Spelling and punctuation correction for multiple languages

Agenda

Task formulation

Naive: Normalize all the words to its vocabulary equivalents

Товарооборот увеличелся до 15 трлн тонн

Tovarooborot uvelichels'a do 15 trln ton

⟹

Товарооборот увеличился до 15 трлн тонн

Tovarooborot uveflichils'a do 15 trln tonn

Погода севодня солничная

Pogoda sevodn'a solnichnaya

⟹

Погода сегодня солнечная

Pogoda sevodn'a solnichnaya

OMГ, какоооой фильм, сто проц его смотртю))

OMG, kakooooi fil'm, sto proc yego smotr'u))

⟹

Какой фильм, сто процентов его смотрю

Kakoi fil'm, sto procentov yego smotr'u

SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

SBER DEVICES

Task formulation

ОМГ, какоооой фильм, сто проц его смотртю))

OMG, kakooooi fil'm, sto proc yego smotr'u))

Какой фильм, сто процентов его смотрю

Kakoi fil'm, sto procentov yego smotr'u

Do we really need to normalize ALL the words?

SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

Task formulation

ОМГ, какооооой фильм, сто проц его смотртю))

OMG, kakooooi fil'm, sto proc yego smotr'u))

Какой фильм, сто процентов его смотрю

Kakoi fil'm, sto procentov yego smotr'u

Do we really need to normalize ALL the words?

- *Context matters;*
- *Keep emotional palette from source text to corrected;*
- *Distinguish intentional violation of spelling;*

Datasets

Motivation

Teach the model to distinguish intentional and unintentional spelling violations

Datasets

Motivation

Teach the model to distinguish intentional and unintentional spelling violations

Issue

Each spelling violation must be examined from language perspective
Can not be done in **automatic fashion**

Datasets

Motivation

Teach the model to distinguish intentional and unintentional spelling violations

Issue

Each spelling violation must be examined from language perspective
Can not be done in **automatic fashion**

Solution

— Gather text sources considering variety and error-proneness;

— Filter out correct sentences;

— Write down corresponding corrections for remaining sentences;

— Check the appropriateness of corrections;

SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

SBER DEVICES

Datasets

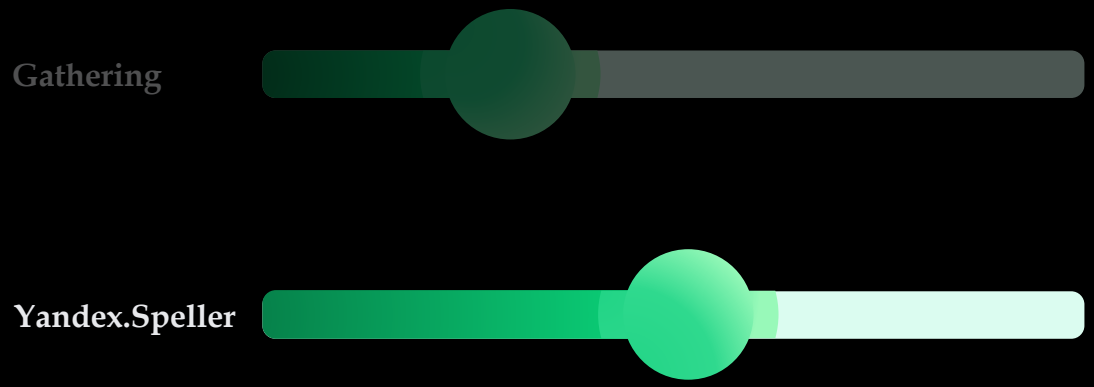
Gathering

Datasets creation

Gathering

We gathered nearly 164k of raw text samples in total.

Datasets



Datasets creation

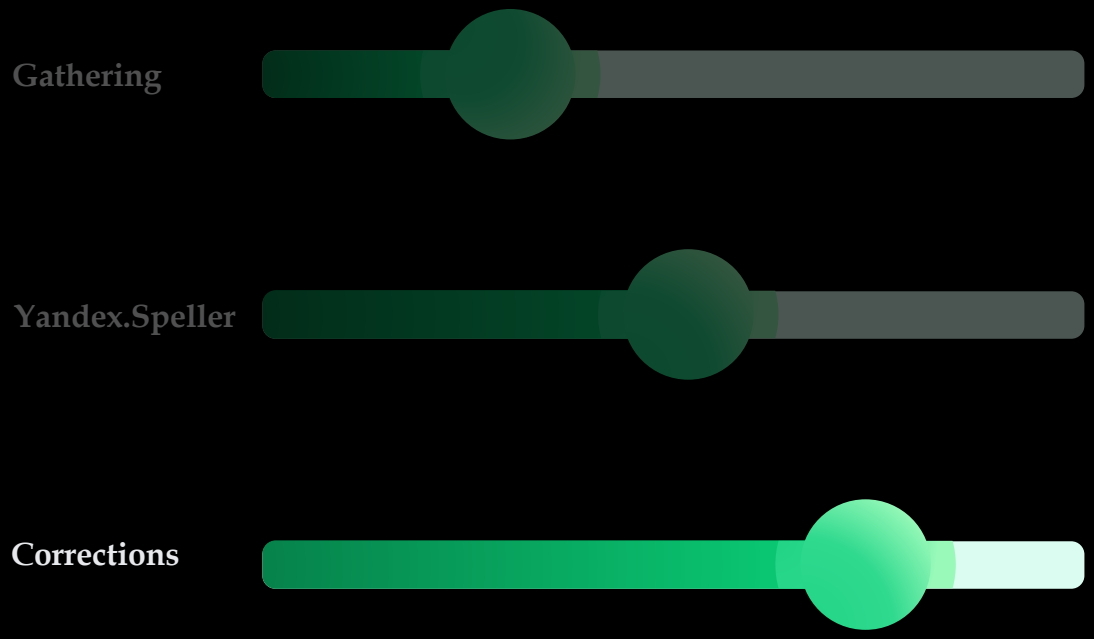
Gathering

We gathered nearly 164k of raw text samples in total.

Yandex.Speller

In order to reduce the amount of manual labor in the next stage, we filtered out correct texts with Yandex.Speller.

Datasets



Datasets creation

Gathering

We gathered nearly 164k of raw text samples in total.

Yandex.Speller

In order to reduce the amount of manual labor in the next stage, we filtered out correct texts with Yandex.Speller.

Correction

We required all potential annotators to have language exam passed. In addition, we came up with test assignments that caused 85 and 86 percent annotators withdrew on first and second stage accordingly.

Datasets

Datasets creation

Gathering

We gathered nearly 164k of raw text samples in total.

Yandex.Speller

In order to reduce the amount of manual labor in the next stage, we filtered out correct texts with Yandex.Speller.

Correction

We required all potential annotators to have language exam passed. In addition, we came up with test assignments that caused 85 and 86 percent annotators withdrew on first and second stage accordingly.

Validation

On each stage of manual labeling we asked annotators to decide the intent of an error considering context and their language sense.

SBER DEVICES

SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

Spelling corruption

— Inverted spelling correction;

— Naïve: insert random characters and gaps;

— Naïve does not really do a great job – one needs something smarter;

SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

Statistic-based spelling corruption

Emulates human's behavior when making an error on a correct text.

Statistics gathering

— Needs parallel dataset with natural errors;

— Accounts for six types of errors: insertion, deletion and substitution of character, transposition of two adjacent characters, insertion and deletion of gaps;

— Builds Damerau-Levenshtein matrix between prefixes of original sentence and its correction, traverse it back from the bottom-right entry and decide type and position of particular error considering its surrounding entries;

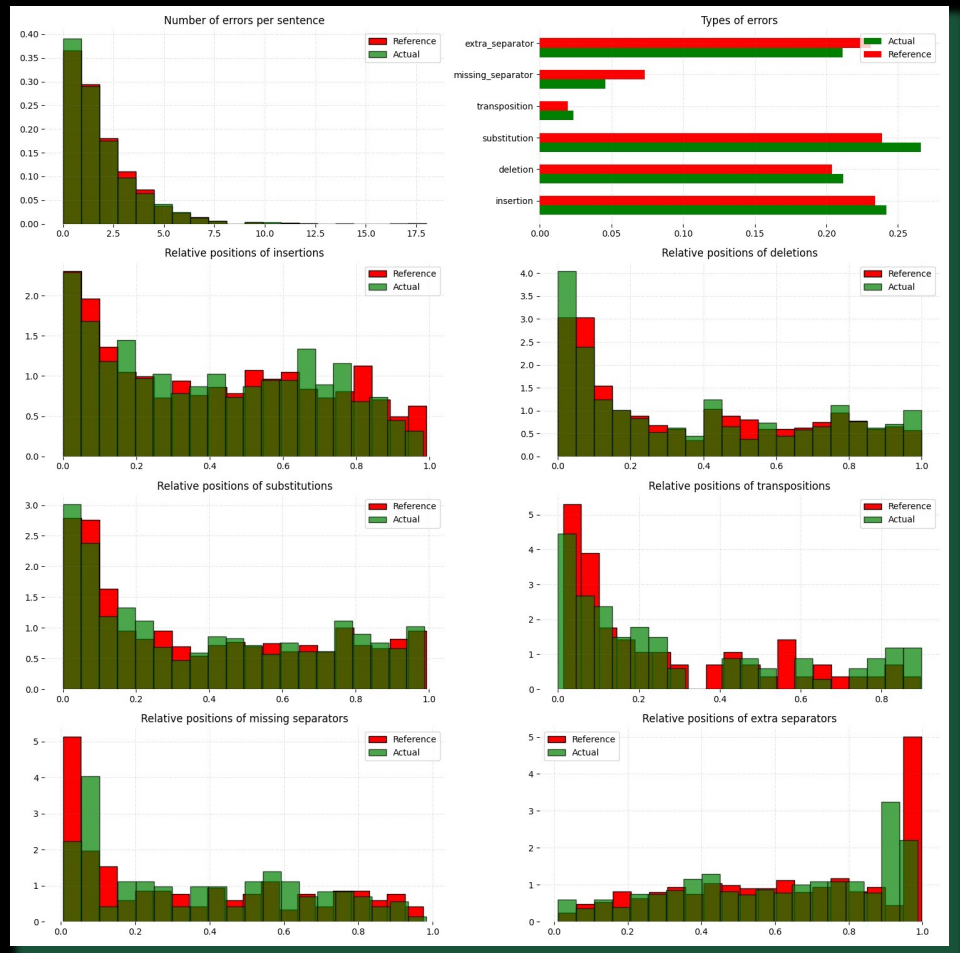
— Gathers statistics along the corpus and normalizes it to valid discrete probability distributions;

Emulation

— Samples number of errors per sentence;

— For each error samples its type and interval;

— Chooses appropriate position within the interval and inserts the error;



SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

SBER DEVICES

Augmentex

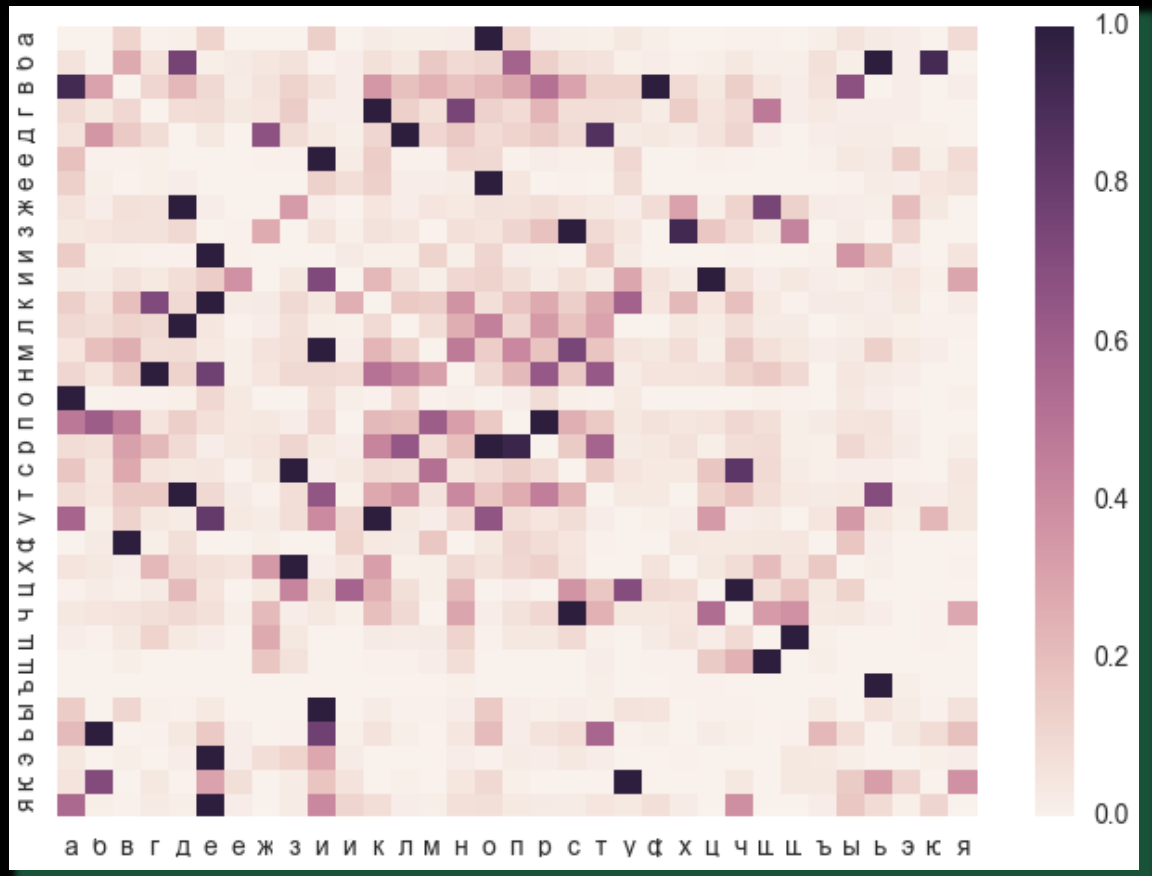
Common errors statistics from KartaSlov project and heuristics-empowered methods for character and word levels.

Character level

- Shift: changes the case of starting character;
- Orfo: character is replaced with its erroneous counterpart according to corresponding statistics;
- Typo: character is replaced with another one as it has been mistyped;
- Delete: deletion of random character;
- Multiply: multiplication of random character;
- Swap: transposition of two adjacent characters;
- Insert: insertion of random character;

Word level

- Replace: replaces random word with its erroneous counterpart;
- Delete: deletion of random word;
- Swap: transposition of two adjacent words;
- Stopword: random insertion of a word from stop-list;
- Reverse: switches the case of a starting characters;



SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

SBER DEVICES

Pre-train

Specifically designed pre-train procedure enables spelling correction models to reach a raise up to 20 F1 points over their counterparts with plain fine-tune

Pre-train procedure for spelling correction task

— Obtain extensive corpus of initially correct texts: Wikipedia and video transcripts (news for the English language);

— Introduce noise into corpus with SBSC initialized from RUSpellRU or Augmentex;

— seq2seq training: noised sentence as input, its correct counterpart as output;

— Range of auxiliary pre-train tasks;

2

auxiliary tasks:
— Multi-label NER [encoder];
— Embedding similarity [decoder]

7

Gb is the size of pre-train corpus. We filtered out texts with non-alphabetic symbols and those shorter than 40 symbols;

5

types of architecture:
— FredT5-small, -large and -XL;
— M2M100-1.2B and -418M;
— T5-large and mT5-large (English language);

>20

percent of characters in original sentences have been corrupted when preparing pre-train corpus;

SBER DEVICES

SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

Pre-train [Multi-label NER]

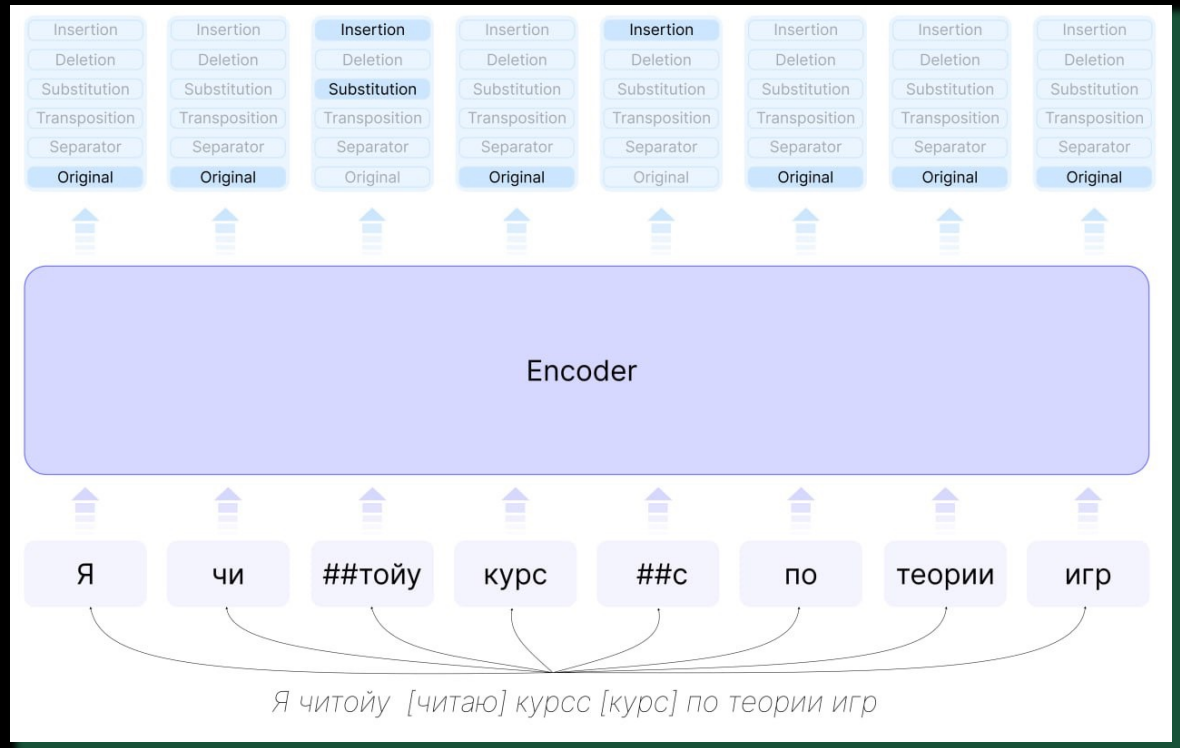
We force the encoder to predict the type of an error for each token in input sequence and optimize the whole architecture for seq2seq and classification objectives.

Mechanics

- Corrupt spelling and / or punctuation with SBSC or Augmentex;
- Remember position and type;
- Label the token accordingly: mark it with corresponding error type or ”clean” if the token does not embody any error;
- One token may probably encompass multiple errors, so the token is annotated with rather multiple lables;

Objective

- Encoder predicts the types of errors that the particular token contains;
- Predicts for each token in a sentence;
- Multi-label NER loss;



SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

Pre-train [CLIP]

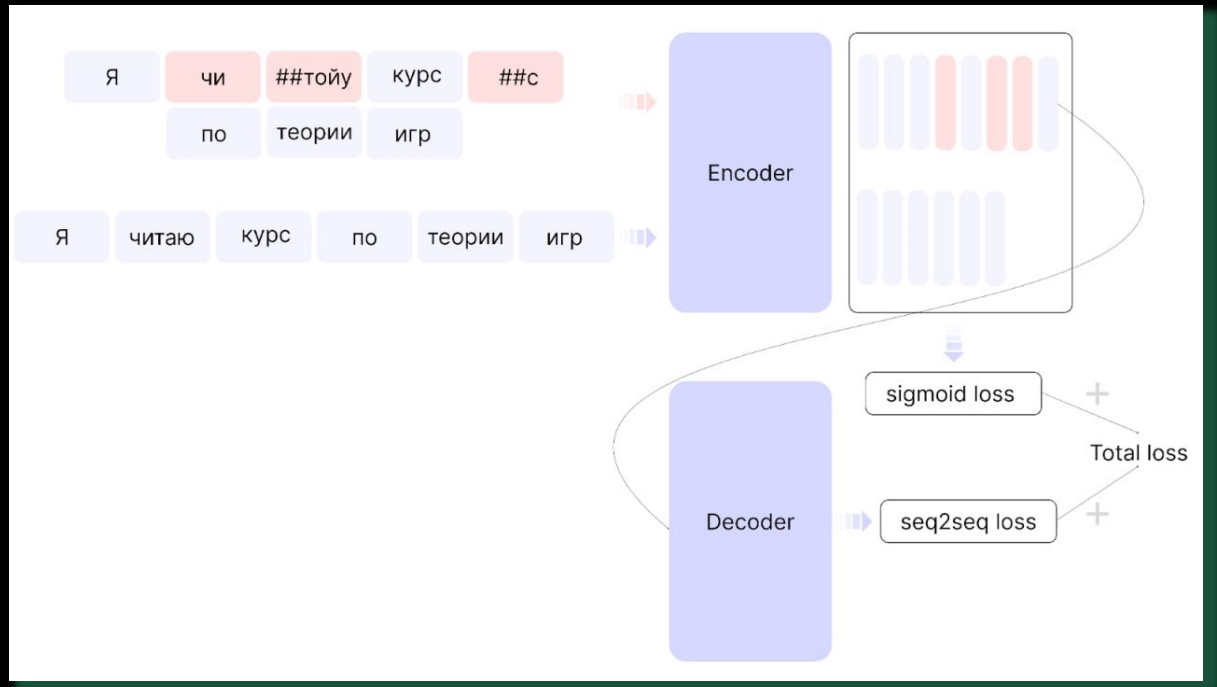
We force the encoder to produce embeddings, which correspond to source and correct sentences, that are close to each other.

Mechanics

— Pass the source and correct to encoder and collect the output;
— Calculate contrastive loss with pairing source and correct as positive examples and any other pair as negative one;
— Pass the output for the source to the decoder;

Objective

— Calculate contrastive loss;
— Calculate seq2seq loss;
— Jointly optimize both encoder and decoder for the sum of objectives;



SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

Fine-tune for the Russian language

We evaluate our models in zero-shot setting and after fine-tuning on target datasets. We also draw a comparison with popular open sourced spellcheckers and proprietary solutions.

Fine-tune

— We fine-tuned our models on a combination of both datasets RUSpellRU and MultidomainGold (MDG);

— Corresponding auxiliary task used while pre-train procedure is mentioned in parenthesis;

— gpt-3.5-turbo-0125 and gpt-4-0125-preview are stable versions of gpt-3.5 and gpt-4 models family by the time of the experiments;

Zero-shot

— We evaluated zero-shot performance of all pre-trained checkpoints (*Pre-train (PT.)*);

Model	RUSpellRU			MultidomainGold			MedSpellChecker			GitHubTypoCorpusRu		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Open source												
Yandex.Speller	83.0	59.8	69.5	52.9	51.4	52.2	**80.6**	47.8	60.0	67.7	37.5	48.3
JamSpell	42.1	32.8	36.9	25.7	30.6	28.0	24.6	29.7	26.9	49.5	29.9	37.3
Hunspell	31.3	34.9	33.0	16.2	40.1	23.0	10.3	40.2	16.4	28.5	30.7	29.6
Proprietary												
gpt-3.5-turbo-0125	39.6	62.3	48.47	17.8	56.1	27.0	15.1	53.6	23.5	23.7	43.9	30.8
gpt-4-0125-preview	69.5	81.0	74.8	31.1	78.1	44.5	48.9	**88.7**	63.1	34.7	**60.5**	44.1
sage-m2m100-1.2B												
Pre-train (PT.)	59.4	43.3	50.1	56.4	44.8	49.9	63.7	57.8	60.6	45.7	41.4	43.5
RUSpellRU+MDG (+PT.)	88.8	71.5	79.2	63.8	61.1	62.4	78.8	71.4	**74.9**	47.1	42.9	44.9
sage-fredt5-large												
Pre-train (NER)	56.7	70.5	62.9	36.5	60.9	45.7	31.8	68.3	43.3	43.3	52.8	47.6
Pre-train (CLIP)	65.6	73.8	69.5	43.1	**80.6**	56.2	25.4	63.7	36.3	39.7	54.4	45.9
RUSpellRU+MDG (+NER)	91.1	77.2	83.6	68.9	67.6	68.3	73.2	73.2	73.2	76.0	48.1	58.9
RUSpellRU+MDG (+CLIP)	90.9	75.0	82.2	70.5	65.3	67.8	74.9	66.5	70.5	79.0	45.7	57.9
sage-fredt5-1.7B												
Pre-train (NER)	69.3	69.8	69.5	40.3	65.0	49.8	34.8	66.0	45.6	47.6	57.3	52.0
Pre-train (CLIP)	72.1	69.1	70.6	45.2	60.7	51.8	40.1	67.3	50.2	56.1	56.0	56.0
RUSpellRU+MDG (+NER)	**93.5**	**82.4**	**87.6**	**70.9**	68.8	**69.9**	73.4	76.2	**74.9**	76.1	51.2	61.2
RUSpellRU+MDG (+CLIP)	92.1	81.1	86.3	70.6	68.4	69.4	72.4	66.5	69.4	**80.5**	51.5	**62.8**

SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

Distillation

Distilled open-sourced FredT5-1.7B version.

Student

— Pre-trained separate FredT5-95m;

— The same corpus as in the original paper;

— The same pre-train tasks and methodology as in the original paper;

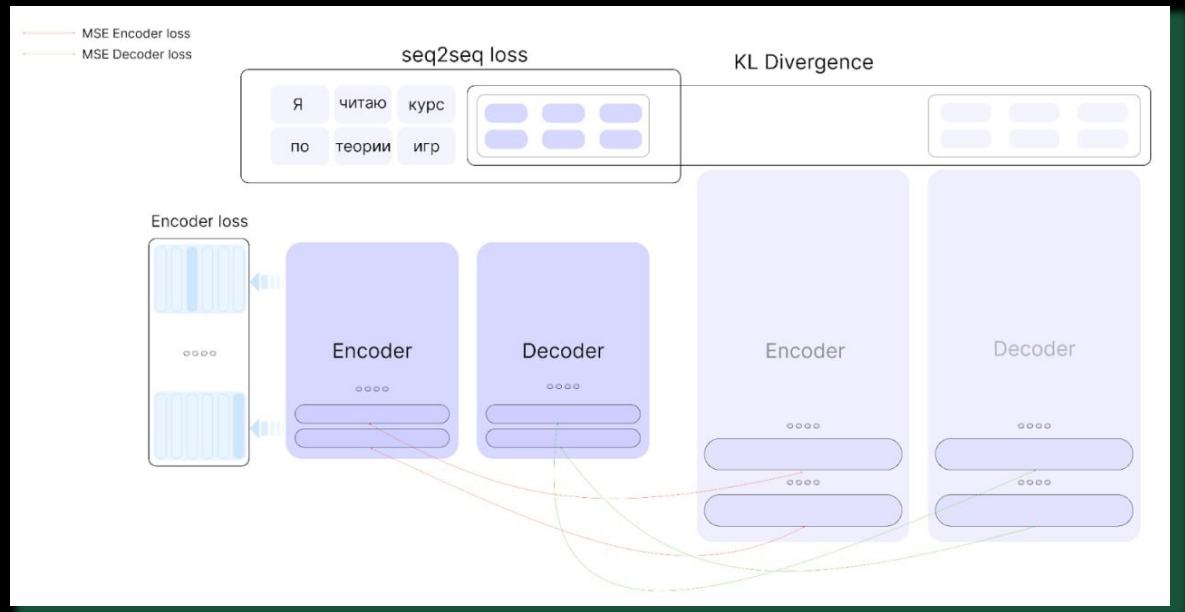
Mechanics

— Three types of objectives;

— Auxiliary task specific loss, KL Divergence, seq2seq loss;

— Teacher model is frozen;

— Distillation is performed on pre-train corpus and fine-tune datasets in two consequent steps;



SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

SBER DEVICES

OpenAI models and open sourced tools

We also evaluated our solutions alongside open spellcheckers and OpenAI models for punctuations and case related errors as well.

Those that do not have an ability to correct punctuation errors or case related errors are mentioned only with spelling.

We use the same metric logic to estimate performance on punctuation and case related errors.

Model	RUSpellRU			MultidomainGold			MedSpellChecker			GitHubTypoCorpusRu		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Yandex.Speller												
Spelling	83.0	59.8	69.5	52.9	51.4	52.2	80.6	47.8	60.0	67.7	37.5	48.3
sage-m2m100-1.2B												
Spelling	88.8	71.5	79.2	63.8	61.1	62.4	78.8	71.4	74.9	47.1	42.9	44.9
sage-fredt5-1.7B												
Spelling	90.3	86.3	88.2	81.6	77.7	79.6	71.3	73.5	72.4	70.8	56.3	62.7
Punctuation	90.3	86.6	88.4	70.2	67.5	68.8	75.1	69.2	72.0	48.9	35.8	41.4
Case	95.2	95.9	95.6	80.5	80.5	80.5	80.9	72.8	76.6	32.9	45.3	38.1
sage-fredt5-large												
Spelling	88.4	80.9	84.5	80.3	75.1	77.6	72.5	72.2	72.3	67.5	53.2	59.5
Punctuation	88.2	85.3	86.8	69.0	66.5	67.7	74.6	66.4	70.3	48.5	38.0	42.6
Case	95.5	94.0	94.7	78.6	80.0	79.3	79.3	85.1	82.1	37.3	50.0	42.7
sage-fredt5-95m												
Spelling	83.5	74.8	78.9	77.2	69.9	73.4	65.1	64.8	64.9	57.8	48.5	52.7
Punctuation	86.8	80.6	83.6	66.8	63.4	65.0	78.6	63.1	70.0	45.2	39.5	42.1
Case	94.4	92.5	93.5	76.8	79.1	77.9	63.5	74.7	68.7	29.9	46.2	36.3
gpt-3.5-turbo-0125												
Spelling	33.6	58.5	42.7	18.8	48.1	27.1	14.7	45.9	22.3	23.7	38.7	29.4
Punctuation	85.9	64.6	73.7	42.0	31.8	36.2	69.9	52.3	59.8	37.6	23.3	28.7
Case	84.9	73.9	79.0	47.1	51.3	49.1	26.4	41.8	32.3	19.6	35.9	25.3
gpt-4-0125-preview												
Spelling	54.9	76.7	64.0	25.4	68.0	37.0	37.8	72.3	49.6	27.0	52.8	35.7
Punctuation	84.0	82.3	83.2	57.8	54.3	56.0	81.4	64.3	71.9	45.9	32.6	38.2
Case	91.5	90.2	90.9	54.0	67.5	60.0	73.0	62.1	67.1	25.7	36.8	30.2

SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

Results for the English language

We employed mT5-large to simultaneously optimize for spelling correction for both Russian and English languages.

- We evaluate our multi-language solution on test suits for both languages;
- We use BEA60K and JFLEG datasets for the English language zero-shot fashion;
- We also draw a comparison with open-sourced tools, proprietary solutions and our mono-language solution;
- Used gpt-3.5-turbo-0125 and gpt-4-0125-preview because they seem to produce better results than newer versions by the time of the experiments;

Model	BEA60K			JFLEG		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Proprietary						
gpt-3.5-turbo-0301	66.9	84.1	74.5	77.8	88.6	82.9
gpt-4-0314	68.6	85.2	76.0	77.9	88.3	82.8
Monolingual SAGE						
sage-t5-large	66.5	83.1	73.9	83.4	84.3	83.8
Multilingual SAGE						
sage-mt5-large	64.7	83.8	73.0	74.9	88.4	81.1

Model	RUSpellRU			MultidomainGold			MedSpellChecker			GitHubTypoCorpusRu		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Open source												
Yandex.Speller	83.0	59.8	69.5	52.9	51.4	52.2	80.6	47.8	60.0	67.7	37.5	48.3
JamSpell	42.1	32.8	36.9	25.7	30.6	28.0	24.6	29.7	26.9	49.5	29.9	37.3
Hunspell	31.3	34.9	33.0	16.2	40.1	23.0	10.3	40.2	16.4	28.5	30.7	29.6
Proprietary												
gpt-3.5-turbo-0125	39.6	62.3	48.47	17.8	56.1	27.0	15.1	53.6	23.5	23.7	43.9	30.8
gpt-4-0125-preview	69.5	81.0	74.8	31.1	78.1	44.5	48.9	88.7	63.1	34.7	60.5	44.1
Monolingual SAGE												
sage-m2m100-1.2B	88.8	71.5	79.2	63.8	61.1	62.4	78.8	71.4	74.9	47.1	42.9	44.9
sage-fredt5-large	91.1	77.2	83.6	68.9	67.6	68.3	73.2	73.2	73.2	76.0	48.1	58.9
sage-fredt5-1.7B	93.5	82.4	87.6	70.9	68.8	69.9	73.4	76.2	74.9	76.1	51.2	61.2
Multilingual SAGE												
sage-mt5-large	88.4	71.6	79.1	65.3	62.7	63.9	77.7	77.5	77.6	69.5	46.0	55.3

SAGE

15

F1-points advance over Open AI models and open sourced tools is demonstrated by our solution

8

open sourced checkpoints pre-trained on spelling correction for Russian and English languages from 95M to 1.2B

4

open-sourced parallel datasets for spelling correction task. Manually annotated, around ten different text domains covered.

2

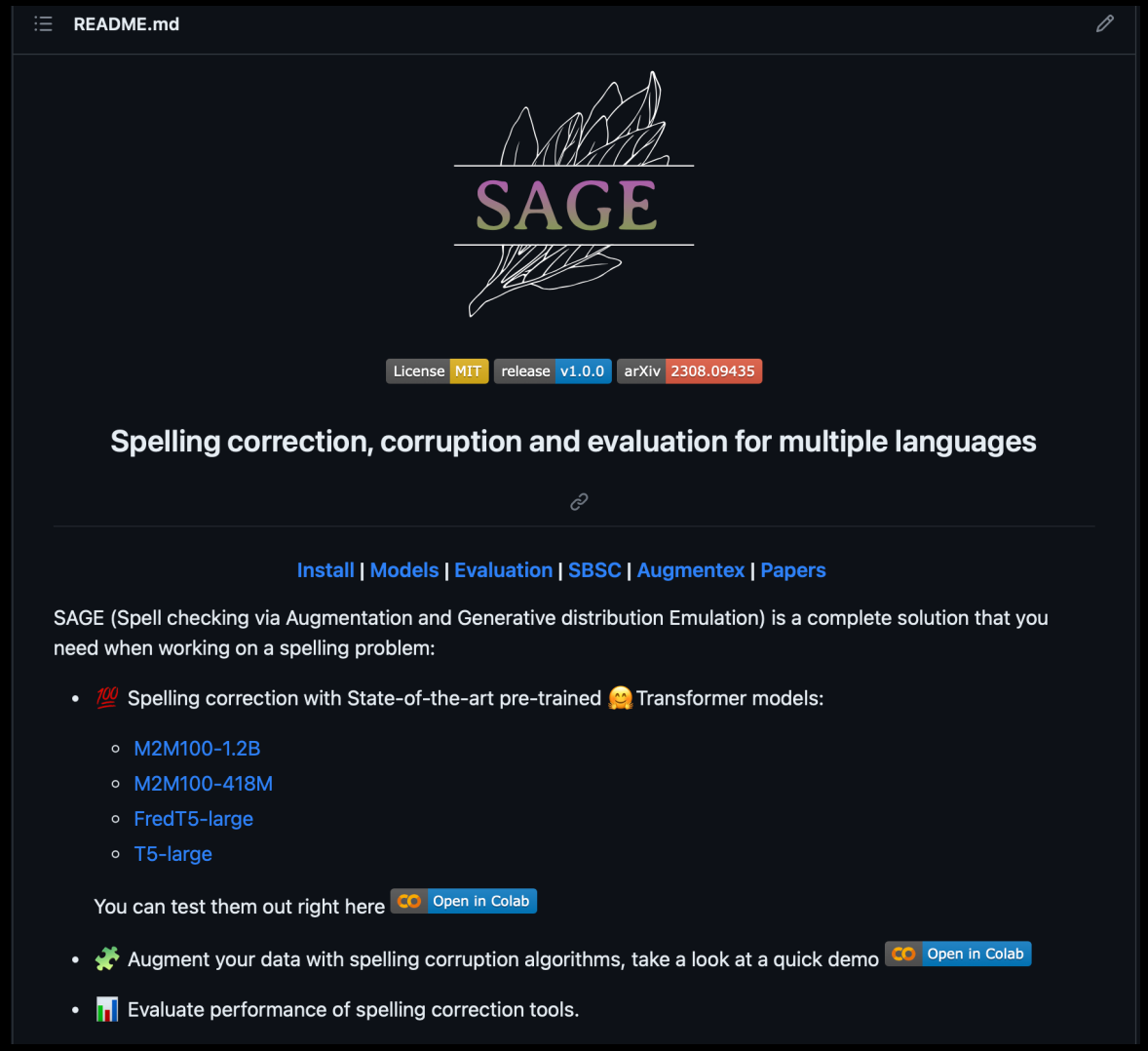
spelling corruption algorithms for text augmentation: Statistic-based spelling corruption and Augmentex.

SAGE: spelling and punctuation correction for multiple languages, Nikita Martynov, IML 2024

SBER DEVICES

SAGE

We would really like to see you on the homepage of our project. Come over and visit SAGE library: learn about our models, datasets and noising strategies and get in touch with our latest updates.




SBER DEVICES