Intelligent Linguistic System for the Grammar of the Romanian Language

EEML Summer School 2021 - Virtual Budapest, Hungary

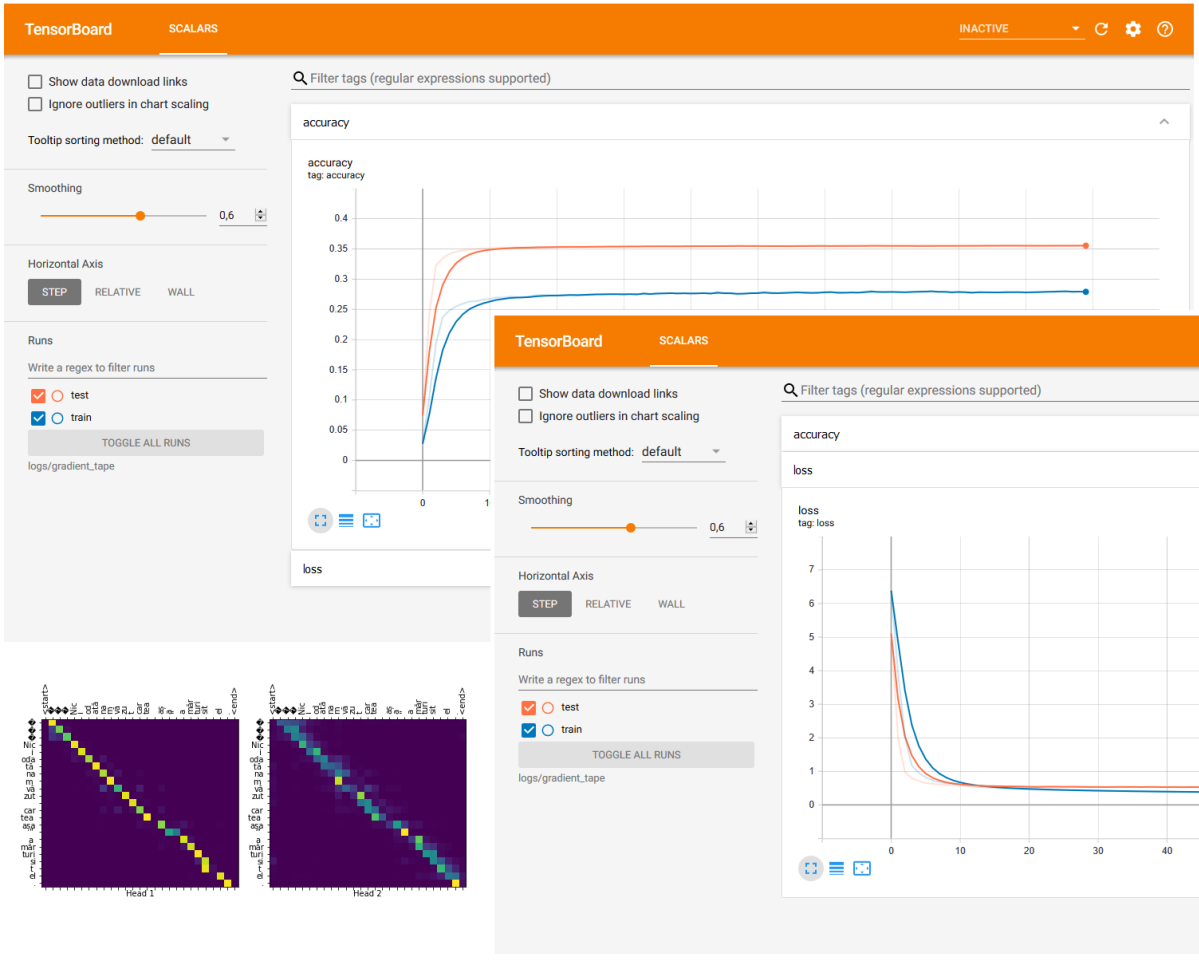
Ioan Florin Cătălin Nițu, Traian Eugen Rebedea

University Politehnica of Bucharest

June - July 2020

Abstract

The proposed correction system receives a sentence with grammatical errors and corrects it, using state-of-the-art technologies to perform this operation such as attention-based Transformers. The paper uses RONACC, the first corpus for grammatical corrections in Romanian for modeling, training, testing, and validating the project. Using a very large data set with over a million learning examples, an average BLEU score of 45.29 points was obtained, in a rather short training time (only two hours for 5 epochs) executed on several GPUs. However, even a small data set of only fifty thousand examples with as many as one hundred epochs achieves an average BLEU score of 33.29 points in three hours.
[Niț u and Rebedea, 2020]


TensorBoard
SCALARS
INACTIVE
Show data download links
Ignore outliers in chart scaling
Tooltip sorting method: default
Smoothing
0,6
Horizontal Axis
STEP
RELATIVE
WALL
Runs
Write a regex to filter runs
test
train
TOGGLE ALL RUNS
logs/gradient_tape
Filter tags (regular expressions supported)
accuracy
accuracy
tag: accuracy
loss
loss
tag: loss
Head 1
Head 2

Outline

Motivation

Correcting texts and natural language, in general, is often encountered in computer science, artificial intelligence, and machine learning. The study of natural language processing has brought technologies, models, and products that can competitively address this issue.

Technologies Used

Python and TensorFlow to implement a Transformer-based solution proposed in the article Attention Is All You Need [Vaswani et al., 2017]:

- Attention layers
- Encoder-Decoder Transformer

Proposed Method

The data set: RONACC corpus (pairs of lines of correct and wrong sentences).
Training is done in several epochs. The model is not pre-trained.
The transformer is an auto-regressive model [Graves, 2013].
As the transformer predicts each word, attention allows it to look at the previous ones in the input sequence to predict the next one [TensorFlow, 2019].

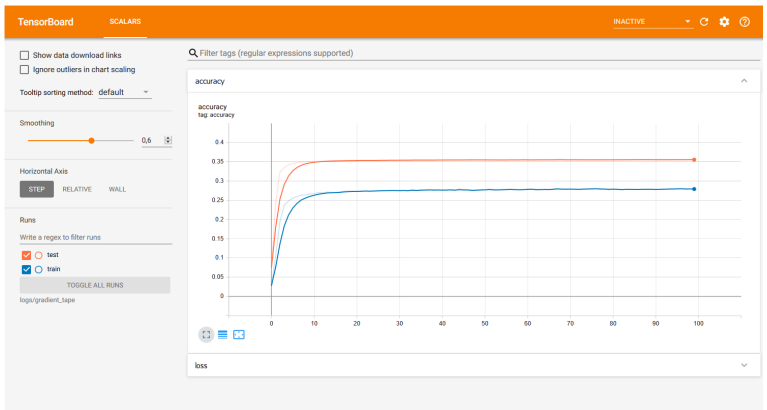
RONACC

Data sets for:

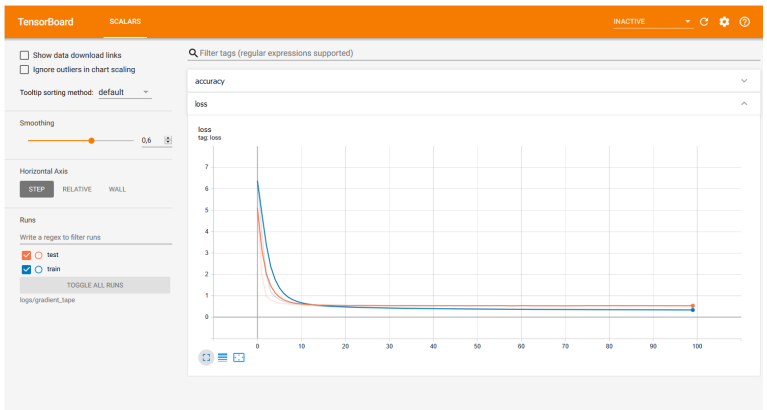
- training:
  - small: 7082 pairs of sentences
  - medium: 7082 + 50000 pairs of sentences
  - large: 7082 + 1000000 pairs of sentences
- testing: 1519 pairs of sentences
- validation: 1518 pairs of sentences

Training

Trained 100 epochs on the (small and) medium data set, testing and measuring accuracy (see Figure 1(a)) and loss (see Figure 1(b)).



(a) Accuracy



(b) Loss

Figure: Metrics for the medium data set

Evaluation

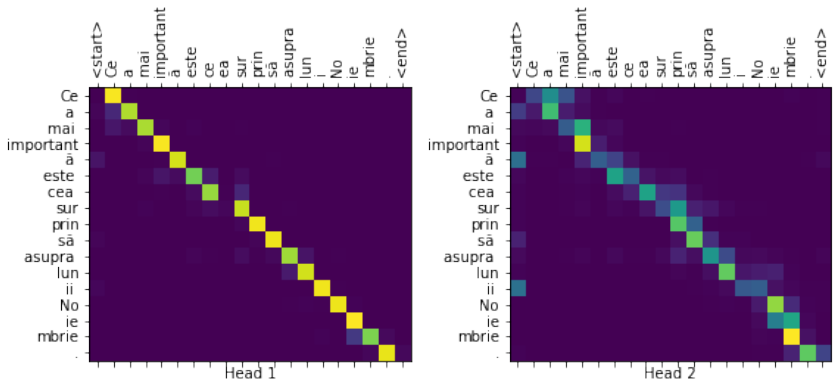
The model implemented in the presented work obtained a competitive BLEU score of 20.56 for text corrections in Romanian with a modest training set. With an increased training set, it got a score of 33.29, as shown in Table 1, representing a substantial improvement.

Table: Training results

Data set	Epochs	BLEU Score	Training time
Small	100	20.56	9.66 sec. / epoch
Medium	100	33.29	99 sec. / epoch
Large	5	45.29	1032 sec. / epoch

Example Attention (1)

- Input: Cea mai importantă este **ceea** surprinsă asupra **luni** Noiembrie.
- Predicted correction: Cea mai importantă este **cea** surprinsă asupra **lunii** Noiembrie.
- Real correction: Cea mai importantă este **cea** surprinsă asupra **lunii** Noiembrie.

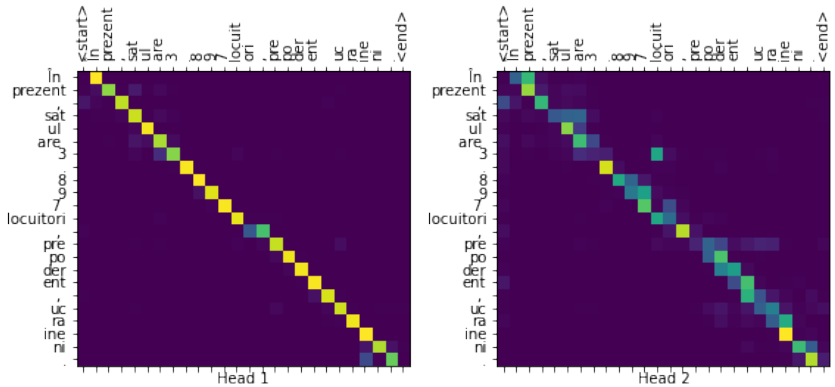


(a) BLEU score: 1.0000

Figure: Fully Corrected

Example Attention (2)

- Input: În prezent, satul are 3.897 locuitori, **prepoderent** ucraineni.
- Predicted correction: În prezent, satul are 3.897 locuitori, **prepoderent** ucraineni.
- Real correction: În prezent, satul are 3.897 locuitori, **preponderent** ucraineni.

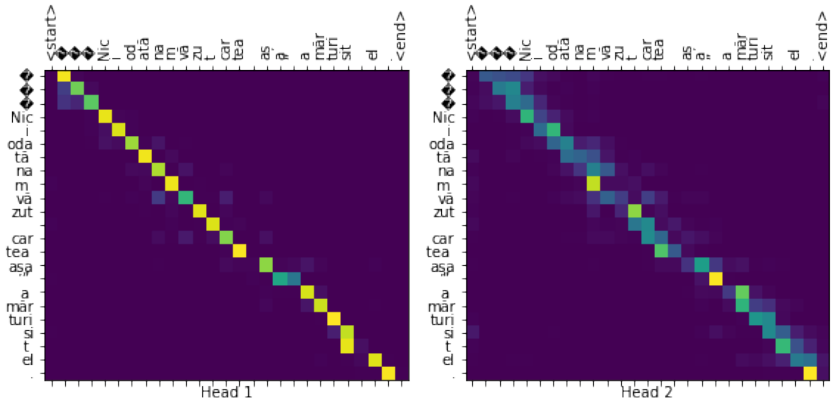


(a) BLEU score: 0.7071

Figure: Partialy Corrected

Example Attention (3)

- Input: **Nici odată nam** văzut cartea așa" a mărturisit el.
- Predicted correction: **Nici odată nam** văzut cartea așa" a mărturisit el.
- Real correction: **Niciodată n-am** văzut cartea așa", a mărturisit el.



(a) BLEU score: 0.2741

Figure: Sentence with Errors or that Could Not be Corrected

Conclusions

- Model proposed in "Attention Is All You Need" [Vaswani et al., 2017]:
  - trained twelve hours on eight P100 GPUs
- Implemented model:
  - on the medium data set: for almost three hours
  - on the large set: for an hour and a half (only five epochs)

Unfortunately, the field development did not make it possible to compare the implemented model with other models. This implementation:

- is a starting point in field development
- can be used in the construction and modeling of high-performance competitive technologies for text corrections in Romanian

Bibliography


Graves, A. (2013).
Generating sequences with recurrent neural networks.
arXiv preprint arXiv:1308.0850.


Niţu, I. F. C. and Rebedea, T. E. (2020).
Intelligent Linguistic System for the Grammar of the Romanian Language.
Bachelor's thesis, University Politehnica of Bucharest.
Faculty of Automatic Control and Computer Science.


TensorFlow (2019).
Transformer model for language understanding.
https://www.tensorflow.org/tutorials/text/transformer.
Accessed: 2020-06-10.


Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017).
Attention is all you need.
arXiv preprint arXiv:1706.03762.

Thank you for your attention!

ioan_florin.nitu@stud.acs.upb.ro