(g) (3 points) (written) 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 positions 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 311-312).

First explain (in around three sentences) what effect 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.

Solution:

(g) The step function takes the multiplicative attention and replaces the masked data with minus infinity. By pushing this value throw a softmax we get a 0 so they won't affect the attention score.involving the padded tokens will result in false attention representation.

Please report the model's corpus BLEU Score. It should be larger than 18.

Solution: 19.394

(i) (4 points) (written) In class, we learned about dot product attention, multiplicative attention, and additive attention. As a reminder, dot product attention is $\mathbf{e}_{t,i} = \mathbf{s}_t^T \mathbf{h}_i$, multiplicative attention is $\mathbf{e}_{t,i} = \mathbf{s}_t^T \mathbf{W} \mathbf{h}_i$, and additive attention is $\mathbf{e}_{t,i} = \mathbf{v}^T \tanh(\mathbf{W}_1 \mathbf{h}_i + \mathbf{W}_2 \mathbf{s}_t)$.

i. (2 points) Explain one advantage and one disadvantage of *dot product attention* compared to multiplicative attention.

ii. (2 points) Explain one advantage and one disadvantage of *additive attention* compared to multiplicative attention.

Solution:

(i) i. One advantage of dot product is the computing resources - no need learning and storing parameters. One disadvantage is the fact that it can't decide what parts it need to pay attention to because it's a simple piece wise similarity.

ii. One advantage of additive attention both h and s has their own learnable weights. One disadvantage is that computation is more expansive.

2. Analyzing NMT Systems (25 points)

(a) (3 points) Look at the `src.vocab` file for some examples of phrases and words in the source language vocabulary. When encoding an input Mandarin Chinese sequence into "pieces" in the vocabulary, the tokenizer maps the sequence to a series of vocabulary items, each consisting of one or more characters (thanks to the `sentencepiece` tokenizer, we can perform this segmentation even when the original text has no white space). Given this information, how could adding a 1D Convolutional layer after the embedding layer and before passing the embeddings into the bidirectional encoder help our NMT system? **Hint:** each Mandarin Chinese character is either an entire word or a morpheme in a word. Look up the meanings of 电, 脑, and 电脑 separately for an example. The characters 电 (electricity) and 脑 (brain) when combined into the phrase 电脑 mean computer.

Solution:

(a) Using a convolutional layer after the embedded layer could help recognize if the word is a word or a morpheme in a word so it will have a better representation. It could find the "computer" mandarine words as one representation for the encoder instead of two different words.

(b) (8 points) 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 reference (i.e., 'gold') English translation, and NMT (i.e., 'model') English translation, please:

1. Identify the error in the NMT translation.
2. Provide possible reason(s) why the model may have made the error (either due to a specific linguistic construct or a specific model limitation).
3. Describe one possible way we might alter the NMT system to fix the observed error. There are more than one possible fixes for an error. For example, it could be tweaking the size of the hidden layers or changing the attention mechanism.

Below are the translations that you should analyze as described above. Only analyze the underlined error in each sentence. Rest assured that you don't need to know Mandarin to answer these questions. You just need to know English! If, however, you would like some additional color on the source sentences, feel free to use a resource like https://www.archchinese.com/chinese_english_dictionary.html to look up words. Feel free to search the training data file to have a better sense of how often certain characters occur.

i. (2 points) **Source Sentence:** 贼人其后被警方拘捕及被判处盗窃罪名成立。
Reference Translation: *the culprits were subsequently arrested and convicted.*
NMT Translation: *the culprit was subsequently arrested and sentenced to theft.*

Solution:

i.

1. **The error is** that the word culprit is in the singular form instead of its plural form.
2. **Why**? The source language doesn't contain plural forms

3. **How to improve**? Add more training data that has plural nouns.

ii. (2 points) **Source Sentence**: 几乎已经没有地方容纳这些人，资源已经用尽。
Reference Translation: *there is almost no space to accommodate these people, and resources have run out.*
NMT Translation: *the resources have been exhausted and resources have been exhausted.*

Solution:

ii.

1. **The error is** that the phrase the resources have been exhausted is repeated.

2. **Why**? Seems like there is an error in the attention mechanism it gave all its attention to the second part of the sentence

3. **How to improve**? We can add penalty to repeated translation or we could change the attention mechanism to multi head attention

iii. (2 points) **Source Sentence**: 当局已经宣布今天是国殇日。
Reference Translation: *authorities have announced a national mourning today.*
NMT Translation: *the administration has announced today's day.*

Solution:

iii.

1. **The error is** that the NMT translate the "national mourning" to "today"

2. **Why**? The model failed to recognize the noun phrase.

3. **How to improve**? Add more training data with noun phrases.

iv. (2 points) **Source Sentence**[4]**:** 俗语有云:"唔做唔错"。
Reference Translation: *" act not, err not ", so a saying goes.*
NMT Translation: *as the saying goes, " it's not wrong. "*

Solution:

iv.

1. **The error is** that the model didn't recognize the Chinese idiom.

2. **Why**? This idiom is not common in the training data. Not a single appearance in the training data.

3. **How to improve**? Provide the model training data that has more idioms in the source language.

i. (5 points) Please consider this example:
Source Sentence **s**: **需要有充足和可预测的资源。**
Reference Translation $\mathbf{r}_1$: *resources have to be sufficient and they have to be predictable*
Reference Translation $\mathbf{r}_2$: *adequate and predictable resources are required*
NMT Translation $\mathbf{c}_1$: there is a need for adequate and predictable resources
NMT Translation $\mathbf{c}_2$: resources be sufficient and predictable to
Please compute the BLEU scores for $\mathbf{c}_1$ and $\mathbf{c}_2$. 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 work (i.e., show your computed values for p_1, p_2, $len(c)$, $len(r)$ and BP). Note that the BLEU scores can be expressed between 0 and 1 or between 0 and 100. The code is using the 0 to 100 scale while in this question we are using the **0 to 1** scale. Please round your responses to 3 decimal places.

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?

Solution:

$$p_n = \frac{\sum_{\text{ngram}\in\mathbf{c}} \min\left(\max_{i=1,\ldots,k} \text{Count}_{\mathbf{r}_i}(\text{ngram}),\ \text{Count}_{\mathbf{c}}(\text{ngram})\right)}{\sum_{\text{ngram}\in\mathbf{c}} \text{Count}_{\mathbf{c}}(\text{ngram})}$$

$$BP = \begin{cases} 1 & \text{if } len(c) \geq len(r) \\ \exp\left(1 - \frac{len(r)}{len(c)}\right) & \text{otherwise} \end{cases}$$

$$BLEU = BP \times \exp\left(\sum_{n=1}^{4} \lambda_n \log p_n\right)$$

c1

1-gram	r1	r2	2-gram	r1	r2
there	0	0	there is	0	0
is	0	0	is a	0	0
a	0	0	a need	0	0
need	0	0	need for	0	0
for	0	0	for adequte	0	0
adequte	0	1	adequte and	0	1
and	1	1	and predictable	0	1
predictable	1	1	predictable resources	0	1

resources	1	1

$$P_1 = \frac{0.5+1.4}{9} = \frac{4}{9} \qquad P_2 = \frac{0.5+1.3}{8} = \frac{3}{8}$$

$$BP = e^{(1-\frac{11}{9})}$$

$$BLEU_{c_1} = e^{-\frac{2}{9}} \cdot e^{0.5\cdot(\log\frac{4}{9}+\log\frac{3}{8})} = 0.327$$

C2

1-gram	r1	r2	2-gram	r1	r2
resources	1	1	resources be	0	0
be	2	0	be sufficient	1	0
sufficient	1	0	sufficient and	1	0
and	1	1	and predictable	0	1
predictable	1	1	predictable to	0	0
to	2	0			

$$P_1 = \frac{6}{6} = 1 \qquad P_2 = \frac{3}{5}$$

$$BP = 1$$

$$BLEU_{c_2} = 1 \cdot e^{0.5(\log 1+\log\frac{3}{5})} = 0.775$$

The BLEU score for c2 is much higher but the translation is not grammatically correct and c1 is much more accurate.

ii. (5 points) Our hard drive was corrupted and we lost Reference Translation $\mathbf{r}_1$. Please recompute BLEU scores for $\mathbf{c}_1$ and $\mathbf{c}_2$, this time with respect to $\mathbf{r}_2$ only. Which of the two NMT translations now receives the higher BLEU score? Do you agree that it is the better translation?

Solution:

C_1: $P_1 = \frac{4}{9}$ $P_2 = \frac{3}{8}$ $BP = 1$ $BLEU_{C_1} = 0.408$

C_2: $P_1 = \frac{1}{2}$ $P_2 = \frac{1}{5}$ $BP = 1$ $BLEU_{C_2} = 0.316$

ii. C1 receives the better BLEU score and I agree it's the better translation

iii. (2 points) 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. In your explanation, discuss how the BLEU score metric assesses the quality of NMT translations when there are multiple reference transitions versus a single reference translation.

Solution:

iii. There could be a diversity between the different translations for the same sentence therefore there is a need to compare the translated sentence to different versions of the target translation because modified n-gram precision is dependent on the maximum of times it appears in any one of the versions, also the brevity penalty takes the closest length from the different versions.

iv. (2 points) List two advantages and two disadvantages of BLEU, compared to human evaluation, as an evaluation metric for Machine Translation.

Solution:

iv. Advantages

1. Automatic, fast and cheap.
2. Independent within the language- can be used for any pair.

Disadvantage

1. Needs multiple translations which not always available
2. Doesn't take word variants and position into account.