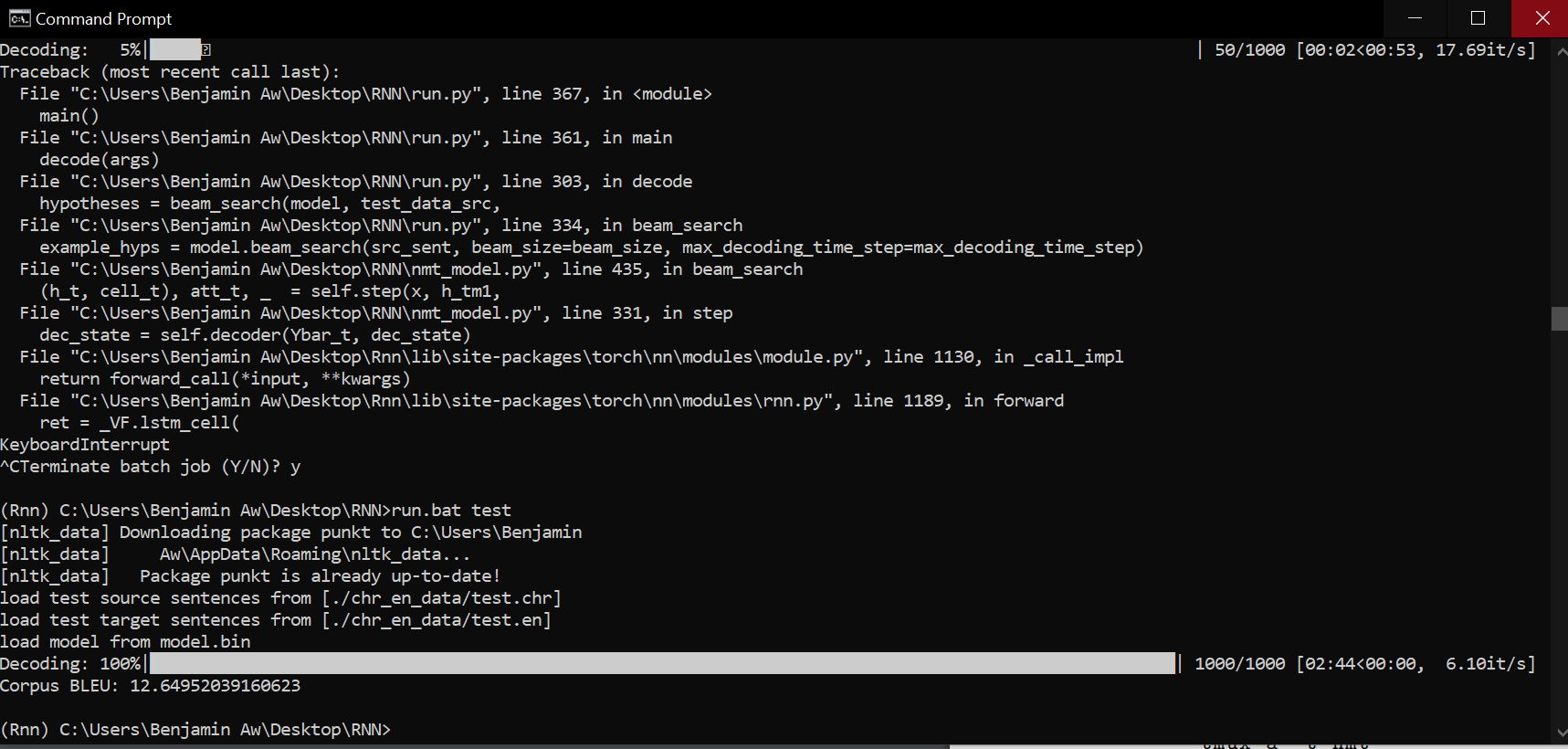
g) If the word is masked, the attention score will be set to -inf. Under the softmax function, the corresponding weight would be set to 0. This is because pad tokens should not contribute to any infomation wrt to the computation of attention, and therefore this method of computation is necessary.

h)



i)

1. Dot product is much more efficient to compute as compared to the additictive

attention mechanisms. Hidden state vectors that are orthogonal to the target hidden

vector will also be removed.

1. Multiplicative is much more efficient to compute as compared to the additictive

attention mechanisms. It also has a learnable parameter W, which allows the NN to

transform the sorurce hidden state which results in a lower loss function.

1. Additive attention mechanism is much more verbose and flexible as compared to the

two other attention mechanisms as there are much more trainable parameters to consider.