Intelligent Linguistic System for the Grammar of the Romanian Language

Ioan Florin Cătălin Nițu, Traian Eugen Rebedea
University Politehnica of Bucharest

June - July 2020

Abstract

The field of natural language processing is not as strongly developed for the Romanian language as it is for others, as is the English language. Writing texts correctly has always been a necessity, and the development of tools that will be useful in this need is critical. 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 neural models like Encoder-Decoder Transformers. These are a cornerstone in the development of intelligent tools for processing – translating, summarizing, or proofreading – texts and are the foundation for this project. 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 five 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. [Nițu and Rebedea, 2020]

Keywords: Romanian language, grammar, transformers, attention, positional encoding

1 Introduction

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.

2 Related Word

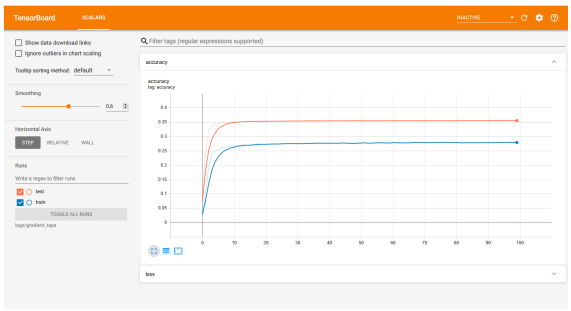
This paper uses Python and TensorFlow to implement a Transformer-based solution proposed in the article Attention Is All You Need [Vaswani et al., 2017]. Attention mechanisms have become an integral part of persuasive sequence models and reasoning models in different tasks, allowing dependency modeling without considering their distance in input or output sequences [Bahdanau et al., 2014]. In most cases, such attention mechanisms are used in conjunction with a recurring network. Transformers are introduced as a model architecture that avoids recurrences and relies on an attention mechanism to attract global dependencies between input and output. The transformer allows a significantly higher parallelization and can reach a new stage in terms of the quality of translation or correction of texts after training [Vaswani et al., 2017].

3 Method

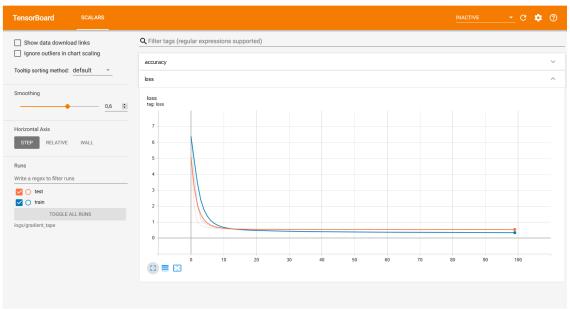
The data set used to train, validate and test the model is the RONACC corpus. The data set consists of pairs of lines of correct and wrong sentences. Training is done in several epochs. From time to time, after some epochs, a checkpoint is made to save the training done. The transformer is an auto-regressive model [Graves, 2013]: it makes predictions one at a time and uses its output so far to decide what to do next [TensorFlow, 2019]. 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].

4 Experiments

The model was evaluated on several sets of training data and a test set with 1519 pairs of sentences: small: 7082 pairs of sentences; medium: 7082 + 50000 pairs of sentences; large: 7082 + 1000000 pairs of sentences.



(a) Accuracy



(b) Loss

Figure 1: Metrics for the medium data set

The model was trained 100 epochs on the (small and) medium data set, at each one testing and measuring accuracy (see Figure 1(a)) and loss (see Figure 1(b)). In the end, the results were validated using a validation set of 1518 pairs of sentences. From the graphs, it can be seen how these values converge, and also, the loss tends to zero. 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 1: Training results

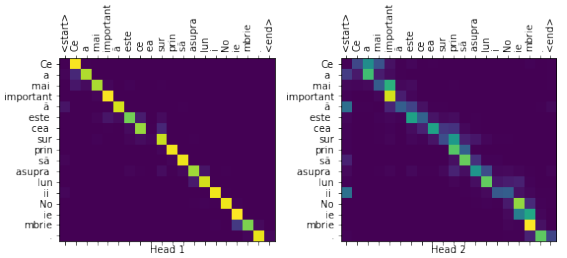
Data set	Epochs	BLEU Score	Training time
Small	100	20.56	9.66 sec. / epoch
Medium	100	33.29	99 sec. / epoch
Large	5	45.29	1032 sec. / epoch

5 Interpretation of the Results

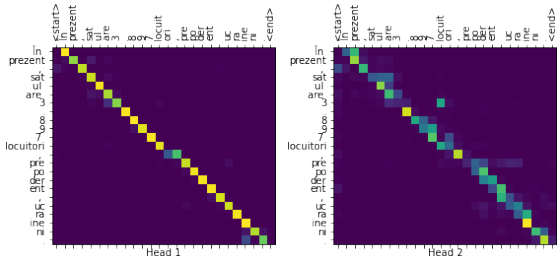
In Figure 2, some examples of running the model on different input sentences are presented. Attention was measured at the output (second substrate) of the last decoder layer of the decoder. The examples presented in the figure on the following page are also fully corrected cases 2(a), partially corrected sentences 2(b), and also sentences with errors or that could not be corrected 2(c).

6 Conclusion

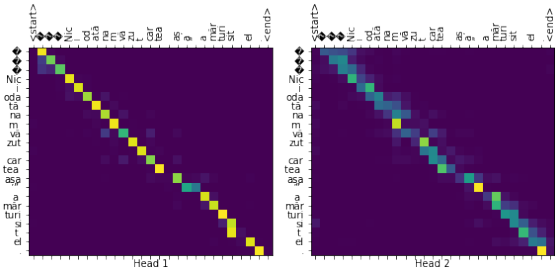
Compared to the model proposed in "Attention Is All You Need" [Vaswani et al., 2017], which was trained twelve hours on eight P100 GPUs, the implemented model was prepared on the medium data set for almost three hours and on the large set for an hour and a half (only five epochs). Only one super data set was used for modeling, and no comparisons could be made with other data sets. Unfortunately, the field development did not make it possible to compare the implemented model with other models. Furthermore, this implementation is a starting point in field development. It can be used in the construction and modeling of high-performance competitive technologies for text corrections in Romanian.



(a) BLEU score: 1.0000



(b) BLEU score: 0.7071



(c) BLEU score: 0.2741

Figure 2: Attention Examples

References

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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

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

[Vaswani et al., 2017] 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*.