(1)

(g)

The masks are designed to put -infinity at all the placeholder values. By putting -infinity values, you are essentially setting the amount of attention to pay 0. So, you aren't going to pay any attention to the placeholder tokens.

(i)

22.9

(j)

$$e_{t,i} = v^T(W_1 h_i + W_2 s_t)$$
$$e_{t,i} = s_t^T W h_i$$
$$e_{t,i} = s_t^T h_i$$

For dot product attention, it makes the most intuitive sense. Sense $e_{t,i}$ is going to be a score of how close the state right now is versus the state in the encoder. A possible disadvantage is that the hidden states in RNN may not correspond to the correct thing.

Multiplicative attention seems to account for the fact that these hidden states come from different RNN's so similar meanings could be mapped to different places. So a weight matrix is used to cover that gap. A possible disadvantage is there may be too much information to transfer from the hidden states.

Additive attention has the most amount of weights, so it allows the model to tune the parameters the best. A possible disadvantage is that there may be overfitting from too many parameters.

(2)

(a)

(i)

one of my doesn't have a lot of meaning in it. Since favorite is a really big sentiment in the sentence. A possible solution to this is to have a normalization, so that no sentiment becomes too big.

(ii)

It links the probably with author. It also links America and most and author. For the general phrase more reading in the US. The context may be too short, and a solution could be to update the RNN cells to keep track of more things.

(iii)

the unknown word token is put instead of the actual unknown word. A possible solution for this scenario is to put the most likely unknown word in the generated sentence.

(iv)

Manzana means both apple and block in spanish. A possible way to circumvent this problem is to use polymorphic embeddings.

(v)

I think there is some implicit bias in the text that teacher are women. So it puts women instead of teachers. A possible solution is to realign the text so that there is no bias.

(vi)

100,000 hectacres is a different measurment than acres. A possible solution is to manually input the different units.

(b)

Ella salv mi vida, mi pareja y yo salvamos la de ella.

She saved my life; I and my partner saved hers.

She saved my life, my partner and I save the one of her.

The problem is that saved hers becomes save the one of her. This is a model limitation. One problem could because of the context, it switches from past tense to present without much thought. Another aspect is that la de ella doesn't become hers, and instead is the word for word translation. More training data would help this.

Bueno, tuve mucho tiempo para pensar en esas 8 horas sin dormir.

Well, I had a lot of time to think during those eight hours and no sleep.

Well, I had a lot of time to think about those eight hours without sleep.

It mistakes during and about. This drastically changes the meaning of the sentences. There is no "about" in the spanish sentence. So it goes to the next most likely word. A possible solution to this problem is to make modifications to the decoder LSTM to make the sentences it spit out have more coherent sense.

(c)

(i)

BLEU score for c_1:0.448 $= e^{1-6/5}e^{0.5\ln 0.6+0.5\ln 0.5}$

BLEU score for c_2:0.632 $= e^{0.5\ln 0.8+0.5\ln 0.5}$

I think that the second one is a better translation.

(ii)

BLEU score for c_1:0.448 $= e^{1-6/5}e^{0.5\ln 0.6+0.5\ln 0.5}$

BLEU score for c_2:0.223 $= e^{1-6/5}e^{0.5\ln 0.4+0.5\ln 0.25}$

(iii)

Just because it doesn't use the same words, doesn't mean it's a bad translation.

(iv)

pros: lack of bias, replicatable

cons: too much emphasis on individual words, doesn't take context into consideration