Model training and Evaluation:

Model training using optimizer as torch.adam.optim a form of stochastic gradient descent which generalizes quickly and when trained the model for around 100 epochs the training loss was reducing, but after 100 epochs it seemed that the model was overfitting. Therefore, stopped the training. The validation loss goes up because it is quite difficult for sequence2sequence model to work well with validation dataset, and transformers would have given better results. However, with optimizer set to the traditional SGD or stochastic gradient descent and learning rate=0.0001 the model seems to train very slowly and validation loss also decreases gradually. For training the dataset size = 20000 -> questions and answers. However, A lot of them are filtered and only pairs of max length up to 10 or less than 10 are used to create vocab and actual training dataset. Again there is room for experimentation.

optimizer as torch.adam.optim and loss function as cross entropy loss

Epoch: 01 | Time: 0m 19s
Train Loss: 6.559 | Train PPL: 705.742
Val. Loss: 6.376 | Val. PPL: 587.314
Epoch: 02 | Time: 0m 18s
Train Loss: 5.965 | Train PPL: 389.376
Val. Loss: 6.281 | Val. PPL: 534.189
Epoch: 03 | Time: 0m 18s
Train Loss: 5.794 | Train PPL: 328.360
Val. Loss: 6.329 | Val. PPL: 560.555
Epoch: 04 | Time: 0m 18s
Train Loss: 5.660 | Train PPL: 287.210
Val. Loss: 6.380 | Val. PPL: 589.781
Epoch: 05 | Time: 0m 18s
Train Loss: 5.589 | Train PPL: 267.541
Val. Loss: 6.434 | Val. PPL: 622.606
Epoch: 06 | Time: 0m 18s
Train Loss: 5.525 | Train PPL: 250.858
Val. Loss: 6.712 | Val. PPL: 822.111
Epoch: 07 | Time: 0m 18s
Train Loss: 5.412 | Train PPL: 224.101
Val. Loss: 6.627 | Val. PPL: 754.968
Epoch: 08 | Time: 0m 18s
Train Loss: 5.324 | Train PPL: 205.199
Val. Loss: 6.733 | Val. PPL: 839.705

Epoch: 09 | Time: 0m 18s
Train Loss: 5.224 | Train PPL: 185.595
Val. Loss: 6.771 | Val. PPL: 872.470
Epoch: 10 | Time: 0m 18s
Train Loss: 5.133 | Train PPL: 169.609
Val. Loss: 6.830 | Val. PPL: 925.088
Epoch: 11 | Time: 0m 19s
Train Loss: 5.092 | Train PPL: 162.645
Val. Loss: 6.948 | Val. PPL: 1040.955
Epoch: 12 | Time: 0m 19s
Train Loss: 5.015 | Train PPL: 150.720
Val. Loss: 7.045 | Val. PPL: 1147.378
Epoch: 13 | Time: 0m 19s
Train Loss: 4.951 | Train PPL: 141.279
Val. Loss: 7.225 | Val. PPL: 1372.701
Epoch: 14 | Time: 0m 19s
Train Loss: 4.931 | Train PPL: 138.452
Val. Loss: 7.319 | Val. PPL: 1508.630
Epoch: 15 | Time: 0m 19s
Train Loss: 4.825 | Train PPL: 124.628
Val. Loss: 7.382 | Val. PPL: 1607.274
Epoch: 16 | Time: 0m 19s
Train Loss: 4.760 | Train PPL: 116.774
Val. Loss: 7.402 | Val. PPL: 1638.939
Epoch: 17 | Time: 0m 19s
Train Loss: 4.729 | Train PPL: 113.220
Val. Loss: 7.414 | Val. PPL: 1659.332
Epoch: 18 | Time: 0m 19s
Train Loss: 4.718 | Train PPL: 111.983
Val. Loss: 7.700 | Val. PPL: 2207.281
Epoch: 19 | Time: 0m 19s
Train Loss: 4.689 | Train PPL: 108.692
Val. Loss: 7.567 | Val. PPL: 1933.021
Epoch: 20 | Time: 0m 19s
Train Loss: 4.596 | Train PPL: 99.069
Val. Loss: 7.523 | Val. PPL: 1850.839
Epoch: 21 | Time: 0m 19s
Train Loss: 4.505 | Train PPL: 90.428
Val. Loss: 7.692 | Val. PPL: 2190.632
Epoch: 22 | Time: 0m 19s

Train Loss: 4.476 | Train PPL: 87.908
Val. Loss: 7.891 | Val. PPL: 2673.096
Epoch: 23 | Time: 0m 19s
Train Loss: 4.395 | Train PPL: 81.048
Val. Loss: 7.776 | Val. PPL: 2382.084
Epoch: 24 | Time: 0m 19s
Train Loss: 4.296 | Train PPL: 73.369
Val. Loss: 7.803 | Val. PPL: 2447.826
Epoch: 25 | Time: 0m 19s
Train Loss: 4.240 | Train PPL: 69.430
Val. Loss: 7.984 | Val. PPL: 2934.726
Epoch: 26 | Time: 0m 19s
Train Loss: 4.167 | Train PPL: 64.489
Val. Loss: 8.201 | Val. PPL: 3645.244
Epoch: 27 | Time: 0m 19s
Train Loss: 4.168 | Train PPL: 64.573
Val. Loss: 8.357 | Val. PPL: 4260.665
Epoch: 28 | Time: 0m 19s
Train Loss: 4.041 | Train PPL: 56.864
Val. Loss: 8.473 | Val. PPL: 4785.048
Epoch: 29 | Time: 0m 19s
Train Loss: 3.996 | Train PPL: 54.384
Val. Loss: 8.534 | Val. PPL: 5085.891
Epoch: 30 | Time: 0m 20s
Train Loss: 3.970 | Train PPL: 52.964
Val. Loss: 8.492 | Val. PPL: 4875.945
Epoch: 31 | Time: 0m 19s
Train Loss: 3.832 | Train PPL: 46.146
Val. Loss: 8.558 | Val. PPL: 5207.905
Epoch: 32 | Time: 0m 19s
Train Loss: 3.819 | Train PPL: 45.567
Val. Loss: 8.554 | Val. PPL: 5186.290
Epoch: 33 | Time: 0m 19s
Train Loss: 3.758 | Train PPL: 42.874
Val. Loss: 8.718 | Val. PPL: 6113.403
Epoch: 34 | Time: 0m 19s
Train Loss: 3.642 | Train PPL: 38.167
Val. Loss: 9.011 | Val. PPL: 8191.395
Epoch: 35 | Time: 0m 19s
Train Loss: 3.540 | Train PPL: 34.471

```
	Val. Loss: 9.168 |  Val. PPL: 9585.885
Epoch: 36 | Time: 0m 19s
	Train Loss: 3.456 | Train PPL:  31.686
	Val. Loss: 9.112 |  Val. PPL: 9064.540
Epoch: 37 | Time: 0m 19s
	Train Loss: 3.385 | Train PPL:  29.530
	Val. Loss: 9.308 |  Val. PPL: 11021.087
Epoch: 38 | Time: 0m 19s
	Train Loss: 3.270 | Train PPL:  26.306
	Val. Loss: 9.473 |  Val. PPL: 13009.255
Epoch: 39 | Time: 0m 19s
	Train Loss: 3.210 | Train PPL:  24.769
	Val. Loss: 9.422 |  Val. PPL: 12354.691
Epoch: 40 | Time: 0m 19s
	Train Loss: 3.072 | Train PPL:  21.581
	Val. Loss: 9.509 |  Val. PPL: 13483.912
Epoch: 41 | Time: 0m 19s
	Train Loss: 3.026 | Train PPL:  20.614
	Val. Loss: 9.547 |  Val. PPL: 14001.701
Epoch: 42 | Time: 0m 19s
	Train Loss: 2.937 | Train PPL:  18.857
	Val. Loss: 9.530 |  Val. PPL: 13769.555
Epoch: 43 | Time: 0m 19s
	Train Loss: 2.877 | Train PPL:  17.764
	Val. Loss: 9.601 |  Val. PPL: 14785.712
Epoch: 44 | Time: 0m 19s
	Train Loss: 2.753 | Train PPL:  15.685
	Val. Loss: 9.813 |  Val. PPL: 18267.009
Epoch: 45 | Time: 0m 19s
	Train Loss: 2.696 | Train PPL:  14.817
	Val. Loss: 10.071 |  Val. PPL: 23647.067
Epoch: 46 | Time: 0m 19s
	Train Loss: 2.635 | Train PPL:  13.941
	Val. Loss: 10.124 |  Val. PPL: 24924.884
Epoch: 47 | Time: 0m 19s
	Train Loss: 2.611 | Train PPL:  13.607
	Val. Loss: 10.187 |  Val. PPL: 26567.874
Epoch: 48 | Time: 0m 19s
	Train Loss: 2.379 | Train PPL:  10.797
	Val. Loss: 10.125 |  Val. PPL: 24966.421
```

Epoch: 49 | Time: 0m 19s
Train Loss: 2.212 | Train PPL: 9.134
Val. Loss: 10.381 | Val. PPL: 32251.782
Epoch: 50 | Time: 0m 20s
Train Loss: 2.132 | Train PPL: 8.430
Val. Loss: 10.485 | Val. PPL: 35772.981
Epoch: 51 | Time: 0m 19s
Train Loss: 1.968 | Train PPL: 7.158
Val. Loss: 10.555 | Val. PPL: 38378.111
Epoch: 52 | Time: 0m 19s
Train Loss: 1.776 | Train PPL: 5.904
Val. Loss: 10.758 | Val. PPL: 46997.231
Epoch: 53 | Time: 0m 19s
Train Loss: 1.575 | Train PPL: 4.831
Val. Loss: 10.860 | Val. PPL: 52038.113
Epoch: 54 | Time: 0m 19s
Train Loss: 1.435 | Train PPL: 4.199
Val. Loss: 11.025 | Val. PPL: 61384.336
Epoch: 55 | Time: 0m 19s
Train Loss: 1.312 | Train PPL: 3.712
Val. Loss: 11.151 | Val. PPL: 69616.895
Epoch: 56 | Time: 0m 19s
Train Loss: 1.166 | Train PPL: 3.209
Val. Loss: 11.215 | Val. PPL: 74264.144
Epoch: 57 | Time: 0m 19s
Train Loss: 1.056 | Train PPL: 2.874
Val. Loss: 11.205 | Val. PPL: 73490.205
Epoch: 58 | Time: 0m 19s
Train Loss: 0.965 | Train PPL: 2.624
Val. Loss: 11.282 | Val. PPL: 79396.259
Epoch: 59 | Time: 0m 19s
Train Loss: 0.855 | Train PPL: 2.351
Val. Loss: 11.467 | Val. PPL: 95506.670
Epoch: 60 | Time: 0m 19s
Train Loss: 0.790 | Train PPL: 2.204
Val. Loss: 11.748 | Val. PPL: 126529.643
Epoch: 61 | Time: 0m 19s
Train Loss: 0.737 | Train PPL: 2.090
Val. Loss: 11.767 | Val. PPL: 128887.845
Epoch: 62 | Time: 0m 19s

Train Loss: 0.665 | Train PPL: 1.944
Val. Loss: 11.630 | Val. PPL: 112456.578
Epoch: 63 | Time: 0m 19s
Train Loss: 0.596 | Train PPL: 1.814
Val. Loss: 11.593 | Val. PPL: 108317.404
Epoch: 64 | Time: 0m 19s
Train Loss: 0.489 | Train PPL: 1.630
Val. Loss: 11.735 | Val. PPL: 124810.446
Epoch: 65 | Time: 0m 20s
Train Loss: 0.403 | Train PPL: 1.497
Val. Loss: 11.931 | Val. PPL: 151848.747
Epoch: 66 | Time: 0m 20s
Train Loss: 0.320 | Train PPL: 1.377
Val. Loss: 12.124 | Val. PPL: 184251.124
Epoch: 67 | Time: 0m 22s
Train Loss: 0.256 | Train PPL: 1.291
Val. Loss: 12.191 | Val. PPL: 196978.577
Epoch: 68 | Time: 0m 20s
Train Loss: 0.200 | Train PPL: 1.221
Val. Loss: 12.323 | Val. PPL: 224785.248
Epoch: 69 | Time: 0m 19s
Train Loss: 0.158 | Train PPL: 1.171
Val. Loss: 12.518 | Val. PPL: 273131.276
Epoch: 70 | Time: 0m 20s
Train Loss: 0.132 | Train PPL: 1.142
Val. Loss: 12.505 | Val. PPL: 269803.527
Epoch: 71 | Time: 0m 19s
Train Loss: 0.104 | Train PPL: 1.110
Val. Loss: 12.571 | Val. PPL: 288023.389
Epoch: 72 | Time: 0m 19s
Train Loss: 0.088 | Train PPL: 1.092
Val. Loss: 12.607 | Val. PPL: 298722.474
Epoch: 73 | Time: 0m 20s
Train Loss: 0.076 | Train PPL: 1.079
Val. Loss: 12.644 | Val. PPL: 309939.272
Epoch: 74 | Time: 0m 20s
Train Loss: 0.065 | Train PPL: 1.067
Val. Loss: 12.614 | Val. PPL: 300803.358
Epoch: 75 | Time: 0m 20s
Train Loss: 0.056 | Train PPL: 1.057

```
	 Val. Loss: 12.633 |  Val. PPL: 306454.307
Epoch: 76 | Time: 0m 20s
	Train Loss: 0.049 | Train PPL:   1.050
	 Val. Loss: 12.661 |  Val. PPL: 315264.167
Epoch: 77 | Time: 0m 20s
	Train Loss: 0.042 | Train PPL:   1.043
	 Val. Loss: 12.755 |  Val. PPL: 346228.817
Epoch: 78 | Time: 0m 20s
	Train Loss: 0.040 | Train PPL:   1.040
	 Val. Loss: 12.820 |  Val. PPL: 369622.377
Epoch: 79 | Time: 0m 19s
	Train Loss: 0.037 | Train PPL:   1.037
	 Val. Loss: 12.902 |  Val. PPL: 401012.430
Epoch: 80 | Time: 0m 20s
	Train Loss: 0.032 | Train PPL:   1.032
	 Val. Loss: 12.942 |  Val. PPL: 417584.677
Epoch: 81 | Time: 0m 20s
	Train Loss: 0.028 | Train PPL:   1.029
	 Val. Loss: 13.025 |  Val. PPL: 453696.899
Epoch: 82 | Time: 0m 20s
	Train Loss: 0.027 | Train PPL:   1.027
	 Val. Loss: 13.015 |  Val. PPL: 449224.848
Epoch: 83 | Time: 0m 20s
	Train Loss: 0.026 | Train PPL:   1.026
	 Val. Loss: 13.008 |  Val. PPL: 445840.337
Epoch: 84 | Time: 0m 20s
	Train Loss: 0.025 | Train PPL:   1.026
	 Val. Loss: 13.103 |  Val. PPL: 490249.963
Epoch: 85 | Time: 0m 20s
	Train Loss: 0.024 | Train PPL:   1.024
	 Val. Loss: 13.040 |  Val. PPL: 460651.325
Epoch: 86 | Time: 0m 20s
	Train Loss: 0.021 | Train PPL:   1.022
	 Val. Loss: 13.108 |  Val. PPL: 492795.284
Epoch: 87 | Time: 0m 20s
	Train Loss: 0.021 | Train PPL:   1.021
	 Val. Loss: 13.101 |  Val. PPL: 489629.465
Epoch: 88 | Time: 0m 20s
	Train Loss: 0.020 | Train PPL:   1.020
	 Val. Loss: 13.162 |  Val. PPL: 520311.426
```

Epoch: 89 | Time: 0m 20s
Train Loss: 0.018 | Train PPL: 1.018
Val. Loss: 13.169 | Val. PPL: 524100.802
Epoch: 90 | Time: 0m 20s
Train Loss: 0.020 | Train PPL: 1.020
Val. Loss: 13.257 | Val. PPL: 572303.289
Epoch: 91 | Time: 0m 20s
Train Loss: 0.026 | Train PPL: 1.026
Val. Loss: 13.245 | Val. PPL: 565161.791
Epoch: 92 | Time: 0m 20s
Train Loss: 0.027 | Train PPL: 1.027
Val. Loss: 13.307 | Val. PPL: 601090.290
Epoch: 93 | Time: 0m 20s
Train Loss: 0.024 | Train PPL: 1.024
Val. Loss: 13.313 | Val. PPL: 605018.905
Epoch: 94 | Time: 0m 20s
Train Loss: 0.020 | Train PPL: 1.020
Val. Loss: 13.311 | Val. PPL: 603946.653
Epoch: 95 | Time: 0m 20s
Train Loss: 0.018 | Train PPL: 1.019
Val. Loss: 13.204 | Val. PPL: 542385.613
Epoch: 96 | Time: 0m 19s
Train Loss: 0.018 | Train PPL: 1.018
Val. Loss: 13.235 | Val. PPL: 559412.823
Epoch: 97 | Time: 0m 20s
Train Loss: 0.016 | Train PPL: 1.016
Val. Loss: 13.400 | Val. PPL: 659957.656
Epoch: 98 | Time: 0m 19s
Train Loss: 0.014 | Train PPL: 1.015
Val. Loss: 13.400 | Val. PPL: 659926.817
Epoch: 99 | Time: 0m 20s
Train Loss: 0.013 | Train PPL: 1.013
Val. Loss: 13.419 | Val. PPL: 672415.380
Epoch: 100 | Time: 0m 20s
Train Loss: 0.013 | Train PPL: 1.013
Val. Loss: 13.499 | Val. PPL: 728614.754
Epoch: 101 | Time: 0m 20s
Train Loss: 0.012 | Train PPL: 1.012
Val. Loss: 13.541 | Val. PPL: 759786.282
Epoch: 102 | Time: 0m 20s

Train Loss: 0.012 | Train PPL: 1.012
Val. Loss: 13.557 | Val. PPL: 771818.154
Epoch: 103 | Time: 0m 20s
Train Loss: 0.011 | Train PPL: 1.011
Val. Loss: 13.528 | Val. PPL: 749810.995
Epoch: 104 | Time: 0m 20s
Train Loss: 0.011 | Train PPL: 1.011
Val. Loss: 13.582 | Val. PPL: 791904.621
Epoch: 105 | Time: 0m 20s
Train Loss: 0.013 | Train PPL: 1.013
Val. Loss: 13.622 | Val. PPL: 824232.263
Epoch: 106 | Time: 0m 20s
Train Loss: 0.016 | Train PPL: 1.016
Val. Loss: 13.478 | Val. PPL: 713606.145
Epoch: 107 | Time: 0m 20s
Train Loss: 0.013 | Train PPL: 1.013
Val. Loss: 13.629 | Val. PPL: 830106.015
Epoch: 108 | Time: 0m 20s
Train Loss: 0.014 | Train PPL: 1.014
Val. Loss: 13.453 | Val. PPL: 696064.411
Epoch: 109 | Time: 0m 20s
Train Loss: 0.015 | Train PPL: 1.015
Val. Loss: 13.508 | Val. PPL: 735132.346
Epoch: 110 | Time: 0m 20s
Train Loss: 0.038 | Train PPL: 1.038
Val. Loss: 13.462 | Val. PPL: 702380.587
Epoch: 111 | Time: 0m 20s
Train Loss: 0.052 | Train PPL: 1.054
Val. Loss: 13.393 | Val. PPL: 655492.242

Input query to the model = ['when', 'did', 'beyonce', 'start', 'becoming', 'popular']
Actual answer or ground truth = ['in', 'the', 'late', '1990s']

The model generated output = “the late 1950s and 1960s EOS EOS”

where EOS refers to the end of sentence tokens, but since model runs on the heuristics and this kind a behavior can be expected. Therefore, thinking of further limiting the generated output from the model, so that it makes more sense. As for future steps it should be able to generate variable length output. However, training process and used hyperparameters have to be factored in as there is a lot of room for experimentation to improve the model further.

Implementation

The implementation involves creation of vocab object for questions or source and for answers or target. These vocab objects provide easy way to convert word to integers or indexes and back to actual words or strings. This is important because model only processes numbers or words in integer form. Where these integers also in way represent one hot encoding vectors. The integers/words are passed through the embedding layer and every integer or word corresponds to embedding layer or embedding weight matrix(under the hood) row, and we get a vector representation for that integer/word. These vector representations then pass through the LSTM layer in the encoder, the LSTM layer have hidden and cell states which are initially tensors of zeros,and after every sequence processing the new hidden and cell state become a input as well for the next sequence or word. We don't care about the generated output of the encoder.However, the last hidden and cell state after processing a input sentence or query completely is sent to the decoder as initial its hidden and cell state, and serves the purpose of context vector(query representation). The decoder also consists of embedding layer, LSTM layer, linear layer and softmax. The first input to the decoder is the <sos> token, which passes through the embedding layer to the LSTM layer to the linear layer and we get <sos> token mapped to predictions against the target vocab size, and select the top1 integer(most likely to be the next word) using output.argmax(1). Then in training we use teacher_forcing with ratio of 0.5 which acts as probability so for the next iteration in the decoder we sometimes use the predicted token as input and sometimes use the actual ground truth or target word/integer. In this way we generate the whole output from the decoder, and when calculating the loss we get rid of both 0th index element in both generated output tensor and target tensor, and calculate the loss. The model is trained on batches of data, where each batch refers to a sentence or question and same process applies to the answer or target as well. The creation of batches process was heavily tested to make sure that getting right amount of batches and that both src and trg have <sos> and <eos> tokens and even padding or <pad> token so that the batch size is consistent through all src batches and trg batches. The batches length or first dimension(row) for both src and trg vary as depends on the max length sentence.