RuSimScore: unsupervised scoring function for Russian sentence simplification quality

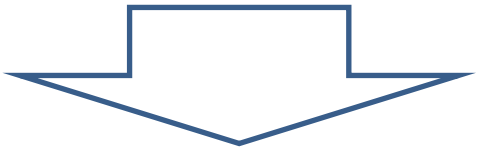
Mikhail Orzhenovskii

Simplification Task

Simplification: make sentence simple, preserve meaning

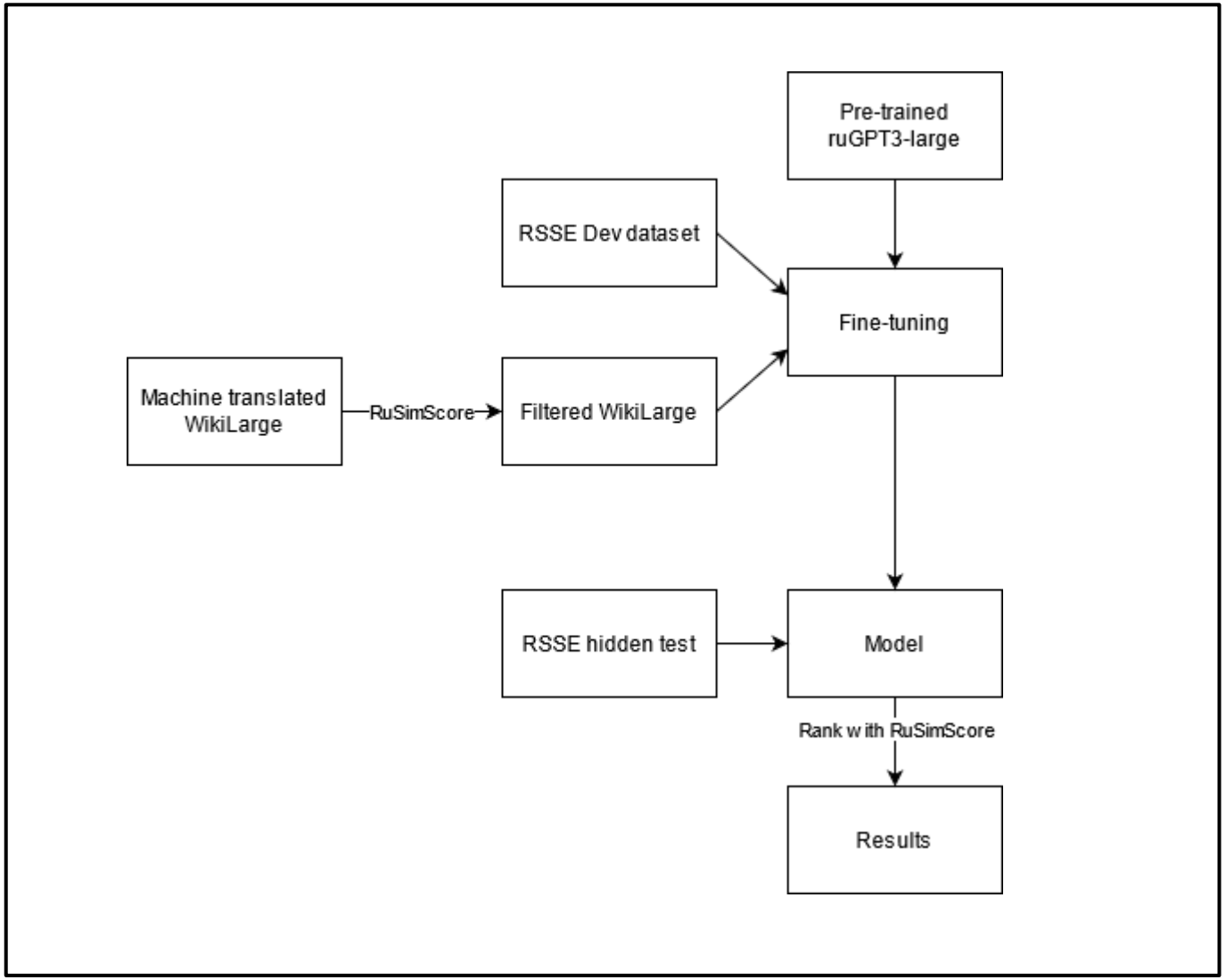
Example:

В качестве образца стала использоваться британская денежная система с делением на фунты стерлингов, шиллинги и пенсы.

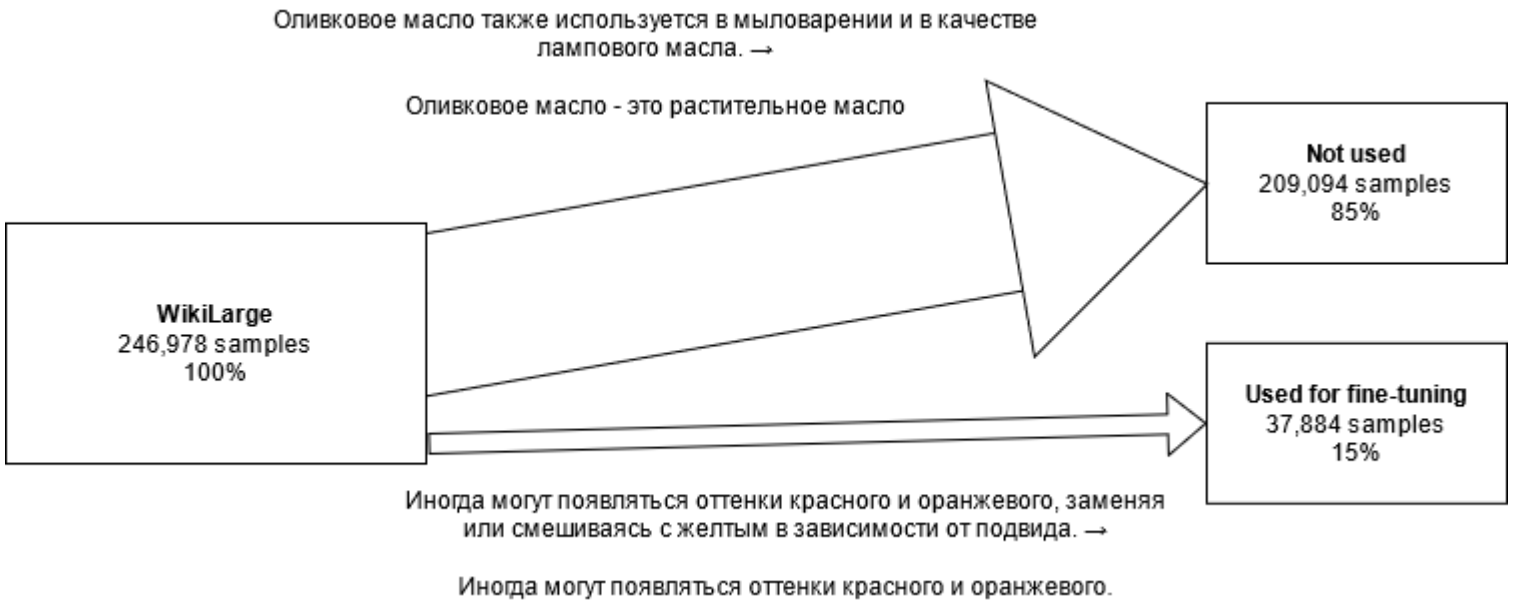


Британская денежная система была принята как образец.

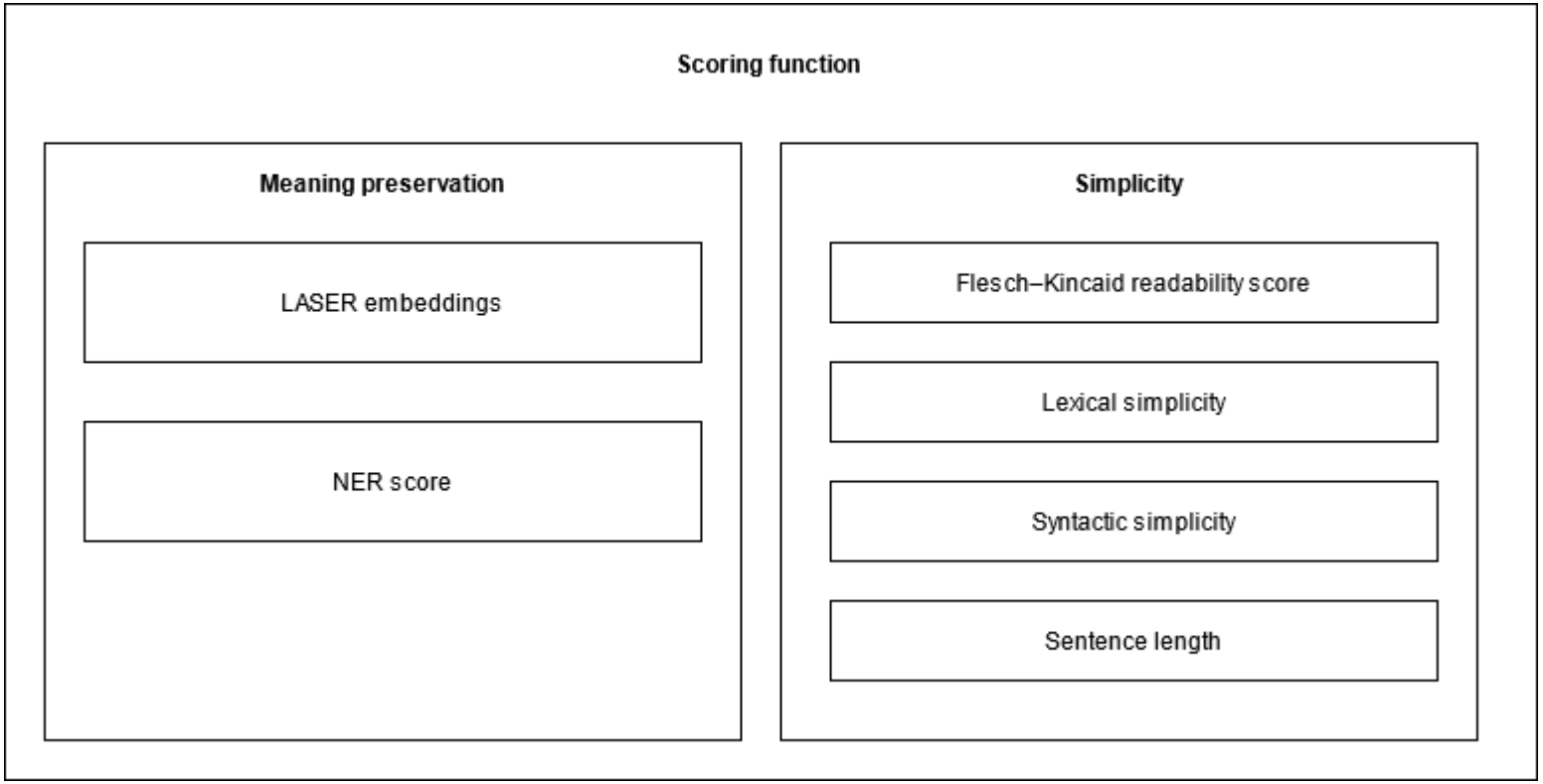
Solution


Pre-trained
ruGPT3-large
RSSE Dev dataset
Fine-tuning
Machine translated
WikiLarge
RuSimScore
Filtered WikiLarge
RSSE hidden test
Model
Rank with RuSimScore
Results

WikiLarge dataset filtering



Scoring function components



Meaning preservation

Similarity score (LASER embeddings cosine distance):

- Accurate
- Weakness: misspelled or replaced names

NER score:

- Number of named entities kept
- Handle possible modification of names:
  Оскар Александрович Энгберг→Энгберг
  Петру → Пётр

Simplicity

- Flesch–Kincaid readability score

$$RS = 0.75 + 0.25\frac{max(-100, min(100, 206.835 - 1.52WPS - 65.14\frac{SC}{WC}))}{100}$$

- Lexical simplicity – average and minimum log word frequency

$$LS = 1 + \alpha_{LS}\frac{\sum_{i=1}^{N} log(f_i)}{N} + \beta_{LS} min(log(f_i))$$

- Syntactic simplicity – dependency tree depth
- Sentence length (number of words)

$$LeS(c, s) = 0.5 \text{ if } WC(s) > WC(c)$$
$$LeS(c, s) = 1 - \frac{WC(s)}{2WC(c)} \text{ if } WC(s) > 6 \text{ and } WC(s) \leq WC(c)$$
$$LeS(c, s) = \frac{WC(s)}{6} \text{ if } WC(s) \leq 6 \text{ and } WC(s) \leq WC(c)$$
where $WC(x)$ is word count in sentence x.

Scoring function weights

$$RuSimScore(c,s) = LS^{\alpha}(c,s)DD^{\beta}(c,s)LeS^{\gamma}(c,s)RS^{\delta}(c,s)SimS^{\epsilon}(c,s)NS^{\zeta}(c,s)$$

Part	Weight
Lexical complexity score, α	1.50
Dependency tree depth score, β	0.21
Length score, γ	1.24
Reading ease score, δ	0.33
Similarity score, ε	1.58
Named entity preservation score, ζ	0.72

Model

Parameters

Base model	rugpt3-large
Fine-tuning epochs	2
Temperature	0.9
top_p	0.95
top_k	Not used
Candidate simplifications per input sentence	100

Filtering results

Repeated words
Толстой сказал сказал...

Starting with determiner
Тот человек был известен...

"Это" in the beginning
Это ухудшило положение...

Starting with pronoun
Они организовали еще несколько...

Adding non-cyrillic words
Август сохранилender-marginalis

Example of model outputs

- Затем выяснилось, что мелатонин образуется и в других отделах желудочно-кишечного тракта, а также во многих других органах.

  →

  Мелатонин образуется и во многих других органах.

- В структуре грузовых перевозок основное место занимают навалочные строительные материалы (песок, щебень, камень и др.).

  →

  Основные грузы - это щебень, песок и камень.

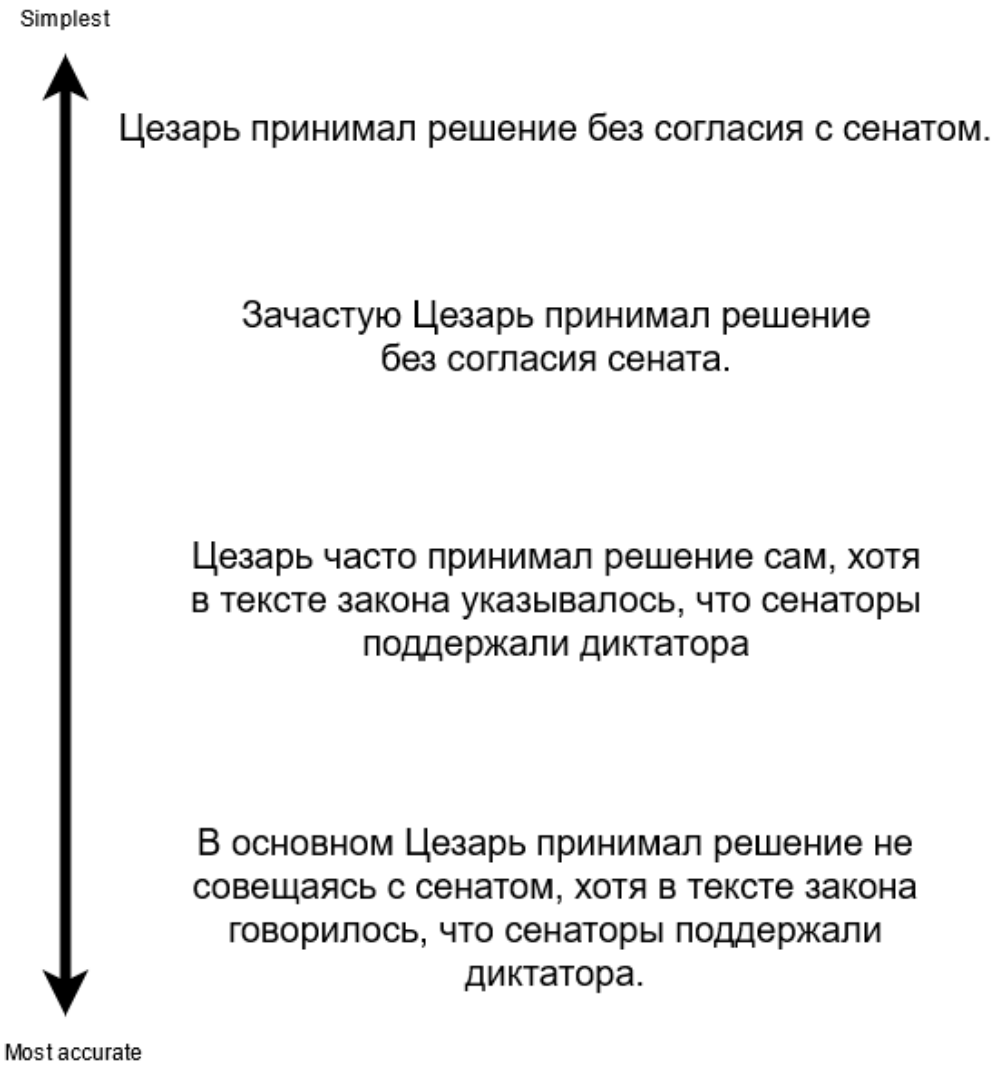
- В Саудовской Аравии официально вступил в силу королевский указ о снятии запрета на вождение автомобиля женщинам.

  →

  В Саудовской Аравии официально стало можно водить машину женщинам.

Controlling simplification

Зачастую Цезарь принимал решение не совещаясь с сенатом, хотя при публичном оглашении текста закона указывалось, будто сенаторы поддержали диктатора.



Results

Model	Hidden test SARI
ruGPT3 on filtered WikiLarge + RuSimScore	**39.28**
ruGPT3 on filtered WikiLarge	38.68
Official benchmark (mBART)	30.15
Iterative deletion with RuSimScore	32.40
First half of source text	30.33
Source text unchanged	11.04

Count	Hidden test SARI
100	39.28
30	**39.39**
10	39.16
1	38.68

Model	Hidden test SARI
ruGPT3-small	38.89
ruGPT3-medium	**39.34**
ruGPT3-large	39.28

Analysis and errors

	Original weights	Better accuracy
Well formed sentences	90	93
Not well formed sentences	10	7
Factually consistent	71	81
Issues with factual consistency, including:	29	19
Omitting significant details	15	7
Wrong entity reference	8	6
Incorrect replacement word used	6	6

Original sentence	Влияние организованной преступности, наркосиндикатов, мафиозных структур, в особенности — в южных регионах — Кампании, Калабрии и на Сицилии.
Omitting significant details	На юге Кампания, Калабрия и Сицилия.
Original sentence	Лейси активно начал поединок, прессингуя и зажимая Джонса у канатов ринга, но Рой умело уходил от ударов и большинство блокировал.
Wrong entity reference	Рой атаковал Джонса и пытался его прессинговать, но Рой успешно отбивался.
Original sentence	Летописцы указывали, что кони уже не могли не ступать по трупам, так как не было чистого места.
Incorrect replacement word used	Летописцы писали, что кони уже не могли не пачкать трупы.

Code and model

https://github.com/orzhan/rusimscore

https://huggingface.co/orzhan/rugpt3-simplify-large