CSC401 Homework Assignment #2
Analysis

Naihe Xiao
Student number: 1006765038
UTORid: xiaonaih
naihe.xiao@mail.utoronto.ca

1 Training Results

1.1 Training Loop Printout

Place your training loop printout in this section.

Model without Attention

1. When cell type is rnn

```
Epoch 1: loss=3.504261497526551, BLEU=0.2308623373987904
Epoch 2: loss=2.8779923957594526, BLEU=0.23931476956151007
Epoch 3: loss=2.6571673715460524, BLEU=0.24735937016611548
Epoch 4: loss=2.5294617084466275, BLEU=0.24519178486474055
Epoch 5: loss=2.449584542798974, BLEU=0.24512966888736532
Run summary: bleu 0.24513, loss 2.44958
```

2. When cell type is lstm

```
Epoch 1: loss=3.44823489624034, BLEU=0.23168984635255666
Epoch 2: loss=2.498243836982865, BLEU=0.2614386297272516
Epoch 3: loss=2.0408068237519936, BLEU=0.2770875587106813
Epoch 4: loss=1.6880922711208524, BLEU=0.2847163008520787
Epoch 5: loss=1.4136399071867813, BLEU=0.2912601501302457
Run summary: bleu 0.29126, loss 1.41364
```

Model with Single-headed Attention

1. When cell type is rnn

```
Epoch 1: loss=3.232511939787195, BLEU=0.2648892566360985
Epoch 2: loss=2.414827988093924, BLEU=0.27982245814906426
Epoch 3: loss=2.113291833365733, BLEU=0.2853473310679333
Epoch 4: loss=1.9296012991092986, BLEU=0.28624702011014086
Epoch 5: loss=1.8095226030863687, BLEU=0.2890986882136451
Run summary: bleu 0.2891, loss 1.80952
```

2. When cell type is lstm

```
Epoch 1: loss=3.2115278051834073, BLEU=0.27319541357178995
Epoch 2: loss=2.1507738430013528, BLEU=0.30485820791700813
Epoch 3: loss=1.6787538851312316, BLEU=0.3183487658666572
Epoch 4: loss=1.3438060935440836, BLEU=0.3235177544835014
Epoch 5: loss=1.0980041828478388, BLEU=0.3263011795049155
Run summary: bleu 0.3263, loss 1.098
```

Model with Multi-headed Attention

1. When cell type is rnn

```
Epoch 1: loss=3.17752643073006, BLEU=0.27037437712348034
Epoch 2: loss=2.419365602778378, BLEU=0.284831395984182
Epoch 3: loss=2.1827273346871507, BLEU=0.2887240764022031
Epoch 4: loss=2.048967465010177, BLEU=0.2909458981107966
Epoch 5: loss=1.9685361359537827, BLEU=0.2922327440035755
Run summary: bleu 0.29223, loss 1.96854
```

2. When cell type is lstm

```
Epoch 1: loss=3.101354992812485, BLEU=0.2830928876342354
Epoch 2: loss=2.096013803791747, BLEU=0.3139189149256519
Epoch 3: loss=1.6819307894708921, BLEU=0.3225688045515997
Epoch 4: loss=1.3947573561978088, BLEU=0.32698895217424023
Epoch 5: loss=1.1878681742118176, BLEU=0.3313305937367558
Run summary: bleu 0.33133, loss 1.18787
```

1.2 Test Set BLEU Score

This section lists the test set BLEU score reported on the test set for each model in table 1.

Model	Test BLEU
Model without Attention	0.29351921860505253
Model with Single-headed Attention	0.3431290494523297
Model with Multi-headed Attention	0.3487122079060109

Table 1: The BLEU score reported on the test set for each model.

1.3 Discussion

From the outputs above I see that the BLEU scores for the training procedure are 0.24513, 0.29126, 0.2891, 0.3263, 0.29223, 0.33133. Therefore the testing result for single and multi headed attention are in fact better than the training ones: they have BLEU score of 0.3431 and 0.3487, respectively. In general, the training and testing results are not very different, but the better performance on the testing might be the consequence of the smaller dataset.

In general, the models with attention have better performance than the models without, since allows the model to focus on the relevant parts of the input sequence as attention mechanism enables the model to focus on specific parts of the input sequence while generating the output sequence, hence the model can better capture the relationships between words in the input sequence and generate more accurate translations. In addition, the models with RNN as cell type have worse performances than the models with LSTM cell type, since LSTM models is better at handling vanishing or exploding gradients.

2 Translation Analysis

2.1 Translations

List all of your translations in this section.

1. Without attention

   (a) '<s> louis david riel university canada canadian'
   (b) '<s> it should be gavel to gavel filming on'
   (c) '<s> the students of guelph wellington said it lot of money'

2. With single-headed attention

   (a) '<s> city is in town city canada canada </s>'
   (b) '<s> the serbs ought to be treated rather than reactive'
   (c) '<s> students students toronto are university successful students were'

3. With multi-headed attention

   (a) '<s> toronto is one town of city of'
   (b) '<s> mothers parents should serving welcomed with of flyer <unk>'
   (c) '<s> students students toronto toronto very toronto crops very'

2.2 Discussion

It's obvious that the model without attention yields the worst results: it basically make no sense. The model with single and multi headed attention yield results of similar quality: they make some sense, but not much. This observation aligns with the model's BLEU scores: the higher the score, the better the translation.