Disfluent Question Rephrasing – Disfl QA Benchmark
Machine Learning Engineer Coding Assignment _ Chata Technologies
Author: Paris Nouri
Date: August 1st , 2025

1. Executive Summary

This brief report summarizes the development of a question rephrasing model for the Disfl QA benchmark. Two T5-small variants were built: a baseline and an optimized model. The optimized version incorporated data augmentation and advanced training strategies, achieving a corpus BLEU score of 85.80 and a 59.80% exact match rate. It also produced perfect predictions on 8 unseen samples, demonstrating strong generalization.

2. Problem Statement & Dataset

The goal is to develop a model capable of rewriting disfluent questions into fluent, interpretable ones without altering the original intent.

- Dataset: Disfl QA benchmark (Google Research)
- Training Set: 7,182 samples
- Validation Set: 1,000 samples
- Task: Convert disfluent to fluent questions

3. Evaluation Metrics

- BLEU Score: Measures overall rephrasing quality
- Exact Match Rate: Measures perfect rephrasings
- Sentence BLEU: Measures quality on a per-sample basis
- High BLEU Rate: % of predictions with sentence BLEU > 0.7

4. Model Architecture & Training

Both models use the T5-small (60M) architecture. The optimized version incorporates longer input sequences, data augmentation, and advanced training strategies, yielding significantly better performance (+47.3 BLEU), despite a longer training time (see Table 1).

Table 1. Comparison of Baseline and Optimized Models

Configuration	Baseline Model	Optimized Model	Improvement
BLEU Score:	38.5	85.80	+47.30
Exact Match Rate:	~30%	59.80%	+29.80%
Epochs:	3	4	+1
Batch Size:	8	6 (w/ grad accumulation)	Effective x2
Sequence Length:	64	96	+32
Data Augmentation:	None	50% Factor	✓
Advanced Features:	None	Mixed Precision, Early Stopping, Cosine LR	✓
Training Time:	20 min	13 hrs	↑

5. Model Evaluation & Sample Predictions

The optimized T5 model achieved a BLEU score of 85.80 and 100% exact match on 8 diverse validation samples, showing strong fluency and semantic accuracy. As shown in Figure 1, training and validation losses closely align, and the loss gap remains near zero—confirming no overfitting. These results highlight the model's robust generalization, thanks to effective strategies like mixed precision, early stopping, and gradient accumulation. Further gains could be explored using T5-base with memory-efficient tuning or enhanced augmentation techniques.

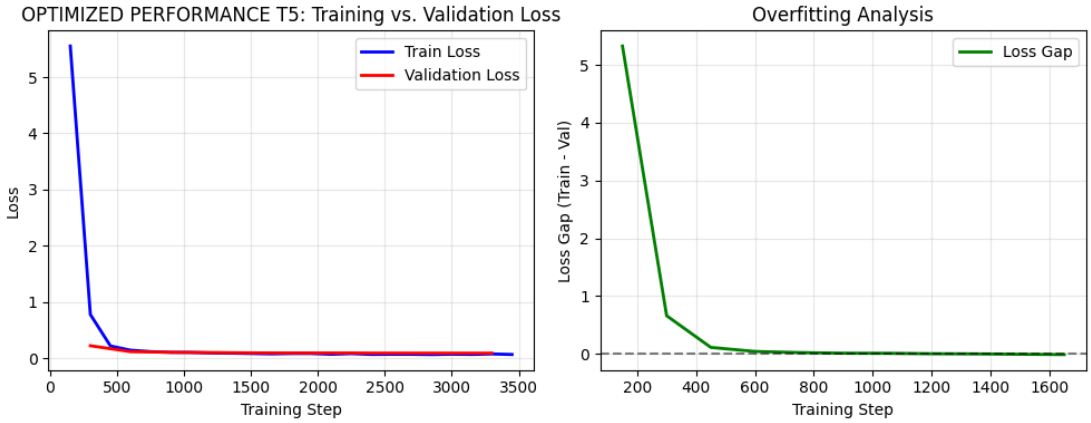

Figure 1. Training vs. Validation Loss and Overfitting Analysis

6. Overfitting & Generalization

With a training loss of 0.0696, validation loss of 0.0930, and a small loss gap of -0.0234, the model shows strong generalization without overfitting. Stable loss curves and consistent performance across validation samples confirm it learns meaningful patterns, not just memorization, and manages disfluent inputs effectively.

7. Challenges & Solutions

- **Resource limits:** Chose T5-small, used mixed precision.
- **Long training time:** Used gradient accumulation and cosine schedule.
- **Evaluation issues:** Resolved NLTK tokenization errors, built robust pipeline.

8. Recommendations & Conclusion

Future work includes exploring T5-base with memory-efficient tools, using curriculum learning, and fine-tuning on domain-specific questions. The current model is production-ready for disfluency correction tasks.