# Title

DDAE: Dynamic Difficulty-Aware Evaluation for Machine Translation

# Problem Statement

Current machine translation (MT) evaluation metrics often treat all sentences equally, failing to account for the varying difficulty of translating different types of content. This can lead to skewed evaluations that don't reflect the true capabilities of MT systems across diverse translation challenges.

# Motivation

Existing metrics like BLEU or COMET provide uniform scoring across all input types, while some metrics attempt to incorporate linguistic features but don't dynamically adjust based on translation difficulty. Human evaluators inherently consider the difficulty of a translation task when assessing quality. An evaluation metric that can simulate this dynamic difficulty assessment could provide a more nuanced and fair evaluation of MT systems. Our proposed DDAE framework aims to address this gap by incorporating translation difficulty into the scoring mechanism, potentially offering a more accurate representation of MT system performance across various linguistic phenomena and difficulty levels.

# Proposed Method

We propose DDAE (Dynamic Difficulty-Aware Evaluation), a novel MT evaluation framework that dynamically assesses and incorporates translation difficulty into its scoring mechanism. DDAE operates as follows: 1) It uses a difficulty estimation module trained on a large corpus of human-annotated translation pairs to predict the difficulty of translating a given source sentence. This module considers factors like sentence length, rare words, idiomatic expressions, syntactic complexity, and required world knowledge. 2) For each source-translation pair, DDAE generates multiple difficulty-specific evaluation criteria, emphasizing different aspects of translation quality based on the estimated difficulty. 3) It employs a neural critic model that evaluates the translation against these dynamic criteria, producing a difficulty-calibrated score. 4) DDAE aggregates these scores across a test set, providing both an overall performance metric and granular analysis of system performance across different difficulty levels and linguistic phenomena.

# Step-by-Step Experiment Plan

## Step 1: Data Collection and Preparation

Collect a diverse dataset of source-translation pairs with human-annotated difficulty ratings and quality judgments. Use existing parallel corpora like WMT datasets and supplement with additional annotations. Aim for at least 10,000 sentence pairs across multiple language pairs and domains.

## Step 2: Difficulty Estimation Module

Train a neural network to predict translation difficulty. Use features such as sentence length, rare word ratio, syntactic parse tree depth, and named entity counts. Train on the human-annotated difficulty ratings from Step 1. Use a transformer-based architecture and experiment with different feature combinations.

## Step 3: Dynamic Criteria Generation

Develop a module that generates difficulty-specific evaluation criteria. Create a set of base criteria (e.g., adequacy, fluency, terminology) and design rules to adjust their weights based on the estimated

difficulty. Implement this as a rule-based system initially, with the potential to make it learnable in future iterations.

### Step 4: Neural Critic Model

Train a neural network to evaluate translations against the dynamic criteria. Use a transformer-based architecture that takes as input the source sentence, machine translation, reference translation, difficulty score, and evaluation criteria. Train on the human judgments from Step 1, optimizing for correlation with human scores.

### Step 5: DDAE Framework Integration

Combine the difficulty estimation module, dynamic criteria generation, and neural critic model into a cohesive framework. Implement a pipeline that processes source-translation pairs and outputs difficulty-calibrated scores.

### Step 6: Baseline Implementation

Implement baseline metrics for comparison, including BLEU, COMET, and BERTScore. Ensure all metrics are applied consistently to the same test sets.

### Step 7: Evaluation

Evaluate DDAE against baselines on a held-out test set from Step 1. Measure correlation with human judgments overall and across different difficulty levels. Analyze performance on specific linguistic phenomena and translation challenges.

### Step 8: Case Study Analysis

Conduct in-depth analysis of specific examples where DDAE performs differently from baseline metrics. Examine how the dynamic difficulty assessment affects scoring in various scenarios.

### Step 9: Meta-evaluation

Assess DDAE's effectiveness in comparing multiple MT systems. Use outputs from top-performing systems in recent WMT shared tasks. Analyze how DDAE's rankings differ from those of standard metrics and human evaluations.

## Test Case Examples

### Baseline Metric Input

Source: 'The cat sat on the mat.' MT Output: 'The feline rested on the rug.' Reference: 'The cat was sitting on the mat.'

### Baseline Metric Output

BLEU Score: 0.223 COMET Score: 0.876

## Baseline Metric Explanation

Standard metrics like BLEU and COMET provide a single score without considering the relative simplicity of this sentence.

## DDAE Input

Source: 'The cat sat on the mat.' MT Output: 'The feline rested on the rug.' Reference: 'The cat was sitting on the mat.'

## DDAE Output

Difficulty Score: 0.2 (Low) Adequacy Score: 0.95 Fluency Score: 0.98 Lexical Choice Score: 0.85 Overall DDAE Score: 0.92

## DDAE Explanation

DDAE recognizes this as a simple sentence and adjusts its scoring criteria accordingly. It gives high scores for adequacy and fluency but slightly lower for lexical choice due to the synonym usage. The overall score is high, reflecting that for a simple sentence, this is a good translation despite not matching the reference exactly.

# Fallback Plan

If DDAE doesn't show significant improvements over baseline metrics, we can pivot to an analysis paper exploring the relationship between translation difficulty and evaluation metric performance. We would conduct a comprehensive study of how different types of translation challenges affect various metrics. This could involve creating a taxonomy of translation difficulties and analyzing how existing metrics perform across these categories. We could also investigate whether certain types of MT systems (e.g., neural vs. statistical) are more or less sensitive to translation difficulty in terms of evaluation scores. Additionally, we could explore the potential of using DDAE's difficulty estimation module as a standalone tool for analyzing dataset complexity or as a method for creating more balanced and representative test sets for MT evaluation.

Ranking Score: 6