**EN - RO Translator**

**I) Naive Approach**

**Method:**

**Pre-processing:**

* Letters were converted to lowercase;
* Punctuation was removed;
* [start] and [end] tags were added to Romanian sentences.

**Tokenizer:**

* Word-based;
* Vocabulary size is 15000;
* Sequence length was truncated to 20 words.

**Model:**

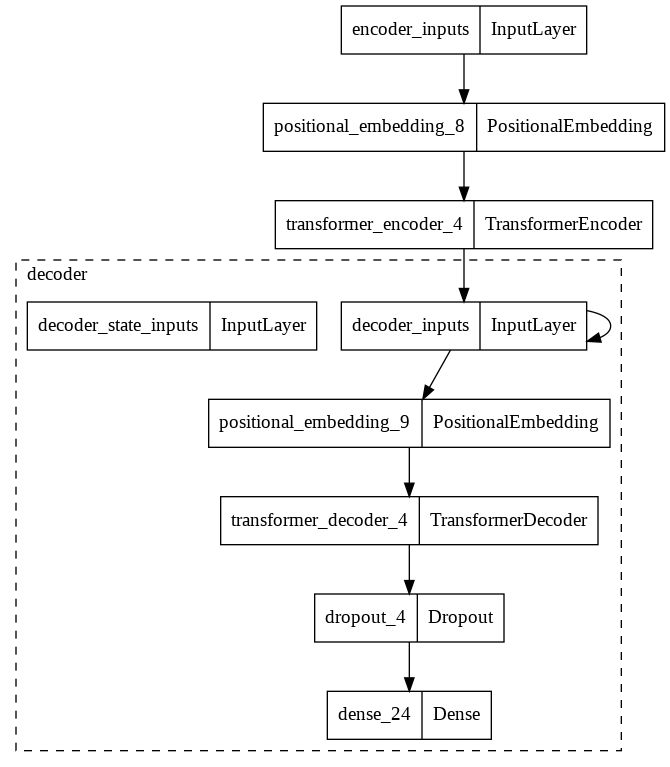
* Transformer;
* ~20 mil parameters.

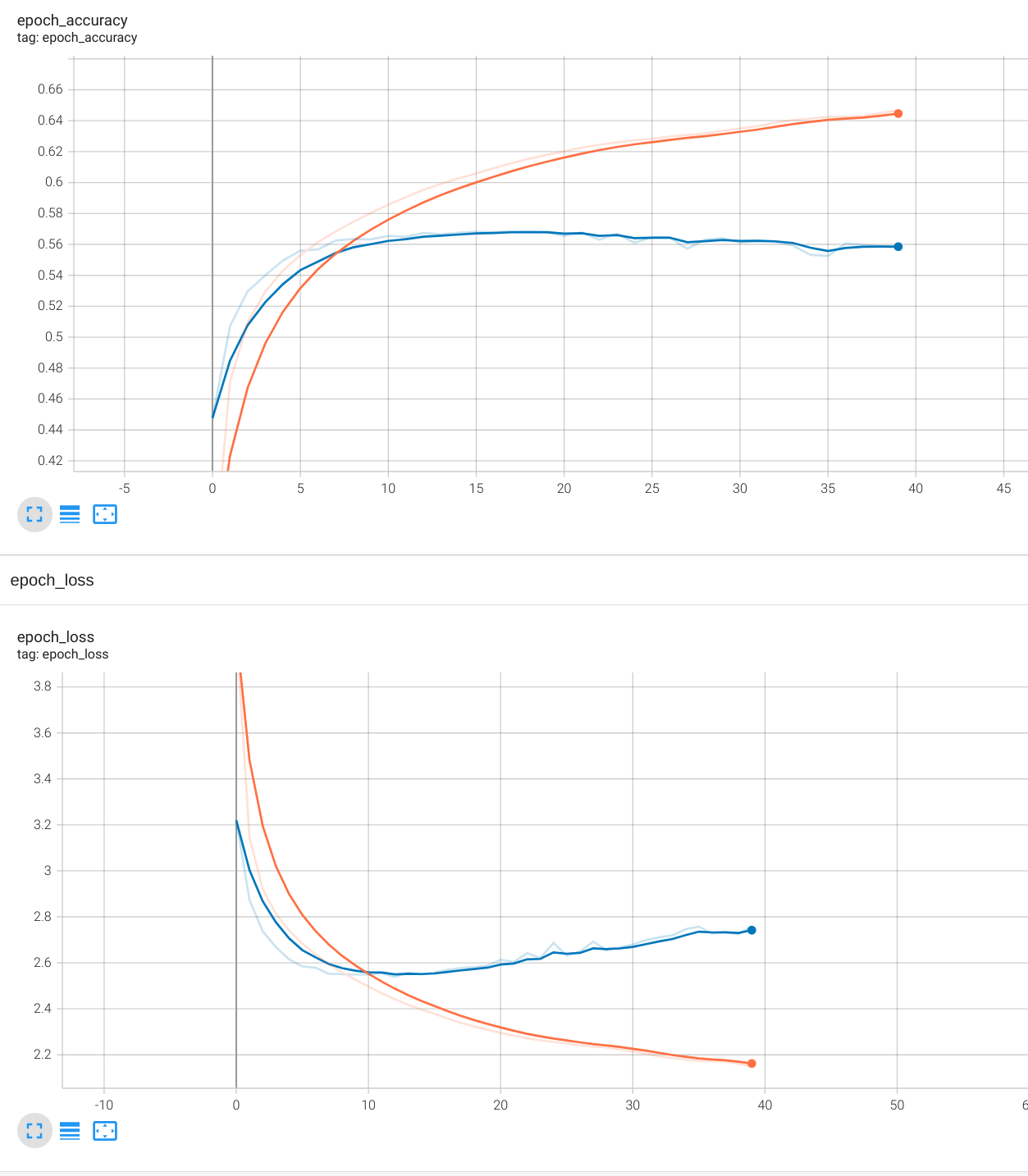
**Training:**

* RMSprop;
* 40 epochs;
* For validation accuracy was tracked.

**Decoding:**

* Greedy.





**Discussion:**

**Metric:**

* The BLEU score obtained from raw sequences was 0.10;
* The BLEU score obtained after minimal post-processing (removing special tokens and punctuation from both sequences) was 1.39;
* Accuracy (on the validation set) was 0.5584.

**Examples:**

* “Madam President, Commissioner, ladies and gentlemen, I would like to thank all of the political groups and everyone who has spoken for the support and the constructive tone of their speeches. [start] doamnă preşedintă domnule comisar doamnelor şi domnilor aş dori să mulţumesc tuturor grupurilor politice şi tuturor celor care au [UNK]”;
* “It does not speak with one coherent voice, but with many voices. [start] nu este vorba de o voce clară [UNK] dar multe [UNK] end”;
* “(The President cut off the speaker) [start] preşedintele a întrerupto pe vorbitoare end”;
* “Stateless people are a separate issue and should be encouraged, using all means available, to apply for citizenship in their host country. [start] [UNK] oameni reprezintă o problemă și ar trebui [UNK] toate mijloacele disponibile pentru a aplica în vederea [UNK] în [UNK]”;
* “The debate is closed. [start] dezbaterea a fost închisă end”.

**Limitations:**

* Many words were replaced with [UNK]. Even basic words were sometimes replaced when used in the plural form. This is partially attributed to the vocabulary size but more importantly, it’s a limitation of the tokenizer that we used. A good solution would be tokenizing sub-words such that grammar does not obstruct word lemmas;
* Sequence length was too short so some sentences were truncated (this can be partially avoided by running the model on multiple slices of the input sentence);
* The model doesn't output punctuations;
* Sometimes, sentences contain mistakes that are a direct result of our model’s limited learning capacity. Performance could be improved by using more recent architectures, pre-trained weights and bigger neural networks.

**Resources:**

* https://keras.io/examples/nlp/neural\_machine\_translation\_with\_transformer/
* https://medium.com/@tskumar1320/how-to-fine-tune-pre-trained-language-translation-model-3e8a6aace9f