are downloaded automatically when needed.

So let 's move on to the conclusion

The following tasks were done. Overview of the Russian medical texts correction is performed. Existing solutions for correction of Russian texts are analyzed. The new method of correcting spelling errors in Russian medical texts is designed. The new spelling correction tool is designed and implemented. The approbation of the developed tool is conducted. Results of the developed tool and existing ones are compared.

Well, that's it from me. Here are links to project parts. And I am ready to answer your questions.