Assignment 4

1. (g)
The mask changes the scores of paddings to -Inf, which changes their softmax probabilities to 0. Thus, the contribution of paddings to the resulting attention vector would be negligible.

(j)
Dot product attention is the most efficient one, but it's too simple that we might fail to capture some patterns using it. The dimension of s_t and h_i must be the same in this case. Multiplicative attention allows us to use s_t and h_i of different dimensions, and it's more flexible than dot product attention since it has a trainable parameter W. However, it's less efficient. Additive attention is the most complex one since it has 3 trainable variables. It can learn the most complex patterns with efficiency trade-off.

2. (a) (i)
Error: favorite of my favorites
Possible reason: the model does not learn one of... phrase well
Possible solution: try to add more training data with one of... in there

(ii)
Error: more reading in the U.S.
Possible reason: sentence structure is different from spanish to English
Possible solution: revise the translation text so phrases align better to the source sentence

(iii)
Error: <unk>
Possible reason: Bolingbroke is not in the vocabulary
Possible solution: Add it into the vocabulary

(iv)
Error: apple
Possible reason: A word might have meanings "apple" and "black", but the model picked the wrong one
Possible solution: Add more training data for this word

(v)
Error: women
Possible reason: gender bias. In the training data, teachers are more likely to be women
Possible solution: try to de-bias the embeddings or provide both genders at the same time in the translation

(vi)
Error: 100,000 acres
Possible reason: the model failed to do the math transforming 100,000 hectareas into 250 thousand acres
Possible solution: change "250 thousand acres" to "100,000 hectares" in the translation

(c) (i)
For c_1
$p_1 = (0+1+1+1+0)/5 = 0.6$
$p_2 = (0+1+1+0)/4 = 0.5$
$c = 5$
$r^* = 4$
$BP = 1$

BLEU = $\exp(0.5\log(0.6) + 0.5\log(0.5)) = 0.7699$

For C_2

$p_1 = (1+1+0+1+1)/5 = 0.8$

$p_2 = (1+0+0+1)/4 = 0.5$

$c = 5$

$r^* = 4$

BP = 1

BLEU = $\exp(0.5\log(0.8) + 0.5\log(0.5)) = 0.8196$

According to BLUE, C_2 is better. I agree it's better.

(ii)

For C_1

$p_1 = (0+1+1+1+0)/5 = 0.6$

$p_2 = (0+1+1+0)/4 = 0.5$

$c = 5$

$r^* = 6$

BP = $\exp(1 - 6/5) \approx 0.8187$

BLEU = $0.8187 \times \exp(0.5\log(0.6) + 0.5\log(0.5)) = 0.6304$

For C_2

$p_1 = (1+1+0+0+0)/5 = 0.4$

$p_2 = (1+0+0+0)/4 = 0.25$

$c = 5$

$r^* = 6$

BP = 0.8187

BLEU = $0.8187 \times \exp(0.5\log(0.4) + 0.5\log(0.25)) = 0.4966$

According to BLEU, C_1 is better. I do not agree.

(iii)

The BLEU score would favor those translations that are more similar to the reference in terms of words instead of meaning.

(iv)

Advantages:

- Way more efficient
- Can be easily integrated into models

Disadvantages:

- Suffer from bias if we don't have enough references
- Cannot control fluency