Section 1

g)

The input sentences are padded before sending it into the encoder, such that they have fixed equal lengths. They are padded with a token that means nothing. The masks essentially mark these padded locations with zero.
The step function now uses this mask to set the attention vector e_t to -inf at these padded locations. This is done so that after the softmax, the attention scores would be 0 at these locations in the α_t calculation.

h)

Corpus BLEU: 12.117

i)

Dot-product attention:

Advantages:

- Fast to compute
- No extra parameters

Disadvantages:

- We might not want to use the whole hidden states to calculate attention scores, but that cannot be controlled
- Both the hidden states need to be of the same size

Multiplicative attention:

Advantages:

- More flexible compared to dot product attention. Can now ignore parts of the hidden states to calculate attention

Disadvantages:

- Too many parameters, grows quadratically with hidden state dimension

Additive attention:

Advantages:

- In high dimensions of hidden states, its more efficient to compute compared to multiplicative attention

Disadvantages:

- Slow to compute
- Extra parameters

Section 2

f)

- i.

  For c_1 with r_1
  $p_1 = 3/5$
  $p_2 = 2/4$
  $len(c_1) = 5, len(r_1) = 6$, Hence $BP = exp(1 - 6/5) = 0.819$
  $\text{BLEU score } = BP * exp(\lambda_1 ln(p1) + \lambda_2 ln(p2)) = 0.448$

  For c_2 with r_2
  $p_1 = 3/5$
  $p_2 = 1/4$
  $len(c_2) = 5, len(r_2) = 4$, Hence $BP = 1$
  $\text{BLEU score } = BP * exp(\lambda_1 ln(p1) + \lambda_2 ln(p2)) = 0.387$

  It seems like c_1 is better, but the score of c_2 is low due to makes not corresponding with make.
  Word stemming might make the BLEU score look better

- ii.

  For c_1 with r_1
  $\text{Same as above - } 0.448$

  For c_2 with r_1
  $p_1 = 2/5$
  $p_2 = 1/4$
  $len(c_2) = 5, len(r_1) = 6$, Hence $BP = exp(1 - 6/5) = 0.819$
  $\text{BLEU score } = BP * exp(\lambda_1 ln(p1) + \lambda_2 ln(p2)) = 0.259$

  No, I do not agree with the BLEU score inference. Although c_1 has a higher score due to similar use of words, c_2 seems closer to the reference phrase

- iii.
  Since BLEU scores rely heavily on n-gram matches, translations that are better semantically, might end up receiving lower scores due to different use of words.

- iv.
  Advantages:
  Faster than humans, as we have a quantitative approach
  Less subjective compared to a human. A sentence will receive the same score is evaluated again

  Disadvantages:
  Heavily relies on n-gram matches. No measure of quality of translation
  Cannot measure semantics, or long dependencies.