Reinforcement Learning for Machine Translation: Fine-Tuning MT5 LLM for English-Indonesian Translation

Arkan Abuyazid
Department of Computer Science
Texas A&M University
College Station, Texas
ata757@tamu.edu

***Abstract*—This proposal outlines an exploratory approach employing reinforcement learning (RL) to fine-tune the MT5-base model for English-Indonesian translation. Our objective is to investigate the feasibility of leveraging RL-based techniques, with translation quality metrics (such as BLEU, ROUGE-L, and METEOR) serving as reward signals. By framing translation as a sequential decision-making process within a Markov Decision Process (MDP), we aim to establish a flexible optimization framework and compare RL algorithms, notably REINFORCE with baseline subtraction and Proximal Policy Optimization (PPO), against more conventional fine-tuning practices. The focus of this work is to explore the potential benefits and underlying mechanisms of reward-driven fine-tuning in low-resource settings.**

***Index Terms*—Neural Machine Translation, Reinforcement Learning, MT5, Low-Resource Languages, Preference Optimization**

I. Proposal

The central aim of this proposal is to explore how reinforcement learning can enhance machine translation, specifically addressing challenges faced when working with low-resource language pairs like English-Indonesian. We provide an initial framework and motivation for applying RL to an existing multilingual text-to-text transformer model (MT5-base) with the goal of improving translation quality using a combination of automatic evaluation metrics. We select MT5-base (580M parameters) over larger alternatives like mT5-XXL (13B) due to its demonstrated balance between multilingual capability and practical deployability. Our approach is designed to further understand the dynamics and potential improvement pathways offered by reward-driven methods.

A. Sequential Nature

Machine translation is by nature a sequential generation task where each output token is influenced by previously generated tokens. Recognizing this inherent sequentiality, our work frames the translation process as a partially observable Markov Decision Process (POMDP). Each decision made by the decoder is critical, as it influences the semantic and syntactic construction of the final sentence. Our exploration focuses on how the sequential dependencies can be more effectively captured by integrating long-term reward signals into token generation decisions, and thus addresses issues of coherence and fluency across lengthy translations.

B. State Space

The state at each time step, denoted as s_t, encapsulates several components:

- **Encoder's Source Representation:** The semantic and syntactic information extracted from the English sentence.
- **Decoder's Hidden State:** The internal state reflecting the decoded sequence up to token t, which carries forward context.
- **Generated Target Tokens:** The sequence of previously generated Indonesian tokens serves both as context and a historical path for current decisions.
- **Attention Weights History:** A record of attention distributions to potentially modulate rewards based on alignment quality.

Expanding on these components, we intend to analyze how each aspect contributes to the overall translation quality and how the RL agent might best utilize this state information for improved decision making.

C. Action Space

In our formulation, the action space A is defined as the full MT5 vocabulary, comprising 250,112 tokens. Each action represents the selection of a next token, with decisions made under a temperature-controlled softmax sampling strategy to balance exploration and exploitation. The extensive vocabulary challenges the RL algorithm to manage a high-dimensional space effectively. Our investigation will shed light on how this combinatorial complexity affects learning dynamics and convergence.

D. Transition Function

Given the transformer architecture, the transition function is modeled as deterministic:

$$s_{t+1} = \mathrm{MT5}_\theta(s_t, a_t) \tag{1}$$

Here, the function encapsulates the state update based on the current state and the token chosen. In our proposed exploration, we will consider how this deterministic mapping interacts with stochastic policy choices, as well as potential modifications that could account for uncertainty and long-term dependencies in the generated sequences.

E. Reward Function

The reward function R_t is a weighted combination of several quality metrics:

$$R = \lambda_1 \text{BLEU} + \lambda_2 \text{ROUGE-L} + \lambda_3 \text{METEOR} - \lambda_4 \text{RepPenalty} \tag{2}$$

Each coefficient λ_i will be tuned via a proposed grid search method to balance the contributions of precise translation match (BLEU), sequence-level overlap (ROUGE-L), and semantic alignment (METEOR), while penalizing undesirable repetitive patterns. Our aim is to explore various weighting schemes to understand their impact on the translation policy, laying the groundwork for RL strategies that can optimally balance multiple metrics.

F. Initial Framework and Baseline

At this stage, our plan is to establish the RL framework by implementing the REINFORCE algorithm enhanced with a baseline subtraction. This will serve as a preliminary method, benchmarked against the standard supervised fine-tuning approach. Our evaluation strategy—though conceptual at this point—will include:

- **Automatic Metrics:** BLEU-4, ROUGE-L, and METEOR scores computed in a post-hoc analysis to quantify translation quality.
- **Human Evaluation:** A plan to incorporate human judgments on translation adequacy and fluency using the Flores-200 evaluation benchmark.
- **Training Dynamics:** Analysis of training stability and convergence characteristics within the RL framework.

These measures are proposed to validate the conceptual benefits of the RL approach for translation tasks, while noting that our current focus is on methodology and potential rather than empirical validation.

G. Stretch Goal

Beyond the initial RL implementation, we propose a stretch goal that involves exploring advanced preference optimization techniques. This includes:

- **Proximal Preference Optimization (PPO):** An investigation into PPO as a scalable method for aligning translation policies with human preferences.
- **Contrastive Preference Optimization (CPO):** Experimenting with pairwise ranking of translations to incorporate nuanced human feedback mechanisms.

The stretch goal is designed to expand our understanding of how complex reward structures can guide large language models (LLMs) towards not only higher quality translations but also more human-aligned outputs. This part of the proposal will explore theoretical considerations and the potential methodological integration of these techniques into the RL framework.

II. Literature Review

Our research builds on a robust set of prior works that have investigated reinforcement learning in natural language processing, preference optimization, and neural machine translation. The following subsections expand on the existing literature and illustrate why our proposed approach is both relevant and promising.

A. Reinforcement Learning Foundations

The REINFORCE algorithm [1] laid the groundwork for policy gradient methods in sequence generation tasks. Despite its pioneering role, the algorithm faces challenges with delayed rewards and large action spaces—issues we aim to address by adapting baseline subtraction and entropy regularization techniques. Our proposal is motivated by the potential to extend these classical methods to handle the intricacies of modern transformer architectures.

B. Preference Optimization and Policy Alignment

Recent advancements in preference optimization, including Direct Preference Optimization (DPO) [4] and Proximal Preference Optimization (PPO) [5], have shown significant promise in aligning language model outputs with human values. The introduction of Contrastive Preference Optimization (CPO) [6] further demonstrates the value of incorporating pairwise comparisons into the learning process. These methodologies highlight the potential of explicitly modeling human preferences alongside traditional reward metrics, and our work proposes to integrate such approaches into the context of machine translation.

C. Evaluation Metrics in Machine Translation

Evaluation metrics such as BLEU [2], ROUGE-L [9], and METEOR [10] have long served as proxies for human judgment in machine translation. The literature reveals both strengths and limitations in these metrics—while BLEU effectively captures n-gram overlap, ROUGE-L provides insights into longer sequence coherence, and METEOR offers semantic alignment considerations. By combining these metrics into a unified reward function, our approach seeks to leverage the complementary strengths of each metric to provide a more holistic assessment of translation quality.

D. Model Architecture and Low-Resource Challenges

MT5 [7] represents a crucial step toward building multilingual models that can handle a broad spectrum of languages, including those with relatively low resources. While the transformer architecture has achieved remarkable success in high-resource settings, its direct application to low-resource language pairs often exposes limitations in translation fidelity and coverage. Our research is driven by the hypothesis that an RL-based fine-tuning strategy can better adapt large models like MT5 to the unique challenges present in low-resource translation scenarios, bridging gaps that conventional supervised fine-tuning might not address.

E. Dataset Considerations

The Flores-200 benchmark [3] offers a standardized evaluation framework specifically designed for low-resource languages. By aligning our evaluation strategy with the methodologies established in Flores-200, we intend to create a consistent and comparable measure of translation quality. Our current proposal focuses on the conceptual design of these evaluation metrics as integral components of the reward function within our RL framework.

REFERENCES

[1] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, 1992.

[2] K. Papineni et al., "BLEU: a method for automatic evaluation of machine translation," ACL 2002.

[3] N. Goyal et al., "The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation," arXiv:2106.03193, 2022.

[4] R. Rafailov et al., "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," arXiv:2305.18290, 2023.

[5] A. Rafailov et al., "Proximal Preference Optimization: A New Approach to Scalable Policy Alignment," Technical Report, 2023.

[6] Z. Sun et al., "Contrastive Preference Optimization: Pushing the Boundaries of LLM Utility in Alignment," arXiv:2401.08417, 2024.

[7] L. Xue et al., "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," NAACL 2021.

[8] A. Radford et al., "Improving Language Understanding by Generative Pre-Training," OpenAI Technical Report, 2018.

[9] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," ACL 2004.

[10] S. Banerjee et al., "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," ACL 2005.