**English to Hebrew Translation using a Sequence-to-Sequence Transformer Model**

page1image44844352

By: Talia Dagan

Dr. Shanker, Advanced Machine Learning

Kent State University

May 9th, 2022

**Objective:**

To improve performance of a sequence-to-sequence translation transformer to translate English to the Hebrew language.

**Data:**

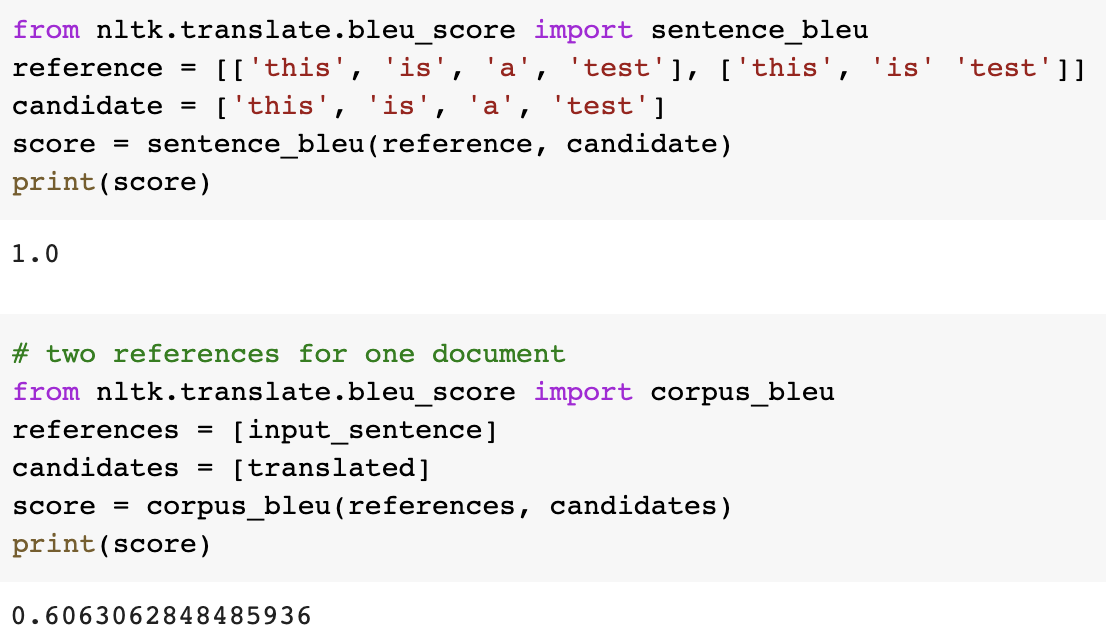
Data is an English to Hebrew translation dataset from Anki. Anki is a computer program with intelligent flash cards. These flash cards include different language translations. There are 126,832 rows of data. The data can be found on <https://www.manythings.org/anki/> along with many other languages translated pairs to English.

**Approach:**

The approach I used to improve the performance of a sequence-to-sequence transformer to translate the language was changing the parameters of different models to choose the best parameters that increased the model’s performance. One of the first changes I did was changed the batch size from 32, to 64, to 128. The second change I made was changing the sequence length from 20 to 15. Since the sequence length is the length of the sequence of input data, I went through my data and found that the max sentence was over 15 words long. I also changed the data partition from 90% training to 70% training. I then used two different versions of the same dataset to see if it would change the accuracy. First, I used the original dataset with 126,833 pairs of translated English and Hebrew words/sentences. I then used a different version of the same data – the only change being that I deleted duplicate words. As seen in Figure B, there are multiple rows for the same word such as “Go on.” This is due to there being multiple ways to say the same thing in Hebrew. I deleted these duplicated words and ended up with 101,655 rows of data. To determine accuracy, I used BLEU which stands for bilingual evaluation understudy as a metric to evaluate the translated sentences. The BLEU score is a number between 0 and 1 and a score of 1 represents a perfectly translated sentence. In Figure C*,* a screenshot is shown of a test I did to show a perfectly translated sentence with a score of 1.

**Figure B**:**Figure C:**

Table

Description automatically generated 

**Results:**

To improve accuracy, I changed multiple parameters and used different models to choose the best parameters with the best accuracy. To start, I used different data partitions. Originally, I used a data partition of 90% training, 5% validation, and 5% test and had an extremely low accuracy. Having too large of a training set lead to problems, which resulted in a change in my data partition to be 70% training, 15% validation, and 15% test. This proved to be more successful. After changing the data partition, the three models I ran were ran for 20 epochs and a 20-sequence length. I changed the batch size from 32, to 64, to 128. When I ran the model at 30 epochs, I got an accuracy of .5453 or 54.53%. When I ran the model at 64 epochs, I got an accuracy of .6093 (60.93%). When I ran the same parameters with the only change being the batch size of 128, I got my highest accuracy of .6997 or 69.97%. As my results indicate, a larger batch size increases the accuracy of the model. Secondly, I decided to use the same parameters or 20 sequence length and 128 batch size to build another model with the only difference being that I ran it for 30 epochs. I ran that model, and an accuracy of 64.18% was found. The next model parameter change I made was that I decreased the sequence length from 20 to a size of 15 sequence length. Running this model at 128 batches and 30 epochs resulted in an accuracy of 63.88%. I then ran an additional model with an altered version of the original dataset with the only difference being that I deleted duplicated words. I ran this model with a batch size of 128 (as this batch size proved to be the most successful) and a sequence length of 20. I ran this model for 15 epochs and resulted at an accuracy of 63.88% at epoch 12. For my highest accuracy model, I used BLEU. As seen in *Figure D,* the BLEU score is 60.63%. If the input sentence and translated the sentence were accurately translated, then the score would have been 1.

***Figure D:***

**Graphical user interface, text, application, email

Description automatically generated**

**Conclusions:**

In conclusion, it is found to get the best accuracy on a sequence-to-sequence transformer model, you should use a model with a batch size of 128 to get the best performance. Additionally, a dataset with thousands of pairs of English and another language sentence and word pairs must be used so that the model has a large enough training to learn the worlds. Lastly, decreasing the sequence length will decrease the accuracy. This is subject to change and is dependent of the dataset that is used for the model, but for my model the max sentence size was 18 which is why a 15-character length had a lesser accuracy. To conclude, to achieve a higher accuracy when using a sequence-to-sequence transformer, use a large batch size, a large dataset, and change the character length to represent your own data.