Bonus Assignment (2%)

Grammar Error Correction – C4 Dataset (200M)

Mandatory Note

The full C4 dataset contains approximately 200 million samples. However, for this assignment:

- You must work on only 1 million samples
- Training on the full 200M dataset is strictly not allowed

The focus is on analysis and linguistic understanding, not dataset scale.

Objectives

This assignment aims to:

- Compare different text representations for Grammar Error Correction.
- Analyze model performance from a linguistic perspective.
- Evaluate the impact of sequential models and Attention mechanisms.
- Link model behavior to English grammar rules.

Required Tasks

1. Bag of Words Experiment

- Apply Bag of Words representation.
- Train a suitable model.
- Present results.
- Provide real input-output examples with explanations.

2. RNN or LSTM Without Attention

- Train an RNN or LSTM model.
- Evaluate and report results.
- Discuss limitations related to context and long sequences.

3. RNN or LSTM With Attention

- Add an Attention layer.
- Train and evaluate the model.
- Compare results with the non-attention model.

4. Results and Outputs

- Present evaluation metrics using appropriate metrics for Grammar Error Correction tasks.
- Show clear input-output examples.
- For each example, explain:
    - The corrected grammar rule
    - Or the grammar rule the model failed to correct

Final Challenge (Mandatory)

Answer clearly:

Are all English grammar rules in the C4 dataset correct, or does the dataset itself contain grammatical issues?

Your answer must be supported by:

- Dataset examples
- Model outputs
- Linguistic reasoning based on results

Submission Requirements

A complete documentation report including:

- Overview of the C4 dataset and justification for using only 1M samples.
- Explanation of all models.
- Results and comparisons.
- Grammar-rule-based analysis.

- Final challenge discussion.
- Final conclusion.