認知フィードバック：眼球運動・脳波による大規模言語モデルの強化学習

原田宥都　大関洋平
東京大学
{harada-yuto,oseki}@g.ecc.u-tokyo.ac.jp

概要

人間のフィードバックを用いて、大規模言語モデルにより好ましい出力を教示する方法が成功しているが、脳波や視線などの認知的なフィードバックの有用性は未だ調査されていない。本研究では、人間のフィードバックにおける問題点を克服するための方法として、認知フィードバックを用いて大規模言語モデルを調整するための手法を提案し、その手法の有効性を調査する。実験の結果、規模の小さなデータセットにおいてもある程度人間のフィードバックを代替、あるいは改善できる可能性を示した。

1　はじめに

近年の大規模言語モデルの成功を支える中心的な技術として、人間のフィードバックを用いた強化学習 (Reinforcement Learning from Human Feedback) がある [1]。これはモデルの出力に対する人間の嗜好に合わせるために効果的な手法であり、ChatGPT[2] のような最新の大規模言語モデルの成功において重要な役割を果たしている。RLHF では、強化学習を用いることで、教師ありファインチューニングだけでは難しいような複雑な目的に達することができる一方で、高品質な人間のフィードバックラベルを得るためには時間とコストがかかるという難点がある。もしコストをかけてラベラーとの綿密なやり取りが実現できたとしても、ラベラーは単純なミスを犯す、あるいは重大なミスを見逃してしまう可能性があり、有害な目標を無意識的に追求してしまうこともあり得る。

RLHF における人ラベリングをテキストに対する明示的なフィードバックとすると、人間が言語を理解する際に記録された脳活動や視線は、テキストに対する暗黙的なフィードバックと呼ぶことができる。本研究では、それらの記録を認知フィードバックと呼び、その有効性を調査する初めての実験を行う。具体的には、テキストについての無意識的で直感的な反応の記録である認知フィードバックから、(1) 明示的なフィードバックラベルを再現・代替できるか？ (2) 熟練していないラベラーのフィードバックを改善できるか？ これら 2 つのリサーチクエスチョンについて答えるために、認知フィードバックを用いた強化学習のための実験の枠組みを提案する。

2　関連研究

2.1　RLHF

Reinforcement Learning from Human Feedback(RLHF) は、対話システムの構築よりも先に要約タスクにおいて実装され [1]、その後 ChatGPT の前進となる Instruct-GPT[3] において対話システムへの導入が進んだ。ターゲットタスクによってやや異なることがあるが、人間のフィードバックを用いて強化学習を行うためには、大体以下のような 3 つの手順を踏む必要がある。

Step1: 人間のフィードバックの収集　まず、人間のラベラーによるフィードバックの収集を行う。1 つの入力に対して複数のモデルの出力を用意し、その中から 2 つを選んだ時、そのどちらがより好ましい出力かを選んでもらい、それを人間によるフィードバックとする。この時、複数の出力を一挙に比較することは難しい採点タスクとなるため、2 つだけを比較することを何度も繰り返すことでより正確なフィードバックを得る目的がある。

Step2: 報酬モデルの訓練　Step1 で収集したデータを正解データとして、報酬モデルを訓練する。報酬モデルとしては、事前学習済みのある程度規模の大きい言語モデルを利用する。先に対象のデータで

ファインチューンしておくことが一般的である。出力層を変更した上で、報酬モデルとして学習する。報酬モデルは、テキストのペアが与えられた時、それぞれのテキストにスコアを割り振る。人間がより好ましいと判断するようなテキストには、より高いスコアを割り振るように学習することで、報酬モデルは人間の代わりにテキストの好ましさを算出するモデルになる。

Step3: ポリシーモデルの強化学習 Step2 で学習した報酬モデルが算出するスコアに基づいて、PPO などのアルゴリズムで強化学習を行う。目的のモデルはより報酬を得られるような出力を行うように学習することで、最終的に人間の嗜好を学習することができる。

2.2 RLAIF

人間のフィードバックが高コストであることから、性能の良い大規模な言語モデルに人間の代わりにフィードバックラベルを作成させることが試みられており、Reinforcement Learning from AI Feedback(RLAIF) として提案されている [4]。人間のフィードバックを AI フィードバックで代替する方法としては単純であり、RLHF における Step1 のフィードバックの収集の段階を、性能の良いモデルに代わりに行わせるだけである。これにより得られるフィードバックの形態は人間のものと変わりがない（どちらのテキストがより好ましいかという二値的な情報である点では同じである）ため、その後のフィードバックの利用方法も RLHF と RLAIF では同じである。また、計算負荷を気にしないのであれば、AI フィードバックを取得するモデル自体を報酬モデルとして用いることも可能である。

RLAIF を実際に収集した結果、人間のフィードバックデータとの一致率は約 75-80%であり、必ずしも全てが一致するわけではなかった。また、その後の報酬モデルによる AI フィードバックの再現率は 75%程度であり、人間のフィードバックの再現率と同じ程度であった。最終的に AI フィードバックと人間のフィードバックのそれぞれで学習したモデルの出力の好ましさを人手で評価したところ、RLAIF のモデルは RLHF のモデルにほとんど遜色がないことがわかった。このように RLAIF は人間のフィードバックを代替する有効な手段であり、その後も RLCD[5] など様々な研究が続いている。

3 実験

被験者 ID	年齢	性別	正解率 (%)
ZKW	25	女性	69.57
ZDN	32	男性	89.13
ZPH	26	男性	89.13
ZMG	51	男性	91.30
ZAB	41	女性	76.09
ZJN	51	女性	54.34
ZKH	41	女性	76.09
ZGW	49	男性	71.74
ZJS	42	男性	91.30
ZKB	26	女性	89.13
ZDM	25	男性	76.09
ZJM	41	男性	80.43
平均	38	-	79.53

表 1 ZuCo の被験者 12 人の映画レビューに対する 5 段階での感情極性ラベリングの正解率。

3.1 データ

認知フィードバックの有効性の検証を直接的に目的としたデータセットは存在しないため、本稿では Zurich Cognitive Language Processing Corpus (ZuCo)[6] というデータセットに含まれる Sentiment Reading データを利用する。ZuCo は自然読書中の視線と脳波を同時に記録したユニークなデータセットであり、Sentiment Reading のデータはその一部である。これは、多くの映画レビューを収録している Stanford Sentiment Treebank(SST) から「非常に肯定的な文章」「非常に否定的な文章」「非常に中立的な文章」合わせて 400 文をサンプリングしたものである。ZuCo の被験者 12 人は、これらの文章 (映画レビュー) を自然に読み、その最中の脳波と視線を記録されている。

また、実験の際に、ZuCo の被験者は各レビューの感情極性を 5 段階で推測するタスクを課されている。SST に記録されているゴールドラベルと比較した際に、ZuCo の被験者のラベリングの正解率は平均で 79.53%であった。正解率が最も高い被験者は 91.3%で、最も低い被験者は 54.3%であった。その他詳細は表 1 に記載した。このように、タスクに熟練しているわけではない ZuCo の被験者は映画レビューの 5 段階の感情極性ラベルを完全に付与できるわけではなかった。

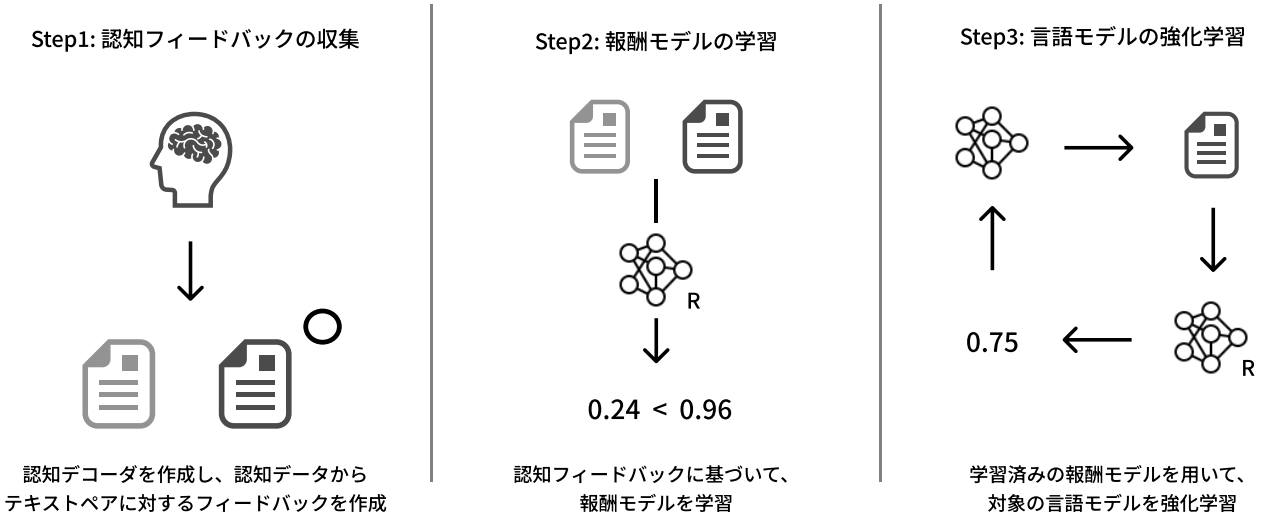

図 1　認知フィードバックの手法の概要。3 つのステップに分けて説明することができる。

3.2　実験設定

認知フィードバックを用いた実験の枠組みとして、図 1 のようなステップで説明する。

Step1: 認知フィードバックの収集　脳波や視線はそのままではフィードバックとして扱うことが難しいため、「認知デコーダ」モデルを訓練する。2 つのテキストを読んだ際の脳波系列データを与えた際に、そのそれぞれにスコアを割り当て、どちらがよりポジティブかを選ぶための認知デコーダを訓練する。認知デコーダは ZuCo に含まれる「非常に肯定的な文章」「非常に否定的な文章」「非常に中立的な文章」の 3 クラスのテキストデータから学習する。具体的には、異なるクラス同士からテキストを 1 つずつピックすることで、1 つのテキストペアができる。予め 8:1:1 にスプリットしておいたデータに対してこれを繰り返すことで、33866 件の訓練用のテキストペアができる。

認知デコーダは、更にテストセットで評価するが、この際のテストセットからは、SST のゴールドラベルを参考に、5 段階の感情極性に基づいたテキストペアの作成を行う。これにより、認知デコーダは「明らかな正例/負例を読んでいる際の脳波、視線によって学習し、実際の運用時にはより曖昧な差を含む正例/負例を判別できるか」が問われることになる (つまり、認知デコーダは教師データとして明示的なフィードバックを必要としない)。このテストセットで得られたフィードバックを「認知フィードバック」とする。今回はデータの全体数が少ないため、この一連の学習、評価 (収集) を 10 分割交差検証を行うことで、映画レビュー全件に対して認知フィードバックを得ることとする。

認知デコーダには、小規模な Transformer によるデコーダ [7] を採用した。

Step2: 報酬モデルの学習　Step1 で収集した認知フィードバックと STS のゴールドデータ (5 段階の感情極性から作成したテキストペアについて) を教師データとして、報酬モデルを学習する。報酬モデルとして事前学習済みの GPT2[8] の 124M パラメータサイズを使用した。学習には TRL[9] を用いた。

Step3: 強化学習を用いた言語モデルの調整　Step2 で学習した報酬モデルを用いて、認知フィードバックを元に作成した「よりポジティブな映画レビューを生成するモデル」を学習することができる。ただし、Step3 については、本稿ではまだ実験を行なっていない。

4　結果

4.1　認知デコーダ

表 2 は認知デコーダのテストセットでの結果である。最も性能が高かったのは脳波のみをデコーダの入力とした場合であり、脳波と視線の組み合わせ、また視線のみの場合の設定では性能が低くなった。これはおそらく、視線データがあまり映画レビューの感情極性に効かない特徴であるか、あるいはテキストペアを作成した際の大幅なデータ拡張によって過学習を起こしやすい状態になっている、などの理由が考えられる。しかし、脳波のみを入力とした場合のデコーダの性能は 70%程度であり、人間のラベ

入力データ	Accuracy - SST Gold Label (%)
視線	54.5
脳波+視線	62.8
脳波	**70.4**

表 2　認知デコーダのテストセットでの結果。ゴールドの感情極性ラベルとの一致度で評価。

フィードバック	Accuracy(%)
SST Gold Label	**81.8**
視線	49.6
脳波+視線	60.9
脳波	73.1

表 3　報酬モデルのテストセットでの結果。元のフィードバックデータの再現率で評価。

リングの精度には及ばないものの、訓練時には与えていない「肯定的な文章同士の比較」「否定的な文章同士の比較」にもある程度汎化できている可能性がある。データを更に増やすことで、より頑健な認知デコーダを作成できることが期待される。

4.2　報酬モデル

続いて、表 3 は報酬モデルのテストセットでの結果である。まず、最も報酬モデルによる再現率が高かったのは、SST データセットから抽出したゴールドラベル (5 値の感情極性) に基づいて作成したテキストペアのフィードバックであった。報酬モデルはサイズはあまり大きくないものの事前学習済みのモデルであり、基本的な言語理解の能力を獲得していると考えられる。そのため、ゴールドラベルの再現率が高いということは、事前学習済みの報酬モデルにとってもより自然なフィードバックの内容であったことが伺える。

また、次に報酬モデルによる再現率が高かったのは、脳波のみを用いた認知デコーダが作成したフィードバックであった。これは、報酬モデルにとって、ゴールドラベルの次に自然なフィードバックの内容であった可能性が高い。認知デコーダの性能検証の結果と併せて考えると、脳波を利用した感情極性のフィードバックの作成はある程度上手く機能しているようである。

その次に再現率が高かったのが脳波と視線、ほぼチャンスレベルとなってしまったのが視線によるフィードバックであった。認知デコーダの性能検証の結果と併せて考えると、少なくとも現段階では視線だけを利用して感情極性のフィードバックを作成することはできていないようである。また、脳波と視線をあわせて用いた際に性能が下がっているのは、視線のデータがノイズとして学習を妨げてしまっているからだと考えることができる。

5　考察

今回の検証では、最後の強化学習のステップのみを除いて、認知デコーダを作成することによる認知フィードバックの収集、また、その認知フィードバックを用いた報酬モデルの学習を行った。入力する認知データについて、視線に関してはノイズのように働いてしまう結果となったが、脳波に関しては、認知デコーダについても、報酬モデルについても、ある程度の性能が確認された。人間の意識的で明示的なフィードバックを完全に代替する、あるいはそれを改善するまでの高い精度は確認されなかったが、学習データをより大きくし、それぞれのモデルの規模も大きくなることで、より頑健な認知フィードバックの作成ができる可能性が示された。

6　おわりに

本稿では、強力な大規模言語モデルの成功を支えている RLHF の技術に着目し、脳波や視線といった非明示的な人間のフィードバックによってそれらを代替することができるのか、それを検討するために小規模な実験を行なった。脳波を用いた実験ではある程度適切なフィードバックを再現することができたが、人間の明示的なフィードバックを改善するほどの性能にはなっていない。ただ、必ずしもゴールドラベルと一致する必要はないのであり、今後は今回作成した報酬モデルを使って言語モデルを実際に強化学習することで、この手法の頑健性を検証していきたい。

謝辞

本研究は、JST さきがけ JPMJPR21C2 の支援を受けたものです。

参考文献

[1] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 3008–3021. Curran Associates, Inc., 2020.

[2] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of chatgpt-related research and perspective towards the future of large language models. **Meta-Radiology**, Vol. 1, No. 2, p. 100017, September 2023.

[3] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744, 2022.

[4] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. **arXiv preprint arXiv:2309.00267**, 2023.

[5] Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. Rlcd: Reinforcement learning from contrast distillation for language model alignment. **arXiv preprint arXiv:2307.12950**, 2023.

[6] Nora Hollenstein, Marius Tröndle, Martyna Plomecka, Samuel Kiegeland, Yilmazcan Özyurt, Lena A Jäger, and Nicolas Langer. The zuco benchmark on cross-subject reading task classification with eeg and eye-tracking data. **Frontiers in Psychology**, Vol. 13, p. 1028824, 2023.

[7] Alex Murphy, Bernd Bohnet, Ryan McDonald, and Uta Noppeney. Decoding part-of-speech from human EEG signals. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2201–2210, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[8] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[9] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.