Assignment 4 Written Solutions

1. Neural Machine Translation with RNNs (45 points):

(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.

Instead of calculating gradients for the whole matrix (batch size by max source sentence length) which include the pad tokens, we calculate the gradients for the non-pad tokens only. This saves a lot of computation time since the batch usually contains one large sentence and many small sentences. This is done by setting the attention for pad tokens to $-\infty$. When applying softmax() to the pad tokens, we simply calculate $\frac{e^{-\infty}}{\sum_{t=1}^{sentece\_max} e^{t}} = 0$ so the pad tokens will be 0.

This is necessary because pads exist in almost all batches with huge counts. If the pads gradients are considered, the model will learn to produce only pads since they will have the highest probability of all words.

(h) (3 points) (written) Once your model is done training (**this should take under 2 hours on the VM**), execute the following command to test the model:

```
sh run.sh test
(Windows) run.bat test
```

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

Corpus BLEU: **20.242372402655874**

(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.

(i) **Advantage:** Dot product is simple and efficient while multiplicative attention has a huge weight matrix that needs time and memory to compute and learn.

Disadvantage: Dot product can be misleading. The hidden states have their own memory storing information about various things like which words appeared before which words, the context of the words it has already seen, etc. Dot product does not differentiate between this information the way multiplicative attention with its learnable parameters can. In fact, multiplicative attention proved to be more successful.

(ii) **Advantage:** Additive attention has more parameters than multiplicative attention and is more general.

Disadvantage: Additive attention is inefficient and complex while multiplicative attention is simple.

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.

The convolutional layer acts as a feature extractor. It collects important information of the given Chinese vectors. It will learn to differentiate if the character is a separate word or a morpheme in a word.

(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.*

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.*

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.*

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. "*

(i)

1. Using singular noun "culprit" instead of a plural noun "culprits".
2. Maybe the data had too many singular nouns that our model tends to produce them more often than plural nouns.
3. Introduce more sentence with plural nouns in the data and train the model again.

(ii)

1. An entire sentence "there is almost no space to accommodate these people" has not been translated.

2. The model may have failed to understand the concept of the Chinese words “容纳” (which means to accommodate) and “人” (which means people). This may be due to a problem with the attention mechanism.
3. Adjust the model so that the model mechanism can differentiate between the given words. Or add more sentences that include both words in different contexts in the data.

(iii)

1. The word “national mourning day” was wrongly translated to “today’s day”.
2. The word “殇” (which means national mourning) is a very specific word and may have been not used enough in the corpus.
3. Add more sentences that include the word “殇” in the corpus.

(iv)

1. The word “act” is wrongly translated to “It’s”.
2. The corpus may be missing Chinese sentences with imperative tense.
3. Add more Chinese sentences that include imperative tense.

(c) (14 points) BLEU score is the most commonly used automatic evaluation metric for NMT systems. It is usually calculated across the entire test set, but here we will consider BLEU defined for a single example.[5] Suppose we have a source sentence $\mathbf{s}$, a set of k reference translations $\mathbf{r}_1, \ldots, \mathbf{r}_k$, and a candidate translation $\mathbf{c}$. To compute the BLEU score of $\mathbf{c}$, we first compute the *modified n-gram precision* p_n of $\mathbf{c}$, for each of $n = 1, 2, 3, 4$, where n is the n in n-gram:

$$p_n = \frac{\sum_{\text{ngram} \in \mathbf{c}} \min\left(\max_{i=1,\ldots,k} \text{Count}_{\mathbf{r}_i}(\text{ngram}),\ \text{Count}_{\mathbf{c}}(\text{ngram}) \right)}{\sum_{\text{ngram} \in \mathbf{c}} \text{Count}_{\mathbf{c}}(\text{ngram})} \tag{15}$$

Here, for each of the n-grams that appear in the candidate translation $\mathbf{c}$, we count the maximum number of times it appears in any one reference translation, capped by the number of times it appears in $\mathbf{c}$ (this is the numerator). We divide this by the number of n-grams in $\mathbf{c}$ (denominator).

Next, we compute the *brevity penalty* BP. Let $len(c)$ be the length of $\mathbf{c}$ and let $len(r)$ be the length of the reference translation that is closest to $len(c)$ (in the case of two equally-close reference translation lengths, choose $len(r)$ as the shorter one).

$$BP = \begin{cases} 1 & \text{if } len(c) \geq len(r) \\ \exp\left(1 - \frac{len(r)}{len(c)}\right) & \text{otherwise} \end{cases} \tag{16}$$

Lastly, the BLEU score for candidate $\mathbf{c}$ with respect to $\mathbf{r}_1, \ldots, \mathbf{r}_k$ is:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{4} \lambda_n \log p_n \right) \tag{17}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are weights that sum to 1. The log here is natural log.

i. (5 points) Please consider this example:
Source Sentence $\mathbf{s}$: **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?

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?

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.

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

(i) For c_1:

Unigrams	C1=Count(NMT)	C2=Count(r_1)	C3=Count(r_2)	C4=Max(C2,C3)	C5=Min(C1,C4)
There	1	0	0	0	0
is	1	0	0	0	0
a	1	0	0	0	0
need	1	0	0	0	0
for	1	0	0	0	0
adequate	1	0	1	1	1
and	1	1	1	1	1
predictable	1	1	1	1	1
resources	1	1	1	1	1
Σ	9	3	4	4	4

From the table, $p_1 = \frac{4}{9}$

Bigrams	C1=Count(NMT)	C2=Count(r_1)	C3=Count(r_2)	C4=Max(C2,C3)	C5=Min(C1,C4)
There is	1	0	0	0	0
is a	1	0	0	0	0
a need	1	0	0	0	0
need for	1	0	0	0	0
for adequate	1	0	0	0	0
adequate and	1	0	1	1	1
and predictable	1	0	1	1	1
predictable resources	1	0	1	1	1
Σ	8	0	3	3	3

From the table, $p_2 = \frac{3}{8}$

$len(c_1) = 9$

$len(r) = \text{argmin}(|len(c_1) - len(r_1)|, |len(c_1) - len(r_2)|)$

$= \text{argmin}(|9 - 11|, |9 - 6)) = \text{argmin}(2,3) = len(r_1) = 11 > len(c_1)$

$BP = \exp\left(1 - \frac{len(r)}{len(c_1)}\right) = \exp\left(1 - \frac{11}{9}\right) = \exp\left(\frac{-2}{9}\right) = 0.800737$

$$BLEU(c_1) = BP \times \exp\left(\sum_{n=1}^{4} \lambda_n \log(p_n)\right)$$

$$= BP \times \exp(\lambda_1 \log(p_1) + \lambda_2 \log(p_2) + \lambda_3 \log(p_3) + \lambda_4 \log(p_4))$$

$$= 0.800737 \times \exp\left(0.5 \log\left(\frac{4}{9}\right) + 0.5 \log\left(\frac{3}{8}\right)\right)$$

$\underline{\boldsymbol{BLEU(c_1) = 0.3269}}$

For c_2:

Unigrams	C1=Count(NMT)	C2=Count(r_1)	C3=Count(r_2)	C4=Max(C2,C3)	C5=Min(C1,C4)
resources	1	1	1	1	1
be	1	1	0	1	1
sufficient	1	1	0	1	1
and	1	1	1	1	1
predictable	1	1	1	1	1
to	1	2	0	2	1
Σ	6	7	3	4	6

From the table, $p_1 = \frac{6}{6} = 1$

Bigrams	C1=Count(NMT)	C2=Count(r_1)	C3=Count(r_2)	C4=Max(C2,C3)	C5=Min(C1,C4)
resources be	1	0	0	0	0
be sufficient	1	1	0	1	1
sufficient and	1	1	0	1	1
and predictable	1	0	1	1	1
predictable to	1	0	0	0	0
Σ	5	2	1	3	3

From the table, $p_2 = \frac{3}{5}$

$len(c_2) = 6$

$len(r) = \text{argmin}(|len(c_2) - len(r_1)|, |len(c_2) - len(r_2)|)$

$= \text{argmin}(|6 - 11|, |6 - 6)) = \text{argmin}(5,0) = len(r_2) = 6 = len(c_2)$

$BP = 1$

$$BLEU(c_2) = BP \times \exp\left(\sum_{n=1}^{4} \lambda_n \log(p_n)\right)$$

$$= BP \times \exp(\lambda_1 \log(p_1) + \lambda_2 \log(p_2) + \lambda_3 \log(p_3) + \lambda_4 \log(p_4))$$

$$= 1 \times \exp\left(0.5 \log(1) + 0.5 \log\left(\frac{3}{5}\right)\right)$$

$$\underline{\boldsymbol{BLEU(c_2) = 0.7746}}$$

According to the BLEU score, c_2 is a better translation than c_1. However, I totally disagree. The second translation is grammatically incorrect and doesn't make any sense while the first translation is way much better.

(ii) For c_1:

Unigrams	C1=Count(NMT)	C2=Count(r_2)	C3=Min(C1,C2)
There	1	0	0
is	1	0	0
a	1	0	0
need	1	0	0
for	1	0	0
adequate	1	1	1
and	1	1	1
predictable	1	1	1
resources	1	1	1
Σ	9	4	4

From the table, $p_1 = \frac{4}{9}$

Bigrams	C1=Count(NMT)	C2=Count(r_2)	C5=Min(C1,C2)
There is	1	0	0
is a	1	0	0
a need	1	0	0
need for	1	0	0
for adequate	1	0	0
adequate and	1	1	1
and predictable	1	1	1
predictable resources	1	1	1
Σ	8	3	3

From the table, $p_2 = \frac{3}{8}$

$len(c_1) = 9$

$len(r) = len(r_2) = 6 < len(c_1)$

$BP = 1$

$$BLEU(c_1) = BP \times \exp\left(\sum_{n=1}^{4} \lambda_n \log(p_n)\right)$$

$$= BP \times \exp(\lambda_1 \log(p_1) + \lambda_2 \log(p_2) + \lambda_3 \log(p_3) + \lambda_4 \log(p_4))$$

$$= 1 \times \exp\left(0.5 \log\left(\frac{4}{9}\right) + 0.5 \log\left(\frac{3}{8}\right)\right)$$

$\underline{\boldsymbol{BLEU(c_1) = 0.4082}}$

For c_2:

Unigrams	C1=Count(NMT)	C2=Count(r_2)	C5=Min(C1,C2)
resources	1	1	1
be	1	0	0
sufficient	1	0	0
and	1	1	1
predictable	1	1	1
to	1	0	0
Σ	6	3	3

From the table, $p_1 = \frac{3}{6}$

Bigrams	C1=Count(NMT)	C2=Count(r_2)	C5=Min(C1,C2)
resources be	1	0	0
be sufficient	1	0	0
sufficient and	1	0	0
and predictable	1	1	1
predictable to	1	0	0
Σ	5	1	1

From the table, $p_2 = \frac{1}{5}$

$len(c_2) = 6$

$len(r) = len(r_2) = 6 = len(c_2)$

$BP = 1$

$$BLEU(c_2) = BP \times \exp\left(\sum_{n=1}^{4} \lambda_n \log(p_n)\right)$$

$$= BP \times \exp(\lambda_1 \log(p_1) + \lambda_2 \log(p_2) + \lambda_3 \log(p_3) + \lambda_4 \log(p_4))$$

$$= 1 \times \exp\left(0.5 \log\left(\frac{3}{6}\right) + 0.5 \log\left(\frac{1}{5}\right)\right)$$

$$\underline{\boldsymbol{BLEU(c_2) = 0.3162}}$$

According to the BLEU score, c_1 is a better translation than c_2. I agree with that

(iii) The process of translation is not a one-way road. The same sentence can be translated into many forms depending on the translator himself. The order of words does not change the meaning of the translated sentence. Consequently, NMT may translate the sentence into any of those forms. Having only one of these forms while NMT uses another one will make its BLEU score low. This makes learning hard and slow.

Having multiple reference sentences will decrease the hardness of learning. The translation process will be more robust and reliable as it takes into account the diversity and variation of many references.

(iv) **Advantages:**

1. Automation: It is an automated process that does not require human intervention.
2. Objectivity: It shows a number based on equations and calculations so there is no bias.

Disadvantages:

1. Many references: It needs a lot of reference translations for the same sentence to be reliable.
2. Quality: It depends on n-grams which are not very reliable in terms of evaluating translations.