FINE-TUNING NEWS SUMMARIZATION MODEL WITH HUMAN PREFERENCES

Rimika Dhara, Nikhitha Gollamudi, Akhil Mavallapalli & Xifan Li
Department of Computer Science and Engineering
University of Minnesota, Twin Cities
{dhara015, golla063, maval003, li003646} @umn.edu

1 INTRODUCTION AND PROBLEM STATEMENT

Traditional news summarization models lack optimization with respect to human preference. Currently, transformer-based summarization models are facing issues including inaccurate information and verbose summaries that are not aligned with how humans would normally prefer their news to be summarized. Given these issues, the problem we aim to tackle is improving how accurately, precisely, and concisely we can tune our model to align more closely with human preferences. We also want to understand how much difference this tuning would make to the established pre-trained models.

Our approach to this problem is to implement reinforcement learning (RL) by training a reward model on human-labeled preferences. More specifically, we employ two RL algorithms, Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO), in hopes of increase the performance of the summarization model. This will help directly tackle current issues by optimizing the model to create summaries that humans prefer. Building on the work of Ziegler et al. (2019) and Stiennon et al. (2020), we apply RL from human feedback (RLHF) specifically to the news domain, with particular attention to factual accuracy and conciseness.

2 SIGNIFICANCE OF THE PROBLEM

Automated summarization plays a vital role in modern information consumption, enabling quick understanding of large texts by highlighting key ideas. This is especially impactful in news summarization, where journalism can benefit from faster, more digestible content. However, current systems face major issues—most notably, factual inaccuracy and verbosity—which limit their effectiveness. Inaccurate or overly long summaries can lead to misinformation, reduce readability, and erode trust in AI-generated content, ultimately defeating the purpose of aiding information processing.

Beyond journalism, accurate and concise summarization has critical applications in domains like healthcare, legal analysis, and finance, where mistakes can be costly and time is a premium resource. Although transformer-based models have advanced the field, existing limitations show the need for better alignment with human preferences. Incorporating human feedback into training—through methods like RL—can improve both reliability and utility. This work has the potential to reshape how AI contributes to information consumption, setting new standards for trustworthy, human-aligned summarization systems.

3 PREVIOUS AND RELATED WORK

The field of language model fine-tuning from human preferences has seen significant developments in recent years. Notably, Ziegler et al. (2019) demonstrated that RL from human feedback (RLHF) can successfully fine-tune language models for tasks including summarization. Their work on the TL;DR and CNN/Daily Mail datasets showed that models trained with 60,000 human comparisons outperformed supervised baselines according to human evaluators, although they observed that the resulting models often acted as ”smart copiers” that selected relevant sentences from the input. Building on this foundation, Stiennon et al. (2020) further refined the approach for summarization

tasks, showing that human feedback can be more effective than traditional supervised learning methods. Their work emphasized the importance of high-quality human feedback and introduced iterative data collection procedures to improve model performance. In parallel, recent work by Böhm et al. (2019) also explored using human evaluations to learn reward functions for summarization, providing detailed investigations of learned policies on the CNN/Daily Mail dataset. Meanwhile, advances in large language models like T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) have established new state-of-the-art results on summarization benchmarks using traditional supervised approaches. More recently, researchers have explored various techniques to improve RLHF, including Constitutional AI (Bai et al., 2022), which combines rule-based and preference-based learning, and Direct Preference Optimization (Rafailov et al., 2023), which reformulates the preference learning problem to avoid the instabilities associated with traditional RL approaches. These advancements provide significant evidence that this is a promising direction.

4 Goals, Objective and Methodology

Our primary goal is to develop a news summarization model that produces accurate, concise summaries aligned with human preferences by fine-tuning a pre-trained transformer summarization model with RL algorithms on human preferences. We will achieve this through the following objectives:

1. Establish a baseline performance using existing pre-trained language models (e.g., T5, BART) on news summarization datasets
2. Identify and tokenize a high-quality human preference dataset for news summaries with specific focus on factual accuracy and conciseness of the summaries
3. Train a reward model from these human preferences that can effectively distinguish between preferred and non-preferred summaries
4. Fine-tune the baseline language model using RL to optimize for the learned reward
5. Evaluate the RLHF-tuned model against baselines and analyze improvements in accuracy and conciseness

5 Plan and Milestones

- **Week 1 (4/6 – 4/12): Dataset collection and preprocessing.**
  - Dissect, preprocess, and tokenize the CNN/Daily Mail dataset.
  - Conduct a literature review on transformer architectures and reinforcement learning.
- **Week 2 (4/13 – 4/19): Reward model training.**
  - Set up the base summarization model and establish benchmark statistics.
  - Train a reward model using collected human preferences.
  - Initialize the reward model from the baseline and use a comparative loss function.
- **Week 3 (4/20 – 4/26): Fine-tune the model with RLHF using PPO.**
  - Fine-tune the baseline model using Proximal Policy Optimization (PPO) with the reward function.
  - Explore Direct Preference Optimization as an alternative fine-tuning method.
  - Apply a KL divergence penalty between the fine-tuned and original models.
- **Week 4 (4/27 – 5/3): Model evaluation.**
  - Evaluate the model for summary accuracy and conciseness.
  - Use ROUGE and BERTScore as primary evaluation metrics.
- **Week 5 (5/4 – 5/10): Results analysis and report documentation.**
  - Analyze the final results and draw conclusions.
  - Prepare and complete the final project report.

REFERENCES

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Florian Böhm, Yang Gao, Christian M Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. Better rewards yield better summaries: Learning to summarise without references. *arXiv preprint arXiv:1909.01214*, 2019.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize from human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.