RuMedSpellchecker: correcting spelling errors for natural Russian language in electronic health records using machine learning techniques

Dmitry Pogrebnoy, Anastasia Funkner, Sergey Kovalchuck

ITMO University

International Conference on Computational Science

3-5 July, 2023

Electronic health records

- Many medical models based on patients' medical records
- Quality of models depends mainly on the quality of source texts
- Patient records is a simple text
  - Unstructured
  - Hand-typed
  - Plain
- Such texts may contain a lot of spelling errors

Spelling errors in EHRs

- Toutanova et al. in their study[1] highlight two main causes of spelling errors.
  - Writer itself
  - Poor-quality typing devices

- Spelling errors greatly reduce the quality of the final models
- Fixing such errors will improve the quality of the medical models

[1]Toutanova, K., Moore, R.C.: Pronunciation modeling for improved spelling correction. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. p. 144–151. ACL '02, Association for Computational Linguistics, USA (2002).

Goal

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

Tool requirements:

- Accepts raw plain text
- Returns the text with the minimal spelling errors
- Can utilize the GPU to accelerate processing
- Comparable in performance to existing tools
- Outperforms existing open source tools
    - Quality of spelling errors correction of Russian medical texts

Spelling errors

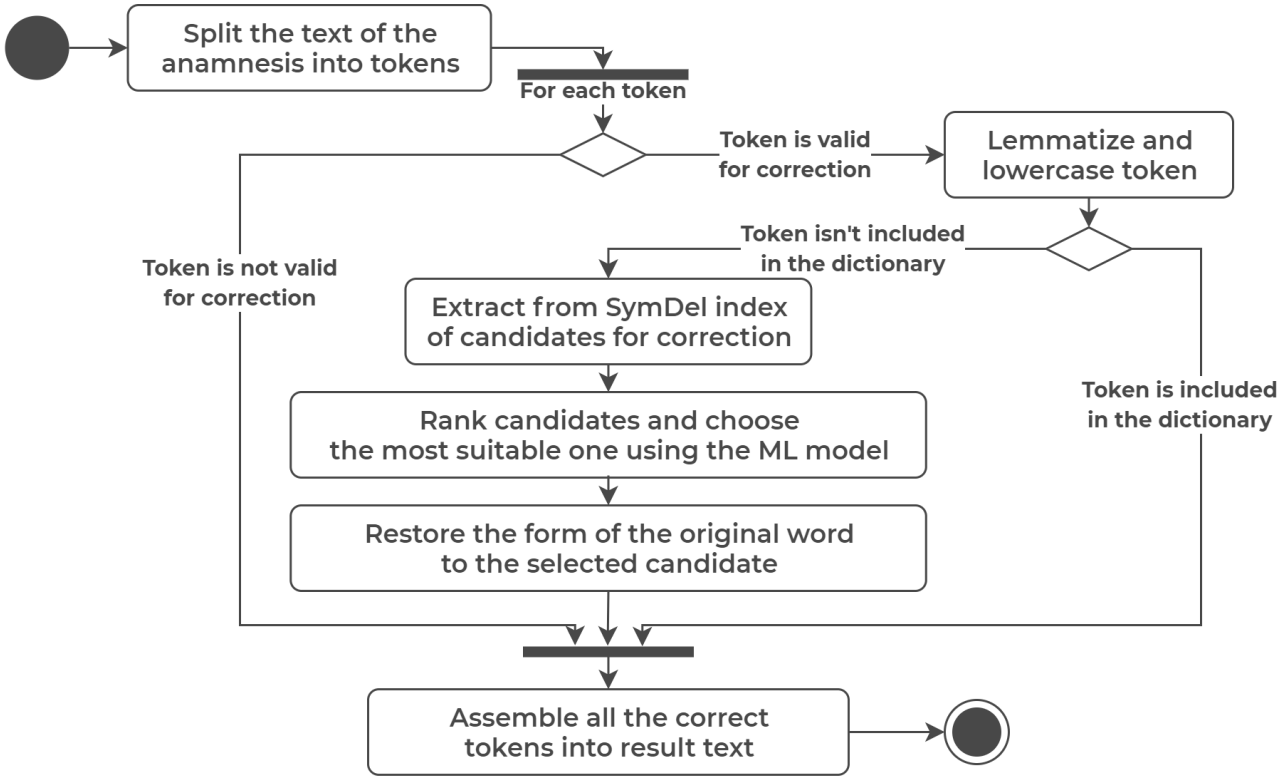
Type of mistake	Incorrect text	Correct text
Wrong characters	тубиркулез	туберкулез[1]
Missing characters	туб□ркулез	туберкулез
Extra characters	туберкпулез	туберкулез
Shuffled characters	тубрекулез	туберкулез
Missing word separator	острый|туберкулез	острый_туберкулез[2]
Extra word separator	туб_еркулез	туберкулез

[1] tuberculosis
[2] acute tuberculosis

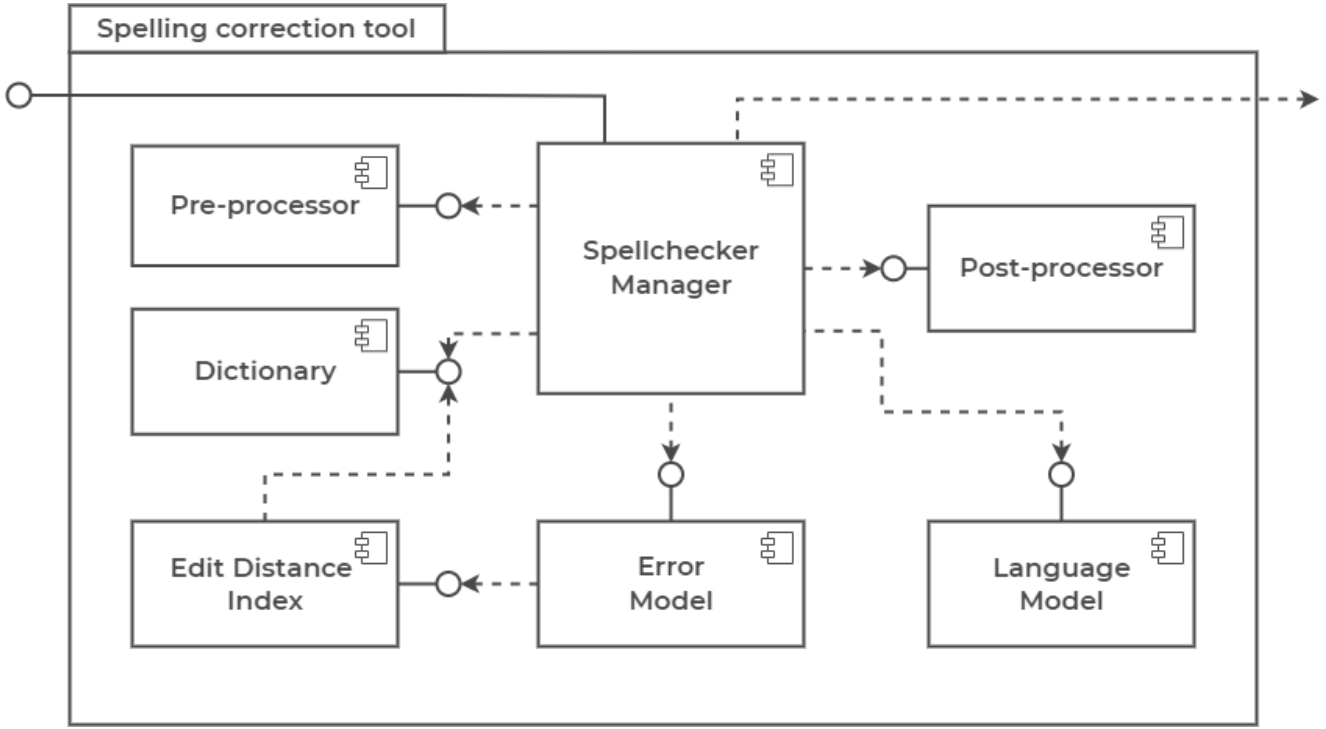
Existing tools

- There are several open source tools for Russian texts
  - Aspell
  - Hunspell
  - Enchant
  - LanguageTool
  - Symspell
  - Jumspell
- Not one is intended for medical texts
- Not one uses advanced language models

Spelling correction process



Tool architecture



Source anamneses datasets

- 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

Assembled anamneses dataset

- All datasets were pre-processed and assembled into final one
  - Tokenization and lemmatization
  - Stop words filtering

- Assembled anamneses dataset
  - Contains 30,737 medical records in Russian
  - Takes about 10.25 Mb

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
- All models are published on the Hugging Face repository

Adopted BERT models

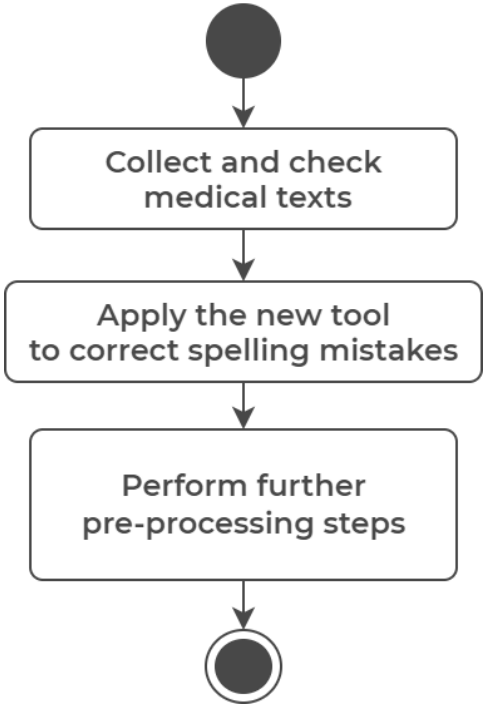
- RuBioBERT
  - Base model is RuBERT by SberDevices
- RuBioRoBERTa
  - Base model is RuRoBERTa by SberDevices

- Both models
  - Fine-tuned on Russian biomedical dataset
  - Produced from research[1] by Yalunin et al. from Sberbank AI Lab

[1]Yalunin, A., Nesterov, A., Umerenkov, D.: Rubioroberta: a pre-trained biomedical language model for russian language biomedical text mining (2022)

Method of tool use

- Only for correction of medical texts in Russian
- Preferably use for medical anamneses
- Use in the preprocessing pipelines
- Use before any other preprocessing steps
- Make sure everything is okay afterwards



Example of correction

Tool	Corrected Result
Original	тревожное расстройство (золофт) и атопичекий дерматит[1]
Aspell-python	тревожное расстройство золота и тапочкой дерматит
PyHunspell	тревожное расстройство золото аи топический дерматит
PyEnchant	тревожное расстройство золото аи атипический дерматит
LanguageTool-python	тревожное расстройство (золото) и утопический дерматит
PySpellChecker	тревожно расстройство (золофт) и атопичекий дерматит
SymspellPy	тревожное расстройство (золофт) и утопический гематит
Jumspell	тревожное расстройство (золофт) и атопичекий дерматит
Tool (MedDistilBERT)	**тревожное расстройство (золофт) и атопический дерматит**

[1] anxiety disorder (zoloft) and atopic dermatitis

Word tests internals

- Single test - error and lexical precision
  - 2700 test samples
- Context test - error and lexical precision
  - 2700 test samples
  - 10 words in each sample
  - One of ten words is incorrect, other words are correct
  - Same incorrect words as in single test
- Test on real anamnesis
  - 100 real anamnesis from Almazov Center dataset
  - Count the correct and unnecessary corrections

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

- **Correct fixes** - the number of correctly fixed errors
- **Unnecessary fixes** - the number of correct words corrected
- **Fixes ratio** - the ratio of the correct fixes metric to the unnecessary fixes

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	**0.991**	0.846	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)	**0.765**	**0.99**	**0.878**	45.5
Tool (GPU, MedDistilBERT)				153.8

Test on real anamnesis

Tool	Correct fixes	Unnecessary fixes	Fixes ratio
Aspell-python	20	171	0.105
PyHunspell	–	–	–
PyEnchant	–	–	–
LanguageTool-python	**21**	135	0.135
PySpellChecker	–	–	–
SymspellPy	–	–	–
Jumspell	–	–	–
Tool (MedRoBERTa)	19	**6**	**0.76**
Tool (MedDistilBERT)	18	9	0.667
Tool (MedBertTiny2)	17	11	0.607

Tool performance

- Tool is written in Python
- Damerau-Levenstein distance is used to reduce candidates for ranking
- SymDel index is used to speed up distance calculations
- Performance-critical operations are delegated to high-performance libraries
  - Edit distance calculation – **editdistpy** package
  - Model inference – **transformers** and **accelerate** packages
  - Tool can use a suitable GPU to accelerate the model inference

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

- New method for correcting medical texts in Russian is presented
- New tool implementing the proposed method is developed
- Tests have shown that the new tool is outperforming existing tools
  - In the task of correcting spelling errors in Russian medical texts
- New tool is open sourced and published as a pip package
- Fine-tuned models are available on HuggingFaces

Links

GitHub project
github.com/DmitryPogrebnoy/MedSpellChecker

Fine-tuned models
huggingface.co/DmitryPogrebnoy

Pip package
pypi.org/project/medspellchecker

Thank you for attention!

ITMO more than a UNIVERSITY