# Title

AFAD: Adversarial Fluency-Adequacy Disentanglement for Improved Machine Translation Evaluation

# Problem Statement

Current machine translation (MT) evaluation metrics often conflate fluency (grammaticality and naturalness) with adequacy (semantic faithfulness), making it difficult to pinpoint the specific strengths and weaknesses of MT systems. This conflation hinders the development of more targeted improvements in MT systems and limits our understanding of translation quality.

# Motivation

Existing MT evaluation metrics typically provide a single score that combines both fluency and adequacy, or attempt to measure them separately using predefined features or heuristics. These approaches often fail to capture the nuanced differences between these two aspects of translation quality. By disentangling fluency and adequacy in a more robust way, we can provide more actionable feedback for improving MT systems and offer a more nuanced evaluation of translation quality. Our proposed method, Adversarial Fluency-Adequacy Disentanglement (AFAD), is inspired by recent advances in adversarial training and disentangled representation learning. By leveraging adversarial training, we can force our model to learn truly independent representations of fluency and adequacy, overcoming the limitations of current heuristic-based approaches.

# Proposed Method

We propose Adversarial Fluency-Adequacy Disentanglement (AFAD), a novel evaluation framework that uses adversarial training to separate fluency and adequacy scores. The framework consists of three main components: 1) A fluency scorer trained on monolingual data in the target language. 2) An adequacy scorer trained on parallel data. 3) An adversarial discriminator that tries to predict the adequacy score from the fluency score. During training, the fluency and adequacy scorers are optimized to maximize their respective metrics while minimizing the discriminator's ability to predict one from the other. This forces the scorers to focus on truly independent aspects of translation quality. We also incorporate a 'difficulty estimator' that adjusts the scores based on the complexity of the source text, ensuring fair comparisons across different text types.

# Step-by-Step Experiment Plan

## Step 1: Data Preparation

Collect and preprocess the following datasets: 1) A large monolingual corpus in the target language (e.g., English) for fluency scoring. We can use the WMT News Crawl dataset. 2) A parallel corpus for adequacy scoring. We can use the WMT14 English-German dataset. 3) Standard MT evaluation datasets such as WMT20 metrics shared task data for testing.

## Step 2: Model Architecture

Implement the AFAD framework using PyTorch. The fluency scorer and adequacy scorer can be based on pre-trained language models like BERT or RoBERTa. The discriminator can be a simple feed-forward neural network. The difficulty estimator can be a regression model trained on human-annotated difficulty scores.

## Step 3: Training

Train the AFAD model using the following steps: a) Pre-train the fluency scorer on the monolingual data. b) Pre-train the adequacy scorer on the parallel data. c) Train the discriminator to predict adequacy scores from fluency scores. d) Jointly train the entire model, optimizing the fluency and adequacy scorers while adversarially training against the discriminator.

### Step 4: Evaluation

Evaluate AFAD on the WMT20 metrics shared task data. Compare its performance to existing metrics that attempt to measure fluency and adequacy separately, such as COMET and BLEURT. Use Pearson correlation with human judgments as the primary evaluation metric.

### Step 5: Human Evaluation

Conduct a human evaluation study to validate the disentanglement of fluency and adequacy. Recruit bilingual annotators to rate a subset of translations for fluency and adequacy separately. Compare these human ratings with the scores produced by AFAD.

### Step 6: Analysis

Analyze the strengths and weaknesses of different MT systems using AFAD. Identify systems that excel in fluency but struggle with adequacy, and vice versa. Use these insights to provide targeted recommendations for improving specific MT systems.

### Step 7: Ablation Studies

Conduct ablation studies to understand the contribution of each component in AFAD. Test versions without the adversarial training, without the difficulty estimator, and with different model architectures for the scorers.

## Test Case Examples

### Baseline Prompt Input

Source: Der Hund jagt die Katze. Translation: The dog chases the cat.

### Baseline Prompt Expected Output

BLEU Score: 1.0 METEOR Score: 1.0

### Proposed Prompt Input

Source: Der Hund jagt die Katze. Translation: The dog chases the cat.

### Proposed Prompt Expected Output

Fluency Score: 0.95 Adequacy Score: 0.98 Difficulty-Adjusted Overall Score: 0.97

### Explanation

While traditional metrics like BLEU and METEOR provide a single score, AFAD disentangles fluency and adequacy, offering more nuanced feedback. The high fluency score indicates grammatical correctness and naturalness in English, while the high adequacy score shows that the meaning is accurately preserved. The difficulty-adjusted overall score accounts for the simplicity of the sentence.

# Fallback Plan

If AFAD doesn't significantly outperform existing metrics, we can pivot to an analysis paper that explores the challenges of disentangling fluency and adequacy. We would conduct a thorough error analysis to understand where and why AFAD fails. This could involve categorizing different types of translation errors and examining how AFAD and other metrics handle them. We could also investigate the correlation between AFAD's fluency and adequacy scores to see if true disentanglement is achieved. Additionally, we could explore how different adversarial training techniques affect the disentanglement and overall performance. This analysis could provide valuable insights into the limitations of current MT evaluation approaches and guide future research in this area.

Ranking Score: 6