Sklokin, Svyatoslav
Dufresne, Éléa

Deliverable 3

22/02/2021
MAIS 202

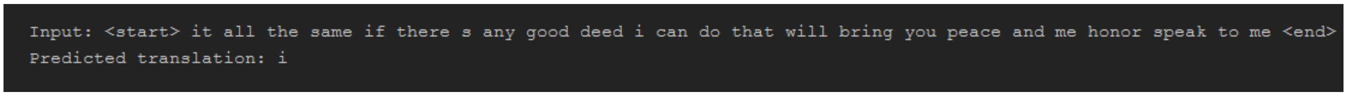
1. Compare your final results to your preliminary ones (from the previous deliverable). Have you changed anything in your model since the previous deliverable? If so, how have your changes improved the results?

   Now, focus on your final results. Once again, present a detailed analysis of your results, provide graphs as appropriate. Analysis requirements differ in every field, but reporting BLEU score with brevity penalty is mandatory for text generation.

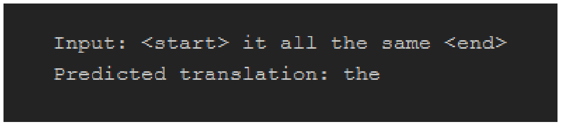
Answer

We trained the model more (we are at 1200 epochs), with hyperparameters being a scaled down version of the optimal ones presented in the Attention is All You Need paper. The algorithm now seems to output sentences whose meaning might be lost but the syntax is generally correct. The preliminary results where sentences but generally did not have proper syntax for longer input. Below are some example outputs from different stages.

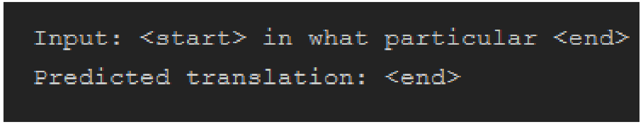
First results



```
Input: <start> it all the same if there s any good deed i can do that will bring you peace and me honor speak to me <end>
Predicted translation: i
```

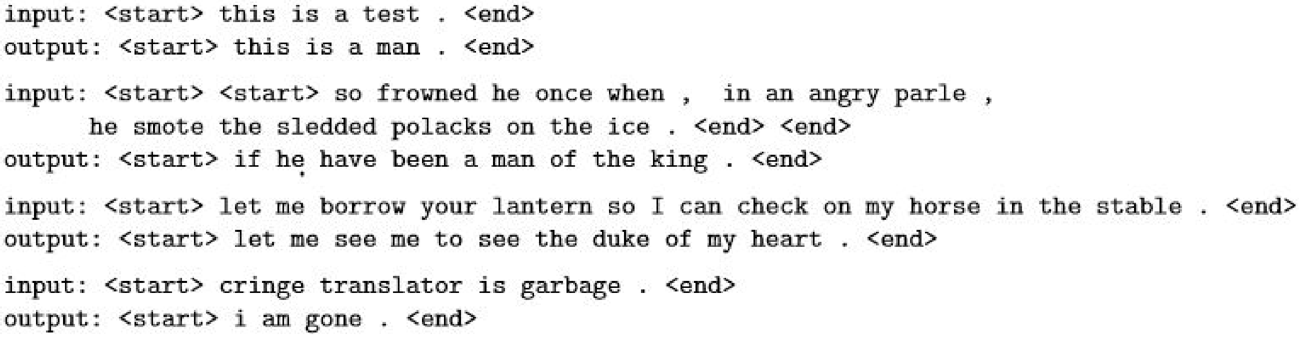


```
Input: <start> it all the same <end>
Predicted translation: the
```



```
Input: <start> in what particular <end>
Predicted translation: <end>
```

Previous results



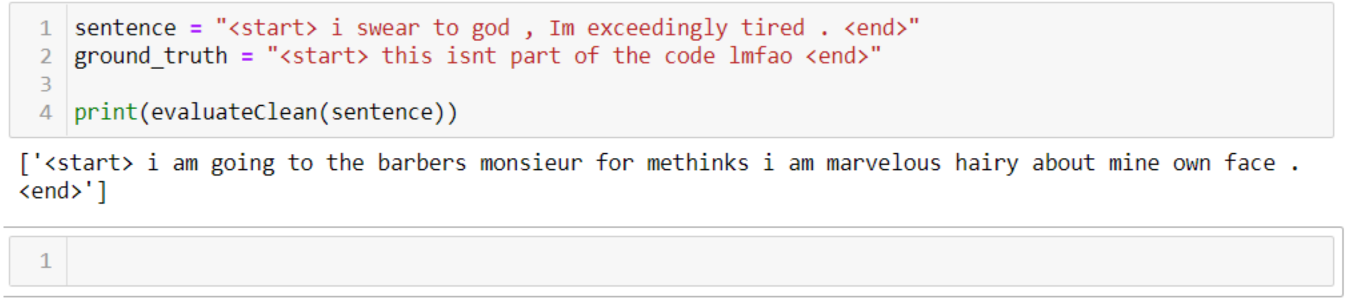
```
input: <start> this is a test . <end>
output: <start> this is a man . <end>

input: <start> <start> so frowned he once when ,  in an angry parle ,
      he smote the sledded polacks on the ice . <end> <end>
output: <start> if he have been a man of the king . <end>

input: <start> let me borrow your lantern so I can check on my horse in the stable . <end>
output: <start> let me see me to see the duke of my heart . <end>

input: <start> cringe translator is garbage . <end>
output: <start> i am gone . <end>
```

Penultimate results

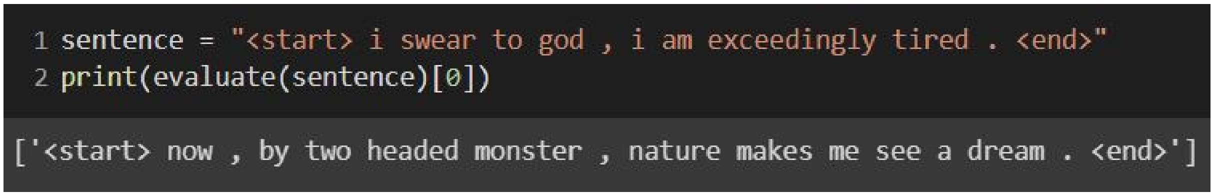


```
sentence = "<start> i swear to god , Im exceedingly tired . <end>"
ground_truth = "<start> this isnt part of the code lmfao <end>"

print(evaluateClean(sentence))
```

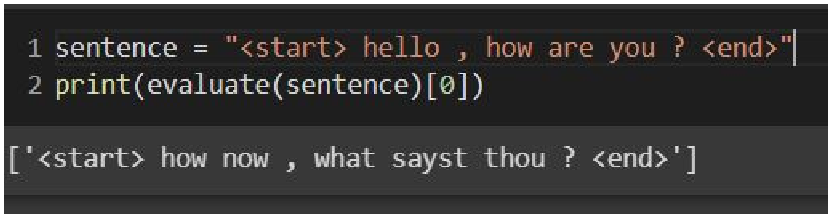
```
['<start> i am going to the barbers monsieur for methinks i am marvelous hairy about mine own face .
<end>']
```

Current results



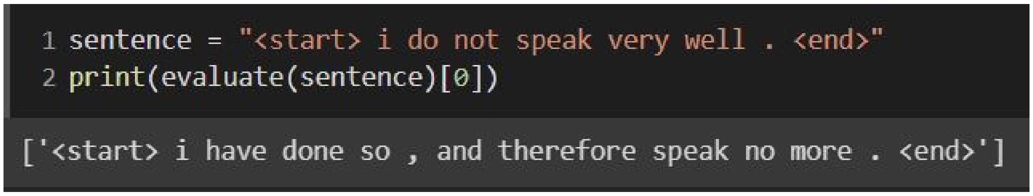
```
sentence = "<start> i swear to god , i am exceedingly tired . <end>"
print(evaluate(sentence)[0])
```

```
['<start> now , by two headed monster , nature makes me see a dream . <end>']
```



```
sentence = "<start> hello , how are you ? <end>"
print(evaluate(sentence)[0])
```

```
['<start> how now , what sayst thou ? <end>']
```



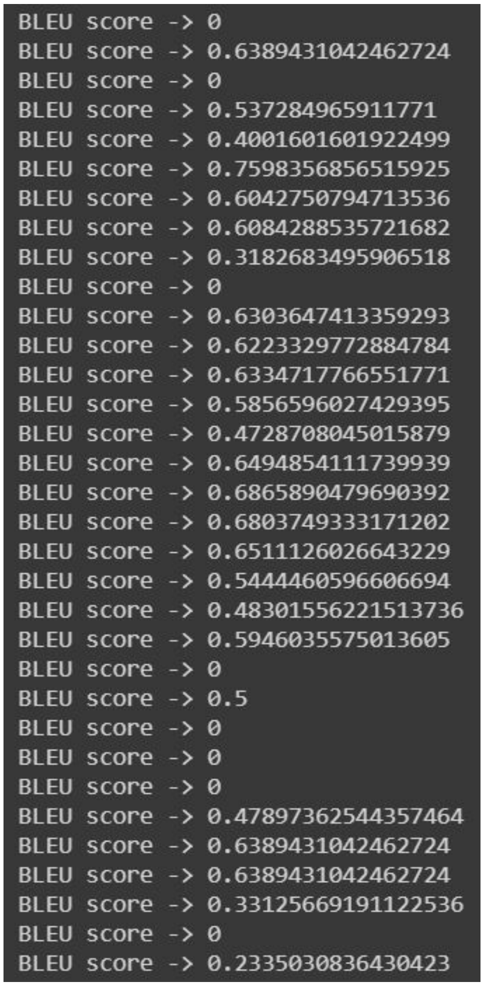
```
sentence = "<start> i do not speak very well . <end>"
print(evaluate(sentence)[0])
```

```
['<start> i have done so , and therefore speak no more . <end>']
```

We have spent hours trying to generate attention plots for our results but are unable to as of yet, so unfortunately we have no such graph to present.

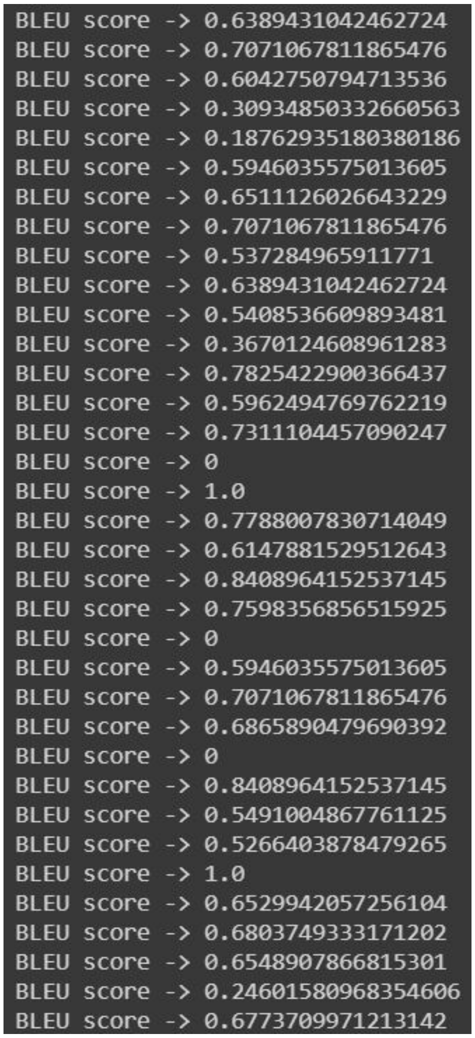
We computed the BLEU score for a small batch of the training set and a test set composed of two Shakespeare plays that we had not use in the original dataset. We obtained a score of 0.5057138533013059 for the former and of 0.4577821393596426 for the later. According to this Google Cloud guide the scores qualify respectively as "Very high quality, adequate, and fluent translations" and "High quality translations" which is more than what we expected. The most notable difference in the computations of the BLEU score is that the test set resulted in 0s for more translations and had fewer 1s than the training set (see below) which indicates potential overfitting. However, as the average scores are not so far apart from each other, the model does not seem to significantly overfit the training set.

test set



```
BLEU score -> 0
BLEU score -> 0.6389431042462724
BLEU score -> 0
BLEU score -> 0.537284965911771
BLEU score -> 0.4001601601922499
BLEU score -> 0.7598356856515925
BLEU score -> 0.6042750794713536
BLEU score -> 0.6084288535721682
BLEU score -> 0.3182683495906518
BLEU score -> 0
BLEU score -> 0.6303647413359293
BLEU score -> 0.6223329772884784
BLEU score -> 0.6334717766551771
BLEU score -> 0.5856596027429395
BLEU score -> 0.4728708045015879
BLEU score -> 0.6494854111739939
BLEU score -> 0.6865890479690392
BLEU score -> 0.6803749333171202
BLEU score -> 0.6511126026643229
BLEU score -> 0.5444460596606694
BLEU score -> 0.48301556221513736
BLEU score -> 0.5946035575013605
BLEU score -> 0
BLEU score -> 0.5
BLEU score -> 0
BLEU score -> 0
BLEU score -> 0
BLEU score -> 0.47897362544357464
BLEU score -> 0.6389431042462724
BLEU score -> 0.6389431042462724
BLEU score -> 0.33125669191122536
BLEU score -> 0
BLEU score -> 0.2335030836430423
```

training set



```
BLEU score -> 0.6389431042462724
BLEU score -> 0.7071067811865476
BLEU score -> 0.6042750794713536
BLEU score -> 0.30934850332660563
BLEU score -> 0.18762935180380186
BLEU score -> 0.5946035575013605
BLEU score -> 0.6511126026643229
BLEU score -> 0.7071067811865476
BLEU score -> 0.537284965911771
BLEU score -> 0.6389431042462724
BLEU score -> 0.5408536609893481
BLEU score -> 0.3670124608961283
BLEU score -> 0.7825422900366437
BLEU score -> 0.5962494769762219
BLEU score -> 0.7311104457090247
BLEU score -> 0
BLEU score -> 1.0
BLEU score -> 0.7788007830714049
BLEU score -> 0.6147881529512643
BLEU score -> 0.8408964152537145
BLEU score -> 0.7598356856515925
BLEU score -> 0
BLEU score -> 0.5946035575013605
BLEU score -> 0.7071067811865476
BLEU score -> 0.6865890479690392
BLEU score -> 0
BLEU score -> 0.8408964152537145
BLEU score -> 0.5491004867761125
BLEU score -> 0.5266403878479265
BLEU score -> 1.0
BLEU score -> 0.6529942057256104
BLEU score -> 0.6803749333171202
BLEU score -> 0.6548907866815301
BLEU score -> 0.24601580968354606
BLEU score -> 0.6773709971213142
```

We can see from the results presented at the beginning that the latest translated sentences not only have proper syntax but mostly (or somewhat) keeps the meaning of what is conveyed while sounding like Shakespearean English. Note that meaning/synonyms is not something that the BLEU score accounts for and so the fact that it is partly conserved raises our confidence in the model.

2. **Final demonstration proposal:** Now that you trained your model, it is time for you to integrate it in a final product. Don't forget to save your trained weights! You will need them for the integration and/or testing of your model.

   Application: We want all of you to at least have a landing page-type website to demo your model and results. For more experienced developers, you are welcome to choose something more advanced.

   Discuss your final product, and final integration approach. Describe and justify the choice of stacks and technologies. Provide diagrams as appropriate. Explain your experiences with the technologies you have proposed. If you do not have any, explain how you would come to learn them (e.g: online tutorials, etc).

   Do not worry if you do not have any experience in webdev or other types of software development. Discuss with your project leaders, or any other execs, and we will help you out!

Answer
Our final product is only a one way translator instead of two-ways as we had planned. It is also more of a generator of somewhat related shakespearean sentences or sometimes of something resembling a metaphorical interpretation of the input, rather than an actual translator. As mentioned in the previous deliverable we will use Flask for the web application, and we will host it on mimi.cs.mcgill. We plan on doing something similar in appearance to the Google translation's interface. We have no prior experience with this software but will watch the workshop recording and go through the repository code. We chose Flask because it was recommended by the TPMs.