ITMO UNIVERSITY

Saint Petersburg, Russia

Development of an automatic spelling correction tool for analyzing clinical text 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

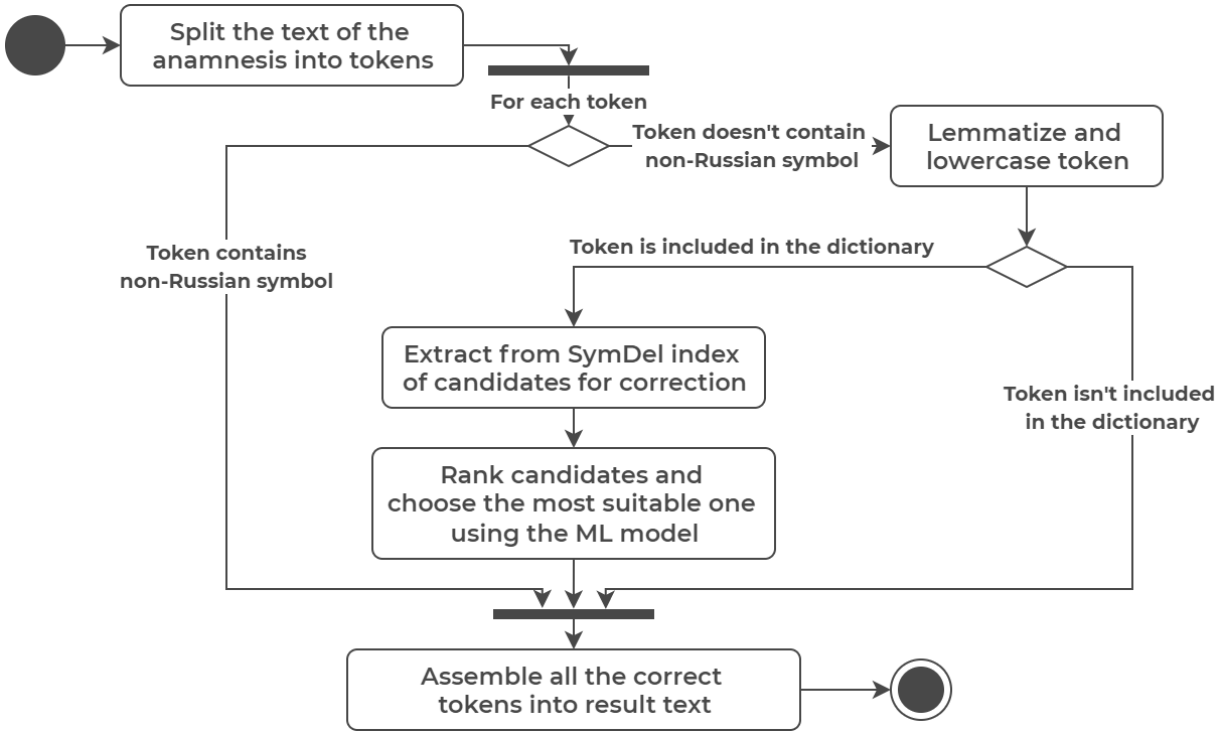
Goal and tasks

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

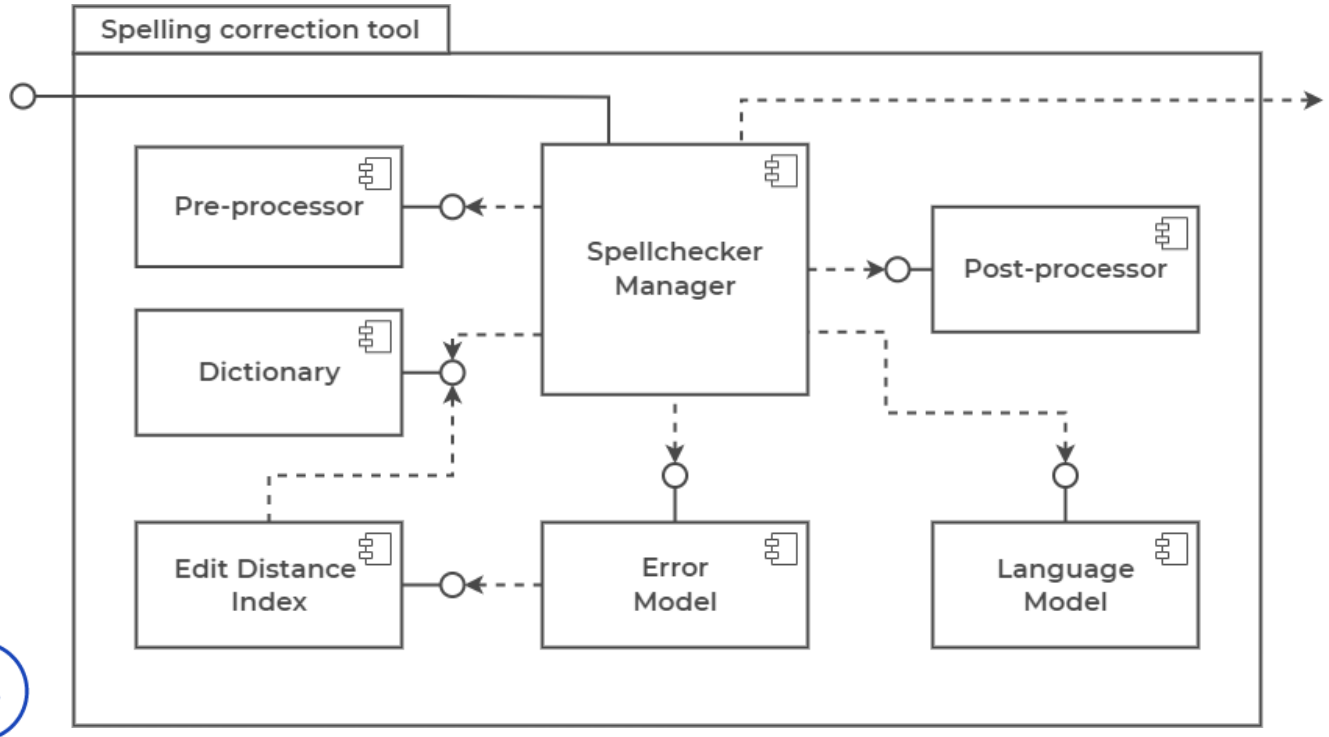
Tasks for second semester:

- Design the spelling correction process
- Design the architecture of a new spellchecker tool
- Implement prototype of new tool
- Conduct testing of the developed tool

Spelling correction process



Tool architecture



Single word test

Tool	Error precision	Lexical precision	Overall precision	Average words per second
Aspell-python	0.65	0.775	0.7125	353
PyHunspell	0.59	0.49	0.54	11.5
PyEnchant	0.6	0.455	0.5275	26.4
LanguageTool-python	0.64	0.845	0.7425	19.1
PySpellChecker	0.335	0.765	0.55	4.3
SymspellPy	0.42	0.78	0.6	15892.1
Jumspell	0.395	0.925	0.66	2043.2
New tool (CPU)	0.45	0.95	0.7	8.4
New tool (GPU)	0.45	0.95	0.7	12.5

Context word test

- Context test contains 60 samples
- Test sample – 10 words, one of which has a spelling error

Metric	CPU mode	GPU mode
Error precision	0.7	0.7
Lexical precision	0.98	0.98
Overall precision	0.84	0.84
Average words per second	51.5	100

Conclusion

- The new spelling correction process is designed
- The architecture of the new spellchecker tool is designed
- The prototype of the spell checker tool is implemented
- Testing of the developed tool is conducted

- The project was presented at the XI Congress of Young Scientists

Further plans

- Improve and optimize the spelling correction process
- Try to fine-tune other language models based on BERT
- Conduct more wide-ranging testing and approbation
- Measure the effect of the developed tool on medical models

Thank you for your attention!

www.ifmo.ru

ITMO more than a UNIVERSITY

Spelling errors in question

Type of mistake	Incorrect text	Correct text
Wrong characters	тубиркулез	туберкулез
Missing characters	туб□ркулез	туберкулез
Extra characters	туберкпулез	туберкулез
Shuffled characters	тубрекулез	туберкулез
Missing word separator	острый|туберкулез	острый_туберкулез
Extra word separator	туб_еркулез	туберкулез

Single word test internals

- **Error precision** - the ratio of the number of correctly corrected words to the total number of incorrect words
- 200 medical words with spelling errors were taken to compute error precision
- **Lexical precision** - the ratio of the number of unchanged modified words to the total number of correct words
- 200 correct medical words were taken to compute lexical precision
- **The performance test** was done on a laptop on Ubuntu 20.04 with 16 GB RAM and Intel Core i7-9750H CPU @ 2.60GHz * 12