Task B: Quality Improvement Strategies for Increasing BLEU Score

NLP Applications - Assignment 2 - Group 5

Team Members

Name	Student ID	Role	Contribution
Karan Sharma	2024AB05145	Documentation, Testing & Visualization	100%
Neerumalla Kavitha	2024AA05879	Data Preprocessing & N-gram Analysis	100%
Selva Pandian S	2023AC05005	Backend API & Model Integration	100%
Shikhar Nigam	2024AA05691	Frontend Development & UI/UX	100%
Suraj Anand	2024AA05731	System Architecture & BLEU Implementation	100%

Introduction

This document analyzes three key strategies to improve BLEU scores in Statistical Machine Translation (SMT) systems.

Strategy 1: Using More Training Data

Concept: SMT systems rely on statistics from parallel corpora. More data equals better probability estimates and vocabulary coverage.

Impact

Training Sentences	Typical BLEU	Improvement
10,000	0.15-0.20	Baseline
1,000,000	0.35-0.42	High

Sources: Europarl, UN Corpus, OpenSubtitles. **Limitation**: Diminishing returns after significant scaling; domain mismatch can hurt quality.

Strategy 2: Better Language Models

Concept: The Language Model (LM) ensures fluency. Higher-order n-grams or Neural LMs capture better context.

Approaches

1. **High-order n-grams**: Moving from 3-gram to 5-gram LMs.
2. **Neural LMs**: Using RNN or Transformer-based LMs for rescoring.
3. **More Monolingual Data**: LMs can be trained on vast amounts of cheap monolingual text.

Impact: +2 to +4 BLEU points by improving output naturalness.

Strategy 3: Domain-Specific Parallel Corpora

Concept: Training on data that matches the target domain (e.g., Medical, Legal) yields the highest ROI.

Comparative Impact

Training Data	Test Domain	BLEU Score
General (News)	Medical	0.15 (Poor)
Medical	Medical	0.42 (Good)

Recommendation: Fine-tune a general model with specific domain data for best results.

Comparative Analysis & Conclusion

Strategy	Effort	Impact	Best For
More Data	High	Medium	General purpose systems
Better LM	Medium	Low-Medium	Improving fluency
Domain Data	High	**High**	Specialized applications

References

1. Papineni et al. (2002). "BLEU: a Method for Automatic Evaluation of Machine Translation"
2. Koehn, P. (2009). "Statistical Machine Translation"