Improving Writing Assistance at JetBrains AI

1. Introduction

This report documents the evaluation of two spell-checking models, BERT and T5, to assess their effectiveness in correcting spelling errors. The goal was to evaluate these models on two key metrics: **Accuracy** (the proportion of correct top suggestions) and **Mean Reciprocal Rank (MRR)** (the average rank quality of the correct suggestions). Through this evaluation, we aimed to understand each model's strengths and weaknesses in terms of precise corrections and the relevance of suggestions.

2. Approach

The evaluation was conducted as follows:

- **Data Preparation**: A dataset of misspelled words with corresponding correct spellings was used. This dataset allowed us to evaluate each model's ability to identify and suggest the correct spelling.
- **Metrics Selection**:
  - **Accuracy** was chosen to measure the proportion of times the model's top suggestion matched the correct spelling.
  - **Mean Reciprocal Rank (MRR)** was selected to assess the quality of the ranking, particularly for cases where the correct answer was not the first suggestion.
- **Model Evaluation**:
  - **BERT-based Spell Checker**: Evaluated for its ability to leverage contextual understanding to generate relevant spelling suggestions.
  - **T5-based Spell Checker**: Assessed for its language generation capabilities in spelling correction, particularly in making context-based adjustments.

3. Results

The performance of each model was measured based on the Accuracy and MRR metrics.

- **BERT Model**:
  - **Accuracy**: 0.25 – BERT provided the correct suggestion as the top choice in 25% of cases.
  - **MRR**: 0.81 – BERT ranked the correct answer highly, indicating that the correct spelling often appeared among the top few suggestions.
  - **Analysis**: The BERT model demonstrated high-quality suggestions, as indicated by the high MRR score, but struggled to consistently rank the correct spelling as the first suggestion.
- **T5 Model**:

    - **Accuracy**: 0.20 – T5 correctly identified the top spelling suggestion in 20% of cases.
    - **MRR**: 0.32 – T5 had a lower MRR, suggesting that the correct spelling was not frequently among the top suggestions.
    - **Analysis**: The T5 model had lower performance on both metrics, indicating that it may lack both the precision and ranking quality required for effective spell checking in this context.

4. Challenges Encountered

Several challenges were encountered during the evaluation:

- **Model Limitations**: BERT and T5 were not fine-tuned specifically for spelling correction, which may have affected their precision. While these models are powerful for language tasks, they may not be optimized for the nuances of spelling correction without further training.
- **Metric Interpretation**: The MRR metric provided valuable insight into the ranking quality but highlighted that even with high MRR, the first suggestion was often incorrect. Balancing MRR with high accuracy proved challenging, especially for BERT.
- **Computational Resources**: Running BERT and T5 on large datasets was computationally intensive, especially in generating multiple suggestions per word. This limited the ability to experiment with larger datasets or more complex ensembling approaches.

5. Improvement Suggestions

To address the observed limitations and improve the spell-checking capability:

- **Fine-Tuning**: Fine-tuning BERT or T5 on a specialized spelling correction dataset could increase their accuracy for providing the correct top suggestions. This would make the models better suited for specific error patterns and more precise in their corrections.
- **Model Ensembling**: Combining the strengths of multiple models (e.g., ensembling BERT with traditional spell-checking tools) could improve both accuracy and MRR by leveraging different approaches to correction.
- **Context-Based Re-Ranking**: A re-ranking system that incorporates linguistic patterns or domain-specific knowledge could improve the placement of correct suggestions at the top, enhancing user satisfaction.

6. Conclusion

The evaluation revealed that while the BERT model provided higher ranking quality with an MRR of 0.81, it struggled to provide correct top suggestions consistently, resulting in a moderate accuracy of 0.25. The T5 model showed limited effectiveness in both accuracy (0.20) and MRR (0.32). Future work could explore fine-tuning and ensembling methods to address these issues and improve performance for practical spell-checking applications.