CSC401 Homework Assignment #2
Bonus

Benson Li

1 Task Definition

In the **Section 2.2 Discussion**, I got the following conclusion:

```
Overall, BLEU score can be a rough indicator of the model's translation quality but
not perfect. It would be better to use in conjunction with some other evaluation
metrics (e.g. ROUGE, METEOR) to get a more comprehensive understanding of the model's
performance.
```

Therefore, I want to implement another evaluation metrics, which is **METEOR** (Metric for Evaluation of Translation with Explicit ORdering) and evaludate the model based on this metrics.

Please refer to the file **a2_bonus_meteor_score.py** for my implementation.

My implementation calculates the METEOR score between a candidate and reference transcription is based on the following paper:

```
Automatic evaluation of machine translation quality using longest common subsequence
and skip-bigram statistics. Chin-Yew Lin and Franz Josef Och. ACL 2004.
```

According to the paper, the METEOR score is a metric for evaluating the quality of machine translation output. It is similar to the BLEU score, but uses a more complex matching algorithm that takes into account synonyms and paraphrases. The score is calculated as follows:

1. Calculate unigram, bigram, and trigram precision between the reference and candidate transcriptions.

2. Calculate unigram, bigram, and trigram recall between the reference and candidate transcriptions.

3. Calculate the precision mean (p_mean) and recall mean (r_mean) using the alpha and beta parameters.

4. Calculate the F-mean using the p_mean, r_mean, and gamma parameters.

5. Calculate the brevity penalty (BP) between the reference and candidate transcriptions.

6. Multiply the F-mean and BP to get the final METEOR score.

2 Experimental Setup & Results

Due to time constraints and computational cost, I cannot conduct rigorous statistical testing to reach my conclusions. However, I make an outline for my experiment.

My null hypothesis (H_0) is: The conclusion for using METEOR score as the indicator is same as the conclusion for using BLEU score.

My alternative hypothesis (H_1) is: The conclusion for using METEOR score as the indicator is different with the conclusion for using BLEU score.

For comparison with BLEU scores, I trained three models as in the analysis.pdf: LSTM without Attention; LSTM with Single-headed Attention; LSTM with Multi-headed Attention. I trained my model for five epochs and pass the model to the test data. The result is shown below:

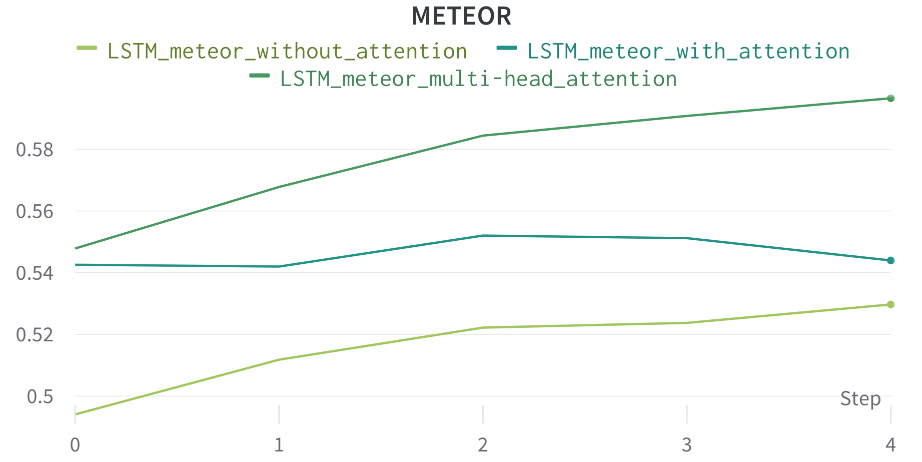

Figure 1: Wandb Training METEOR Score for Three Models

Model	Final Train METEOR	Test METEOR
LSTM without Attention	0.528	0.534
LSTM with Single-headed Attention	0.544	0.550
LSTM with Multi-headed Attention	0.597	0.602

Table 1: Comparison for Training and Testing METEOR for each model

3 Analysis

Comparing the three models, we can see that the LSTM with Multi-headed Attention achieved the highest METEOR score, both in the training and testing sets. Since the model with multi-headed attention can generate more accurate translations, we get the same conclusion with the method of using BLEU score.

Even though our statistical testing is not rigorous enough, after doing more experiment, I expect that we don't have enough evidence to reject H_0. I conclude that both METEOR and BLEU are useful metrics for evaluating machine translation, and it's would be better to use both of them together to get a more comprehensive evaluation.