Assignment 4

1g) The generate_sent_masks() function in nmt_model.py produces a tensor called enc_masks. It has shape (batch_size, max_source_sentence_length) and contains 1s in position corresponding to ‘pad’ tokens in the input, and 0s for non-pad tokens. Look at how the masks are used during the attention computation in the step() function (lines 295-296). First explain (in around three sentences) what effects the masks have on the entire attention computation. Then explain (in one or two sentences) why it is necessary to use the masks in this way.

Ans) Masks make sure that the pad token is not included in the attention computation. They are used to map the pad token to a score of –inf before computing the attention vector. We do not want to include the pad token in the attention mechanism because it contains no relevant information; it just used to make sure that all the sentences have the same length.

i) Please report the model’s corpus BLEU score. It should be larger than 21.

Ans) BLEU score = 22.477

j) In class we learned about dot product attention, multiplicative attention and additive attention. Please provide one advantage and disadvantage of each attention mechanism with respect to either of the other two attention mechanisms. As a reminder, dot product attention is $e_{t,i} = s_t^T h_i$, multiplicative attention is $e_{t,i} = s_t^T W h_i$ and additive attention is $e_{t,i} = v^T(W_1 h_i + W_2 s_t)$.

Ans) Dot product and multiplicative attention mechanisms are faster than the additive attention. Additive attention performs better than multiplicative attention for large dimensions. Dot product attention is too simple and may not perform well.

2a) Here we present a series of errors we found in the outputs of our NMT model (which is the same as the one you just trained). For each example of a Spanish source sentence, reference (i.e. gold) English translation, and NMT (i.e. ‘model’) English translation, please:

1. Identify the error in the NMT translation.
2. Provide a reason why the model may have made the error (either due to a specific linguistic construct or specific model limitations).
3. Describe one possible way we might alter the NMT system to fix the observed error.

Below are the translations that you should analyse as described above. Note that out-of-vocabulary words are underlined.

i) Source Sentence: Aqu´ı otro de mis favoritos, “La noche estrellada”.

Reference Translation: So another one of my favorites, “The Starry Night”.

NMT Translation: Here’s another favorite of my favorites, “The Starry Night”.

Ans) The error is that instead of ‘one of my favorites’, the model outputted ‘favourite of my favorites’. The reason for this could be because the word ‘one’ doesn’t appear explicitly in the Spanish sentence. One possible way to fix this error is to use beam search in the decoder with an increase in the beam size.

ii) Source Sentence: Ustedes saben que lo que yo hago es escribir para los ni˜nos, y, de hecho, probablemente soy el autor para ni˜nos, ms ledo en los EEUU.

Reference Translation: You know, what I do is write for children, and I'm probably America's most widely read children's author, in fact.

NMT Translation: You know what I do is write for children, and in fact, I'm probably the author for children, more reading in the U.S

Ans) The error is that the output of the NMT model is not grammatically correct. The reason for this is that the model does a more word to word translation, rather than looking at the sentence or a part of the sentence and translating it. One possible way to fix this error is to pass the output to a language model to make the sentence grammatically correct.

iii) Source Sentence: Un amigo me hizo eso – Richard Bolingbroke.

Reference Translation: A friend of mine did that – Richard Bolingbroke.

NMT Translation: A friend of mine did that – Richard <unk>

Ans) The error is that output contains the unknown word token (<unk>). The reason for this is that the model has never seen the word Bollingbroke before, so it models it as an unknown word and hence outputs <unk>. One possible way to fix this error is to use a hybrid model i.e. a word level model combined with a character level model, so for an unknown word we can use the character level model to construct that word. Another way is to have some mechanism such that the model can copy a word from the source sentence whenever it needs to.

iv) Source Sentence: Solo tienes que dar vuelta a la manzana para verlo como una epifan´ıa.

Reference Translation: You've just got to go around the block to see it as an epiphany.

NMT Translation: You just have to go back to the apple to see it as a epiphany.

Ans) The error is that the NMT model output has the word 'apple' instead of 'block'. The reason for this is that the model fails to correctly translate the Spanish idiom; the model does the more literal translation and hence gives a wrong translation. One possible way to fix this is to train the model on more sentences with idioms.

v) Source Sentence: Ella salv´o mi vida al permitirme entrar al ba˜no de la sala de profesores.

Reference Translation: She saved my life by letting me go to the bathroom in the teachers' lounge.

NMT Translation: She saved my life by letting me go to the bathroom in the women's room.

Ans) The error is that the model output has the word "women's" instead of "teachers'". The reason for this is that the model sees the female context 'Ella' and therefore translates 'profesores' to 'women's', this suggests gender bias in the model. One possible way to fix this is to 'debias 'the model (word embeddings).

vi) Source Sentence: Eso es m´as de 100,000 hect´areas.

Reference Translation: That's more than 250 thousand acres.

NMT Translation: That's over 100,000 acres.

Ans) The error is that the model output has '100,000' instead of '250 thousand'. The reason for this is that the model is unable to convert from the unit 'hect'acreas' to 'acres'. One possible way to fix this is to give the model more training data that requires the model to perform unit conversions.

b) Now it is time to explore the outputs of the model that you have trained! The test-set translations that your model produced in 1-i should be located in outputs/test_outputs.txt. Please identify 2 examples of errors that your model produced. The two examples you find should be different error types from another and different error types than the examples provided in the previous question. For each example you should:

1. Write the source sentence in Spanish. The source sentences are in the en_es_data/test.es.
2. Write the reference English translation. The reference translations are in the en_es_data/test.en.
3. Write your NMT model's English translation. The model-translated sentences are in the outputs/test_outputs.txt.
4. Identify the error in the NMT translation.
5. Provide a reason why the model may have made the error (either due to a specific linguistic construct or specific model limitations).
6. Describe one possible way we might alter the NMT system to fix the observed error.

Ans) Source Sentence: El rosa es mi color favorito.

Reference Translation: Pink is my favorite color.

NMT Translation: The rose is my favorite color.

The error the model makes is that it translates 'rosa' to 'rose' instead of 'pink'. The reason for this is because 'rosa' has two meanings 'rose' as well as 'pink' but the model is not able to distinguish between words having more than one meaning, given a certain context, as in this case 'el rosa' means pink but 'la rosa' means rose. One possible way to fix this is to give the model more training data which contains words with multiple meanings.

Source Sentence: Tenemos esta lista de cosas para hacer antes de morir, estas cosas que queremos hacer en vida, y pens en toda la gente a las que quera llegar y no lo hice, todas las cercas que quera reparar, todas las experiencias que he querido tener y nunca tuve.

Reference Translation: We have this bucket list, we have these things we want to do in life, and I thought about all the people I wanted to reach out to that I didn't, all the fences I wanted to mend, all the experiences I wanted to have and I never did.

NMT Translation: We have this list of things to do before they die, these things that we want to do in life, and I thought about all the people I wanted to get and I did all the fences that I wanted to <unk> all the experiences I've wanted to have and never <unk>

The error the model makes is instead of using the idiom 'bucket list'; it uses the meaning of the idiom 'list of things of things to do before they die'. The reason for this is that the model is unable to directly translate Spanish phrases into commonly used English idioms. One possible way to fix this is to use more training data that contains such idioms, or pass the output of the NMT model into another model that specifically converts phrases into idioms.

c) i) Please consider this example:

Source Sentence s: el amor todo lo puede

Reference Translation r1: love can always find a way

Reference Translation r2: love makes anything possible

NMT Translation c1: the love can always do

NMT Translation c2: love can make anything possible

Please compute BLEU scores for c1 and c2. Let $\lambda_i = 0.5$ for $i \in \{1, 2\}$ and $\lambda_i = 0$ for $i \in \{3, 4\}$ (this means we ignore 3-grams and 4-grams, i.e., don't compute p_3 or p_4). When computing BLEU scores, show your working (i.e., show your computed values for p1, p2, c, r* and BP). Which of the two NMT translations is considered the better translation according to the BLEU Score? Do you agree that it is the better translation?

Ans) For c1:

$p_1 = 3/5 = 0.6$

$p_2 = 2/4 = 0.5$

c = 5

r* = 6

BP = 0.819

BLEU = 0.819 *0.548 = 0.449

For c2:

$p_1 = 3/5 = 0.6$

$p_2 = 2/4 = 0.5$

c = 5

r* = 4

BP = 1

BLEU = 1*0.548 = 0.548

According to BLEU score c2 is the better translation.

I think c2 is the better translation, therefore I agree with the BLEU score.

ii) Our hard drive was corrupted and we lost Reference Translation r2. Please recompute BLEU scores for c1 and c2, this time with respect to r1 only. Which of the two NMT translations now receives the higher BLEU score? Do you agree that it is the better translation?

Ans) For c1:

p_1 = 3/5 = 0.6

p_2 = 2/4 = 0.5

c = 5

r* = 6

BP = 0.819

BLEU = 0.819 *0.548 = 0.449

For c2:

p_1 = 2/5 = 0.4

p_2 = 1/4 = 0.25

c = 5

r* = 6

BP = 0.819

BLEU = 0.819*0.316 = 0.259

According to BLEU score c1 is the better translation.

It can clearly be seen that even though c_1 gets a higher BLEU score than c_2, c_2 is clearly the better translation.

iii) Due to data availability, NMT systems are often evaluated with respect to only a single reference translation. Please explain (in a few sentences) why this may be problematic.

Ans) This is problematic because many output translations that are not grammatically correct but have some matching words will get a higher BLEU score, just like the example above where the bad translation got a higher BLEU score.

iv) List two advantages and two disadvantages of BLEU, compared to human evaluation, as an evaluation metric for Machine Translation.

Ans) Two advantages of BLEU are that it is fast and inexpensive as compared to human evaluation. Two disadvantages are that it is sometimes unreliable and compares the machine generated translation only against a few human translations.