報告書

1 今週の進捗

- DPO について
- Neo4j について
- ポスターについて

2 Direct Preference Optimization (DPO)

Direct Preference Optimization (DPO) [1] は機械学習における強化学習の手法の 1 つであり, 特に大規模言語モデル (LLM) などの人間の好みに基づくモデルの微調整に適している. DPO の特徴は, 報酬関数の推定を省略し, 直接的に人間の選好データからポリシーを最適化するアプローチを採用する点にある. これにより, 従来の手法に比べて効率的かつ解釈可能な形でモデルの調整が可能となる.

2.1 DPO の説明

Reinforcement Learning from Human Feedback (RLHF) は人間のフィードバックを用いて強化学習をする手法であり, 言語モデルに対して報酬モデルを使い方策を学習する. 一方, DPO では, 直接データセットを用いてその差分により方策を更新することで言語モデルを効率的かつ効果的に学習させる手法である. 図 1 に RLHF および DPO のモデル概略図を示す. 表 1 に DPO に用いる学習データセットのフォーマットの例を示す.

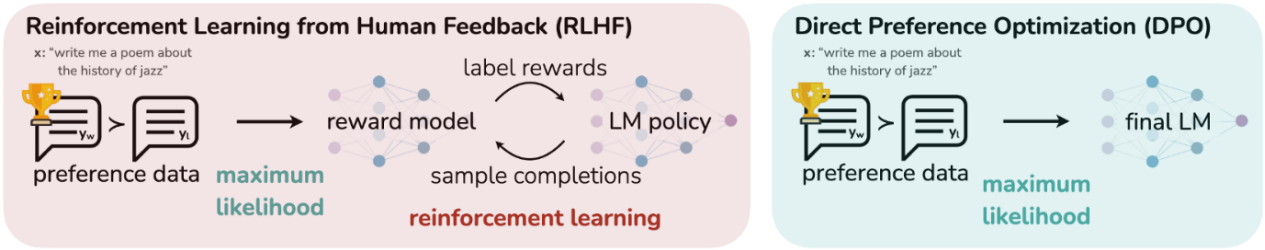

図 1: RLHF と DPO の違い (参考文献 [1] の Figure 1 より引用)

表 1: DPO の学習データセットのフォーマット

Prompt	Chosen	Rejected
What is the capital of France?	The capital of France is Paris.	France has many beautiful cities.
How do I boil an egg?	Place it in boiling water for 10 min.	Boiling an egg involves hot water.
Translate 'hello' to Spanish.	'Hello' in Spanish is 'Hola'.	'Hello' can be translated into Spanish.

文章からのナレッジグラフ生成において, どのように "Prompt", "Chosen", "Rejected" を作成すればよいか, 考える必要がある.

3 Neo4j

3.1 利点

- RAG への応用可能
- ナレッジグラフの構造に最適
- リアルタイム解析が可能

4 情報知識学会 (JSIK)

参加申し込み (11/27 まで). ポスター発表者も必要.

4.1 概要

「研究データエコシステム × 地域資料の保存・継承」 〜災害を乗り越え地域資料継承に貢献する研究データエコシステムの未来〜

- 10 月 18 日（金）発表原稿提出期限 12:00
- 11 月 30 日（土）情報知識学フォーラムの開催

4.2 11 月 30 日（土）当日 プログラム

- 12:30 受付開始（2 階・研修室 1）
- 13:00 - 13:10 開会宣言・開会挨拶
- 13:10 - 16:00 講演等
- 16:00 - 16:05 休憩
- 16:05 - 16:25 ポスター概要発表（1 件 90 秒以内)
- 16:25 - 17:10 ポスター発表
- 17:10 - 17:15 閉会宣言・事務連絡
- 18:30 - 20:00 情報交換会（要予約 11 月 26 日 12:00 まで, 6000 円）

5 今後

- ポスターの製作
- DPO のためのデータセット作成および実装
- Neo4j の実装

参考文献

[1] 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. 2024.