

The background is a dark blue-grey color. It is decorated with various geometric elements: orange circles of different sizes, some with dotted patterns; white circles and hexagons; orange hexagons; and white dotted patterns arranged in circles, hexagons, and rectangular grids. Thin white and orange lines also crisscross the background.

Project CALL

Group CALL 1

Project CALL

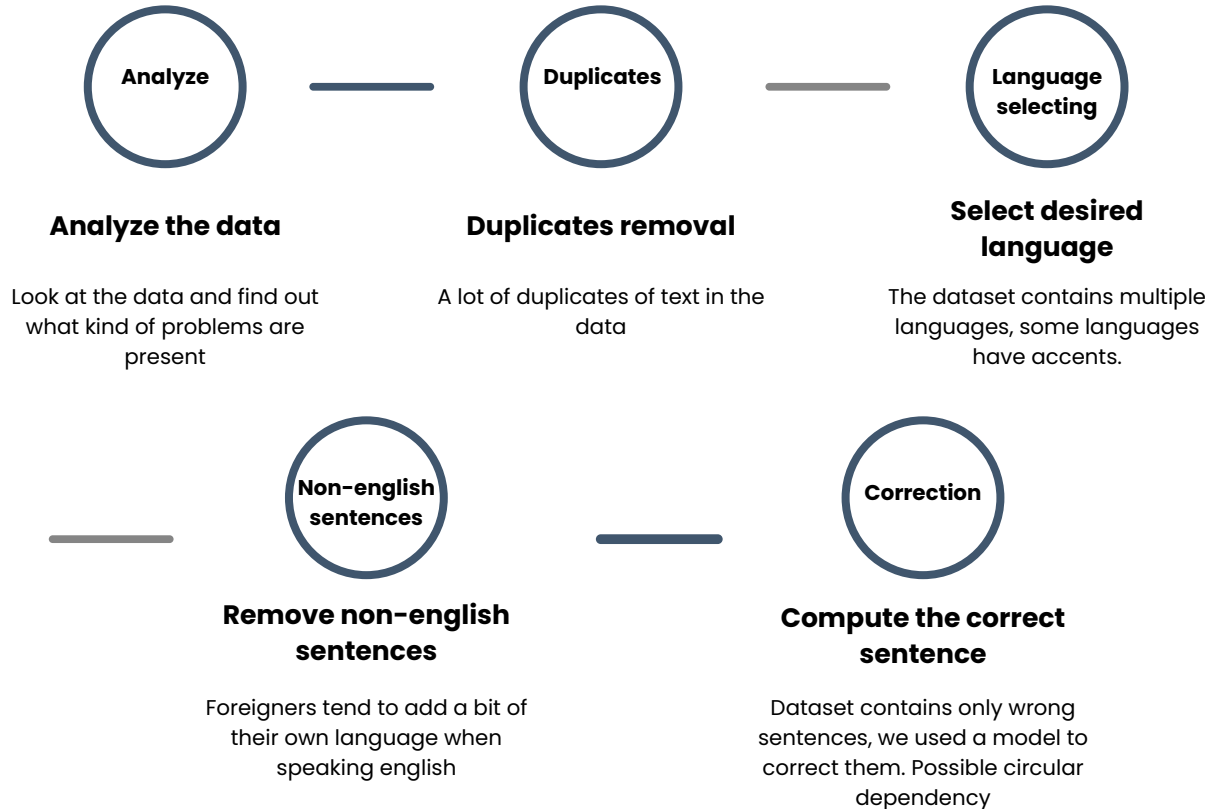
Main task: Grammatical Error Correction (GEC) for non-native language learners

Sub-task: Create an GEC for latin languages

Sub-task: Test our methodology to create a language specific GEC



Data Quality



Error Tagging

- ERRANT (grammatical ERRor ANnotation Toolkit)
- Error categories: Unnecessary, Missing and Replacement

Corrupt: This are gramamtical sentence.

Correct: This is a grammatical sentence.

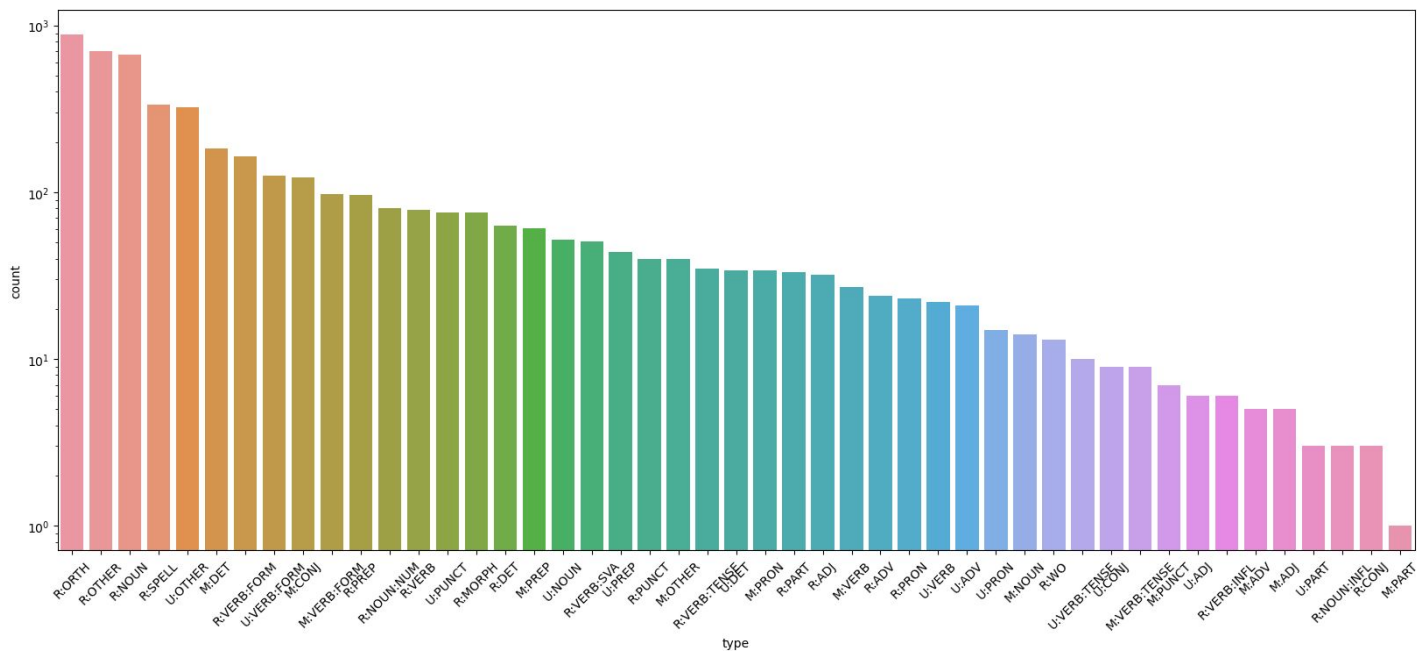
Predicted tags: R:VERB:SVA, M:DET and R:SPELL

Code	Meaning
ADJ	Adjective
ADJ:FORM	Adjective Form
ADV	Adverb
CONJ	Conjunction
CONTR	Contraction
DET	Determiner
MORPH	Morphology
NOUN	Noun
NOUN:INFL	Noun Inflection
NOUN:NUM	Noun Number
NOUN:POSS	Noun Possessive
ORTH	Orthography
OTHER	Other
PART	Particle
PREP	Preposition
PRON	Pronoun
PUNCT	Punctuation
SPELL	Spelling
UNK	Unknown
VERB	Verb
VERB:FORM	Verb Form
VERB:INFL	Verb Inflection
VERB:SVA	Subject-Verb Agreement
VERB:TENSE	Verb Tense
WO	Word Order

Training data: Part 1

Step 1: Get corrupt and correct sentence pairs for language subset

Step 2: Apply error tagging and retrieve the error frequencies

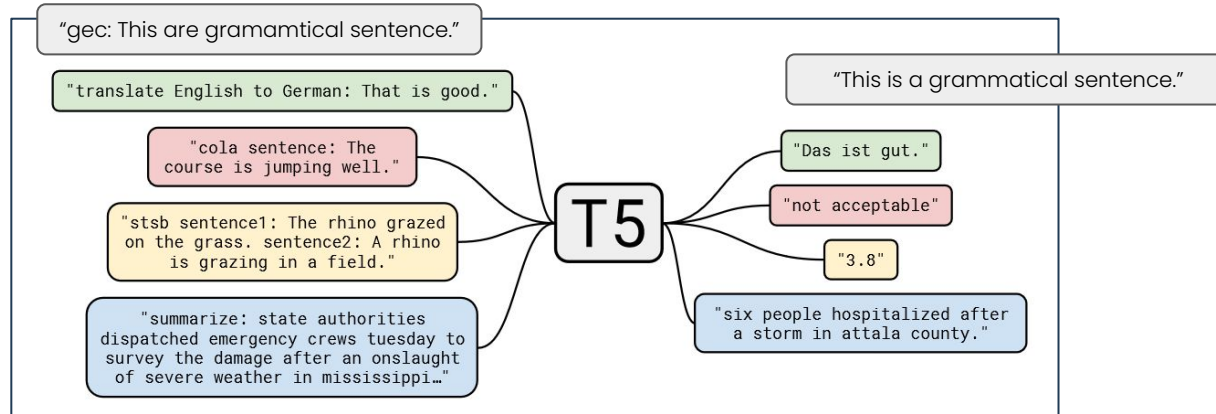


Training data: Part 2

- Take sample from C4 200M dataset
 - Needed 120000 to have the exact same frequencies
- Calculate error tags
- Create a training set with the frequencies for the specific language set
 - Sentences have multiple tags
 - Run once to see what frequencies are missing most, select those first

T5 model

- T5 = Text-to-Text Transfer Transformer
- Encoder-Decoder Model
- Prefix specifies task



Text-to-text framework

Evaluation

Original: 4. I get nourish myself so help me to help others.

Correct Old: 4. I nourish myself so help me to help others.

Correct New: 4. I get to nourish myself so I can help others.

Original: We hope see our future anglershops There!

Correct Old: We hope to see our future anglershops there!

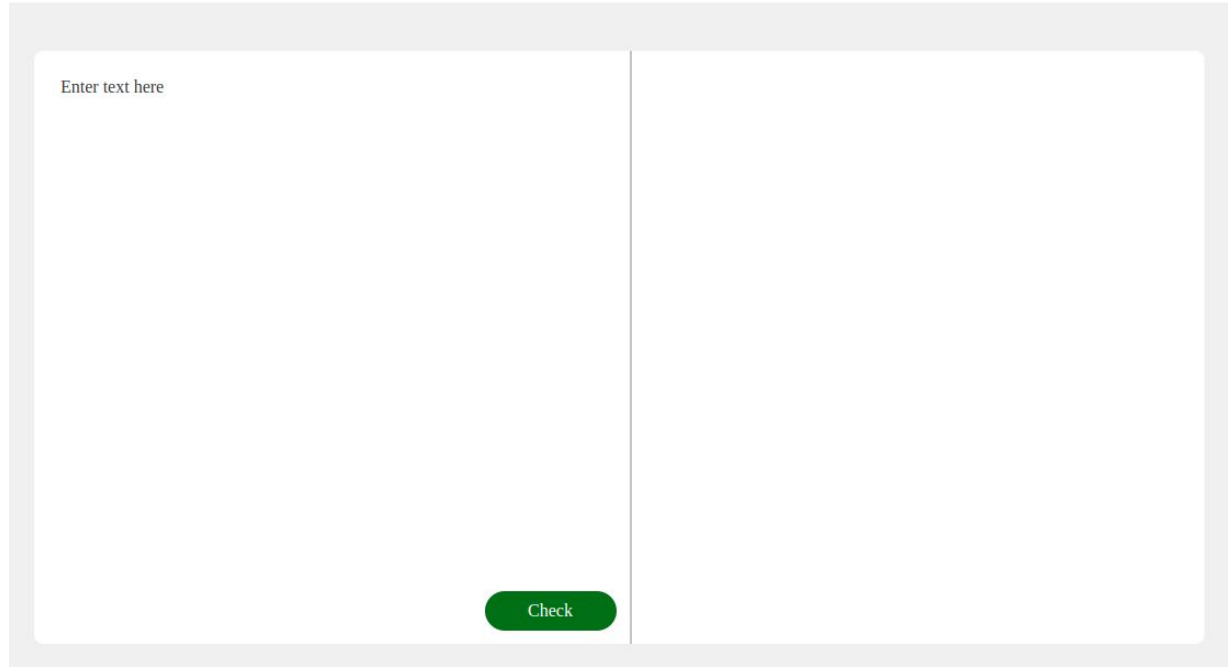
Correct New: We hope to see our future anglershops!

Loss 0.76 → 0.43 for a dataset with the same frequencies

Demo

- Created a web app
 - FastApi
 - HTML, CSS and JS
 - T5 model

CALL



The image shows a web application interface. It consists of a light gray rectangular frame. Inside the frame, there is a white rectangular area. At the top left of this white area, the text "Enter text here" is displayed in a small, gray font. At the bottom center of the white area, there is a green rounded rectangular button with the word "Check" written in white text. A thin vertical gray line divides the white area into two equal-width sections.

Future

- Improve the dataset from which the feature frequencies are extracted
- Exploring other model architectures than T5
- Fine-tune a better model
- Create an automated script that would do the whole process for you with the provided parameters
- Test how big of a difference more data would do on the fine-tuned model
- Test how the length of the sentence influences the model fine-tuning
- Getting a better understanding of the results
- Test other tools for generating synthetic training data from the frequencies such as Errgent
- Select data from a different dataset than C4
- More models, for example Hindi and Urdu

The background is a dark blue-grey color. It is decorated with various geometric elements: orange circles of different sizes, some with white dotted patterns; white circles and hexagons; orange hexagons; and white dotted patterns arranged in circles, hexagons, and triangles. There are also thin white lines and orange dots scattered throughout.

Questions?