



Machine learning technology for correcting electronic medical texts in Russian

Student: Dmitry Pogrebnoy, J41332c

Supervisor: Sergey Kovalchuk, PhD

Electronic medical records

- In healthcare, there are various predictive and decision-making models based on information from patients' medical records
- The quality of such models strongly depend on the quality of the source texts
- Electronic patient data is usually presented in plain text and contains a lot of spelling errors
- Spelling errors in the source texts greatly reduce the quality of the final models and therefore require correction

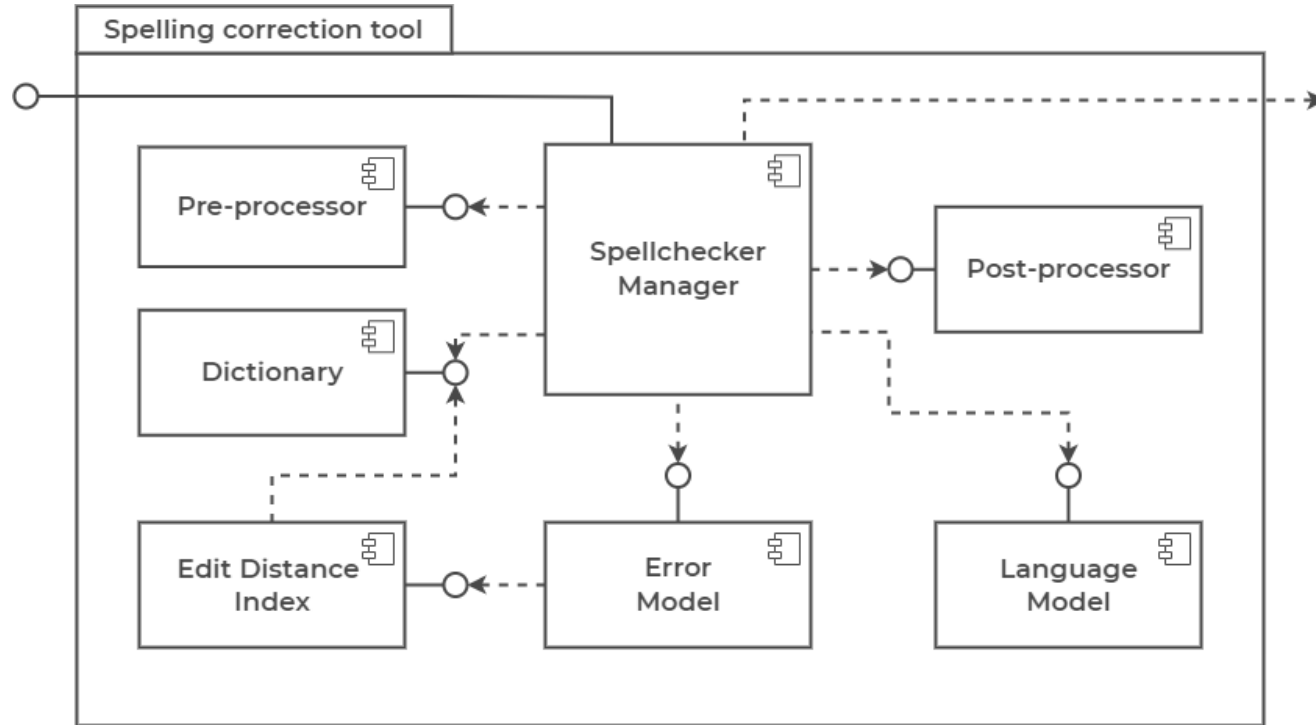
Goal and tasks

Goal: Design a method and implement an automatic spelling correction tool for clinical texts in Russian.

Tasks for third semester:

- Collect and prepare data for training language models
- Select and fine-tune BERT models for ranking task
- Conduct extensive testing of the developed tool
- Assemble the tool into a package and publish it

Tool architecture



Anamneses dataset

- Public datasets
 - RuMedNLI – 14716 records
 - RuMedPrimeData – 15249 records
- Private datasets
 - Almazov National Medical Research Center – 2355 records
 - Research Institute of the Russian Academy of Sciences – 161 records
- All datasets were pre-processed and assembled into final one
 - Tokenization and lemmatization
 - Stop words filtering

Fine-tune BERT models

- sberbank-ai/ruRoberta-large → MedRuRobertaLarge
 - Size – 1.4 Gb
- distilbert-base-multilang-cased → MedDistilBertBaseRuCased
 - Converted from multilang to Russian model
 - Size - 217 Mb
- cointegrated/rubert-tiny2 → MedRuBertTiny2
 - Size – 117 Mb
- Train/test/eval datasets – 0.8/0.1/0.1
- All models are published on the Hugging Face repository

Word tests internals

- Single test - error and lexical precision
 - 100 test sample per each error type
- Context test - error and lexical precision
 - 100 test samples per each error type
 - 10 words in each sample
 - One of ten words is incorrect, other words are correct
 - Same incorrect words as in single test
- Performance
 - Laptop with Ubuntu 20.04
 - 24 GB RAM and Intel Core i5-10750H CPU @ 1.60GHz * 12

Single word test

Tool	Error precision	Lexical precision	Average precision	Average words per second
Aspell-python	0.86	0.859	0.859	283.7
PyHunspell	0.812	0.539	0.675	9.4
PyEnchant	0.829	0.541	0.685	20
LanguageTool-python	0.762	0.904	0.833	25.1
PySpellChecker	0.354	0.86	0.607	3.4
SymSpellPy	0.399	0.813	0.606	9702.8
Jumspell	0.267	0.947	0.607	2552.1
Tool (CPU, MedDistilBert)	0.701	0.981	0.841	12.7
Tool (GPU, MedDistilBert)				39.7

Context word test

Tool	Error precision	Lexical precision	Average precision	Average words per second
Aspell-python	0.739	0.93	0.835	357.3
PyHunspell	0.706	0.719	0.713	11.8
PyEnchant	0.721	0.719	0.72	24.3
LanguageTool-python	0.727	0.942	0.835	43.6
PySpellChecker	0.304	0.868	0.586	6.7
SymSpellPy	0.37	0.913	0.642	26060.2
Jumspell	0.307	0.969	0.638	4322.3
Tool (CPU, MedDistilBert)	0.734	0.984	0.861	45.47
Tool (GPU, MedDistilBert)				134.453

Python package

- Assembled the pip python package
- Package contains
 - Source code
 - Dictionary with correct words
 - No models included
- Models are loaded dynamically as needed
- Published package name – [medspellchecker](#)

Conclusion

- Dataset for training language models is collected
- Three different BERT models are fine-tuned for ranking task
- Extensive testing of the developed tool is conducted
- Package with new tool is assembled



Fine-tuned models

huggingface.co/DmitryPogrebnoy



Pip package

pypi.org/project/medspellchecker

Further plans

- Improve and optimize the spelling correction process
- Try to fine-tune smaller language models and test them
- Evaluate the effect of the developed tool on medical models

Thank you for your attention!

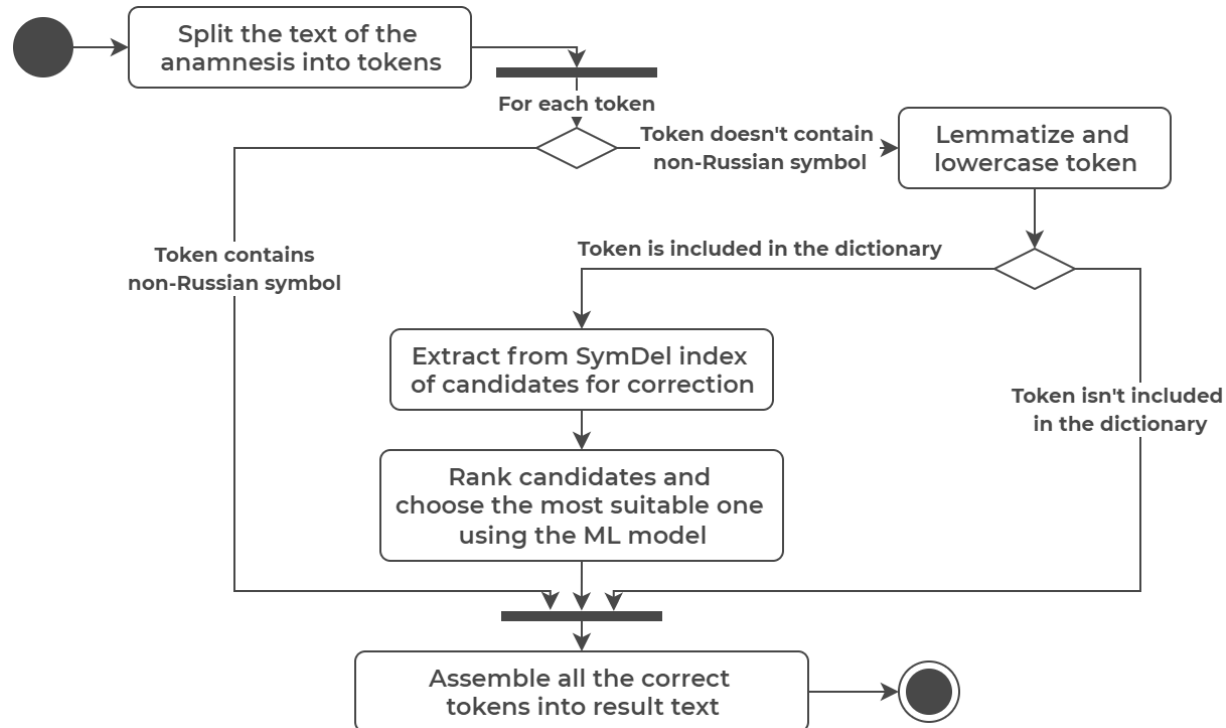
ITMO *re than a*
UNIVERSITY

Ваши контакты

Spelling errors in question

Type of mistake	Incorrect text	Correct text
Wrong characters	туб и ркулез	туб и <u>е</u> ркулез
Missing characters	туб о ркулез	туб <u>е</u> ркулез
Extra characters	туберк д улез	туберкулез
Shuffled characters	туб е <u>р</u> кулез	туб <u>е</u> ркулез
Missing word separator	острый туберкулез	острый <u>_</u> туберкулез
Extra word separator	туб_ е ркулез	туберкулез

Spelling correction process



Metrics

- **Error precision** – the ratio of the number of correctly corrected words to the total number of incorrect words
- **Lexical precision** – the ratio of the number of unchanged modified words to the total number of correct words
- **Average precision** – the average of error precision and lexical precision
- **Performance** – the number of words processed by the tool per second