(a)

(b)

(c)

(d)

(e)

(f)

(g) In attention computation, *enc_masks* is used to remove attention from padded tokens. If we do in this way, the attention from padded tokens will attend in the prediction, which disobeys intuition and may cause unknown affects.

(h) The base model has 28.12 avg loss on trainset, 17.5 ppl on devset and 18.52 BLEU score on testset.

(i) *Dot product* is more computationally easy to compute, but it doesn't allow different sizes of s_t and h_i. *Multiplicative attention* allows s_t and h_i in different lengths, but it requires more computations. *Additive attention* has richer representation capability, but it requires even more computations.

1. (i) *Aqui* is translated into *So* and the grammar is incorrect. Possible fix is to include a loss related to grammar to avoid the mistake.
   (ii) Alignment error. Possible fix is to adopt better attention mechanism.
   (iii) Out of vocabulary error. Possible fix to extend the vocabulary.
   (iv) Grammar alignment error. Possible fix to combined fixes in i and ii.
   (v) Direct translation error. Possible fix to extend the dataset to include this kind of sentences.
   (vi) Number translation error. Possible fix to extend the dataset to include this kind of sentences.

2. (a) por 3 aos. For three years. She did it for three years. The context is ignored here and possible fix is to integrate context information in NMT.
   (b) Ella salv mi vida, mi pareja y yo salvamos la de ella. She saved my life, my partner and I took out of it. She saved my life; I and my partner saved hers. Kinda one to many error. Possible fix is to extend the dataset to include this kind of sentences.

3. (i) $BP_1 = exp(1 - \frac{6}{5}) = exp(-\frac{1}{5})$, $BP_2 = 1$, since c_2 is longer than r_2;
   $BLEU_1 = exp(-\frac{1}{5} + 0.5log\frac{3}{5} + 0.5log\frac{2}{4}) = 0.45$
   $BLEU_2 = exp(0.5log\frac{5}{5} + 0.5\frac{2}{4}) = 0.71$.
   The second one is better. I also agree with it.
   (ii) $BLEU_1 = exp(-\frac{1}{5} + 0.5log\frac{3}{5} + 0.5log\frac{2}{4}) = 0.45$
   and $BLEU_2 = exp(0.5log\frac{2}{5} + 0.5log\frac{1}{4}) = 0.31$.
   Now the first one seems to be better but I think the second one should be better.
   (iii) It will introduce the bias into BLEU metric since we don't consider other possible sentences as references.
   (iv) BLEU is more objective than human and tends to be a fair metric when we have multiple reference sentences. However, when the data availability is poor, BLEU is less fair than human evaluation. Also, higher BLEU score can't 100% guaranttee a better translation since the N-gram computation has some issues (e.g., doesn't consider relative position).