Cognitive Preference Optimization: 脳情報による言語モデルの選好最適化

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

概要

近年の大規模言語モデルの性能向上において、Direct Preference Optimization (DPO) によるモデルのアラインメントの成功が大きな役割を果たしている。しかし、学習で用いられるラベルの作成には多大なコストが必要であり、作業者の負担も大きい。本研究では、作業者がテキストを読んでいる際の脳波から選好情報を抽出し、それに基づきモデルを学習する手法として Cognitive Preference Optimization (CPO) を新たに提案する。CPO では、作業者はテキストを読むだけでラベルを作成でき、負担の軽減が期待できる。人間のフィードバックを用いた場合と比較して、手法の妥当性を検証した。

1　はじめに

大規模言語モデルの開発において、安全でユーザに好まれる出力を生成するためには、モデルが人間の意図や価値観に沿って動作するように調整するアラインメントが不可欠である。事前学習済みモデルに対する事後学習としては、Supervised Fine-tuning (SFT) を行い、さらに人間の選好を取り入れる Direct Preference Optimization (DPO) [1] による最適化を行うという流れが主流であり、最先端のモデルでもこれらの手法が変わらず採用されている [2]。しかし、DPO などのポリシー最適化手法で必要とされる選好ラベルの作成には大きな負担が伴い、作業者の選抜や教育、作業者との信頼関係の構築など、長期的なコストがかかる点は大きな課題である [3]。このような状況を踏まえ、近年では AI Feedback[4] のように、従来の人手によるフィードバックを補完・代替する手法も開発されてきた。本研究では、これらのアラインメントの根幹となる「人間の選好情報」の取得方法について、脳活動データを活用できないかという新たな観点から検討を行う。特に、DPO などのポリシー最適化手法を利用する場合においても、脳波から得られる選好情報はモデル学習に活用できるのかを検証する。

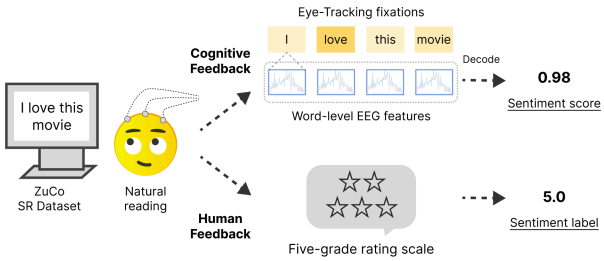

図 1: ヒューマンフィードバックと認知フィードバックの概要図

具体的には、DPO が最初に提案された際のベンチマークタスクの一つである感情コントロール生成（Controlled Sentiment Generation）を対象とし、被験者がテキストを読んでいる際に計測される脳波を用いてモデルの選好最適化を行う手法として Cognitive Preference Optimization (CPO) を提案する。CPO では、人間がテキストに対して直接ラベル付けを行うのではなく、脳波計測によって推定された選好情報を利用するため、従来の人手によるフィードバックを大幅に削減できる（図 1）。本研究では、従来の人間が明示的に作成するヒューマンフィードバックと、脳活動から作成される暗黙的な認知フィードバック、二つのフィードバックに基づく学習手法との比較を行うことで、その妥当性を検証した。実験の結果、提案手法を用いたモデルは、ベースラインモデルと比較してよりポジティブな出力を生成するポリシー最適化が可能であることが示された。これらの結果は、LLM のアラインメントにおける新たなアプローチを示唆するものである。

2 関連研究

自然言語処理において、人間の生理的な情報を活用してモデルの性能を向上させようとする研究は Cognitively Inspired NLP と呼ばれる分野として数多く報告されている。これらの研究では、脳波や視線などから得られる人間の生理データを利用することで、さまざまな下流タスクにおいて性能改善が得られることが示されている。具体的には、品詞解析 [5]、依存構造解析 [6]、感情分析 [7]、固有表現認識 [8, 9]、関係抽出 [10] など非常に幅広いタスクに適用されており、生理データによる補助情報はタスクに依存せず有益な効果をもたらすことが示唆されている。一方で、これらの研究の多くは LSTM ベースのモデルを対象としている。近年主流となっている Transformer ベースの大規模言語モデルへの適用例は限定的であるうえ、適用のためにアーキテクチャの変更が必要 [11] であり、既存の事前学習済モデルを活用できない。本研究では、事前学習済みモデルのアーキテクチャを直接変更することなく、人間の生理データから選好情報を作成する手法を提案する。事後学習の段階で生理データを用いることにより、既存の大規模言語モデルにも適用可能な手法として機能することが期待される。

3 手法

3.1 Cognitive Preference Optimization

手法全体の流れを図 2 に示す。

Step1: 脳波デコーダの学習 まず、ある文書を読んでいる際の EEG 特徴系列を $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ と表し、各 $\boldsymbol{x}_t \in \mathbb{R}^m$ は文書中の単語 t に対応する EEG 特徴ベクトルとする。文書の単語数を T とする。脳波デコーダはこの EEG 系列を入力とし、スカラー値を出力する関数 $s_\phi(\boldsymbol{X}) \in \mathbb{R}$ として定義する。ここで，ϕ はデコーダの学習パラメータである。学習サンプルとして，「好ましい (chosen)」と「好ましくない (rejected)」の文書ペアを $\left\{(\boldsymbol{X}^{(i)}_{\text{chosen}}, \boldsymbol{X}^{(i)}_{\text{rejected}})\right\}_{i=1}^{N}$ の形で用意する。i 番目のペアに対して脳波デコーダが出力するスコアを $s_\phi(\boldsymbol{X}^{(i)}_{\text{chosen}})$, $s_\phi(\boldsymbol{X}^{(i)}_{\text{rejected}})$ とし，その差を $\text{score\_diff}^{(i)} = s_\phi(\boldsymbol{X}^{(i)}_{\text{chosen}}) - s_\phi(\boldsymbol{X}^{(i)}_{\text{rejected}})$ と定義する。脳波デコーダの目標は，chosen のほうが rejected よりも高いスコアを出力することであり、この条件を満たすように次の損失関数を最小化する：

$$\mathcal{L}(\phi) = \sum_{i=1}^{N} \log\Big(1+\exp\big(-\big[s_\phi(\boldsymbol{X}^{(i)}_{\text{chosen}}) - s_\phi(\boldsymbol{X}^{(i)}_{\text{rejected}})\big]\big)\Big). \tag{1}$$

この損失を最小化することで、脳波デコーダは chosen を高く、rejected を低くスコアリングするよう学習される。

Step2: 認知フィードバックの収集 次に、学習した脳波デコーダを用いて認知フィードバックの収集を行う。Step1 ではデコーダの教師データとして選好データが必要であったが、Step2 では脳波データのみから認知フィードバックを収集することが可能である。具体的には、ある二つの文章に対して脳波デコーダを用いてスコアを算出する。スコアが高い文章を chosen、スコアが低い文章を `rejected` とみなし、これらの文章ペアと対応する選好情報を収集する。この方法により、従来必要とされていた人間による選好ラベル付けの手間を削減することが可能となる。

Step3: 認知フィードバックに基づく DPO Step3 では、Step2 で収集した選好データ（認知フィードバック）を用いて、DPO によるモデルの最適化を行う。DPO は、参照モデルと比較して、好ましい出力が選択される確率を最大化するようにモデルを学習する手法であり、報酬モデルなしで、選好データから直接最適化を行うことができる。具体的には、モデル π_θ の条件付き確率分布と参照モデル π_{ref} の条件付き確率分布を用いて、次の損失関数を最小化する：

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log\sigma\left(\beta\left(\log\frac{\pi_\theta(\boldsymbol{y}^{(i)}_{\text{chosen}}|\boldsymbol{x}^{(i)})}{\pi_{\text{ref}}(\boldsymbol{y}^{(i)}_{\text{chosen}}|\boldsymbol{x}^{(i)})} - \log\frac{\pi_\theta(\boldsymbol{y}^{(i)}_{\text{rejected}}|\boldsymbol{x}^{(i)})}{\pi_{\text{ref}}(\boldsymbol{y}^{(i)}_{\text{rejected}}|\boldsymbol{x}^{(i)})}\right)\right), \tag{2}$$

これにより、Step2 で得られた認知フィードバックを活用し、生成モデルを被験者の選好に適合させることが可能となる。

3.2 選好ペアデータの作成

使用データセット 本実験では，データセットとして Zurich Cognitive Language Processing Corpus (以下 ZuCo) [12] に収録されている Sentiment Reading (SR) データセットを用いた。このコーパスはアイトラッキングと EEG の記録を同時に行ったデータセットであり、12 人の被験者によって読まれた約 400 文を含む。文章は全

て映画のレビュー文であり、Stanford Sentiment Treebank (SST) から抽出されたものである。脳波は視線データ（Gaze Duration）に基づいて抽出しており、840 次元のベクトルとして使用する。

ペアデータへの変換　本研究では、ZuCo SR データセットを用いてペアデータを作成する。このデータセットには、SST（Stanford Sentiment Treebank）[13] が作成された際に付与された Ground Truth ラベルが含まれており、400 文のうち 140 文が positive、137 文が neutral、123 文が negative に分類される。データリークを防ぐため、これらのラベル分布を維持したまま、10-fold に分割してペアデータを作成する。ペア作成の際には、positive > neutral の関係、さらに positive > negative の関係に基づいてペアを構築する。理論上、positive な文と neutral な文の組み合わせ数、あるいは positive な文と negative な文の組み合わせ数に基づき最大のペアを作成可能であるが、訓練時には各文につき最大 5 ペアのみを作成する。これにより、データ全体のバリエーション低下による過学習を防ぐ。一方、テスト時には可能な限りすべてのペアを生成する。脳波デコーダの学習は 10 分割交差検証により行い、各 fold のテストデータにおける結果である全 3640 件を認知フィードバックとして使用する。

ヒューマンフィードバックの取得　ZuCo SR データセットに含まれる 400 文のうち、47 文には被験者が回答した感情ラベル（5 段階評価）が記録されている。これらの 47 文の内訳は、Ground Truth ラベルで positive が 22 文、neutral が 6 文、negative が 17 文である。Ground Truth のラベルに基づき、これらの 47 文から計 506 ペアを作成する。作成されたペアに対して、被験者が付与した 5 段階の感情ラベルを基に選好情報を構築し、これをヒューマンフィードバックとして利用する。なお、ヒューマンフィードバックは認知フィードバックに比べてデータ数が少ないため、両者を比較する際には、認知フィードバックデータの中から同じ 506 件のテキストペアを抽出して使用する。これにより、データが異なることによる影響を排除し、フェアな比較を行う。

3.3　学習設定

使用モデル　脳波デコーダとして、[5] を参考に、小規模な Transformer を用いた。ハイパーパラメータは、事前にグリッドサーチを用いて決定した。num layers は 2、nhead を 8、dmodel を 512、dim feedforward を 2048 に設定して実験を行った。また、ポリシーモデルとして、IMDb データセット [14] でファインチューニング済みの GPT-2 medium[1] を用いた。以降これを SFT モデルと呼ぶ。IMDb データセットは SST と同様に映画レビューで構成されているが、両者は異なるデータを基に作成されおりデータリークの懸念はない。

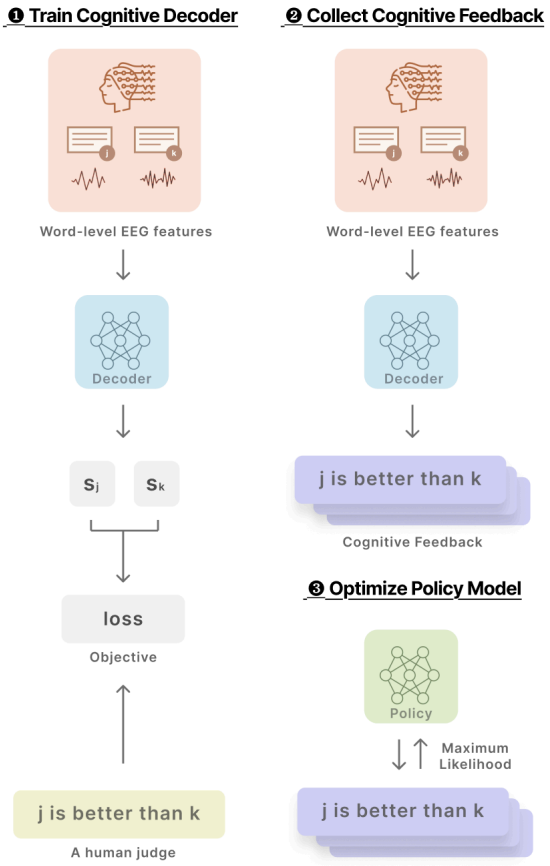

図 2: 提案手法の全体の流れの概要図。脳波データのデコーダを学習する際には教師データとして選好ラベルが必要になるが、認知フィードバックは脳波のみから収集することができ、作業者はテキストを読むだけでよい。

ポリシーモデルの評価　作成したポリシーモデルの評価では、SST データセットの中から、ZuCo SR データセットに収録されていない例を、neutral 文を 50、negative 文を 50 抽出し、それらの初期 10 単語のみポリシーモデルの入力として与え、続きを生成させた。生成させた文章について、gpt-4o-2024-11-20 による llm-as-a-judge を実施した。よりポジティブな文章を出力したモデルを選ばせ、それに基づく 2 モデル間の Win

1）　https://huggingface.co/edbeeching/gpt2-medium-imdb

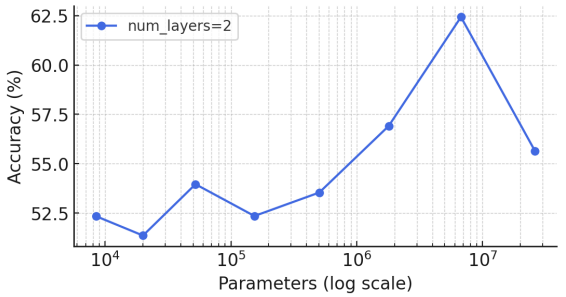

図 3: モデルサイズごとの脳波デコーダの性能

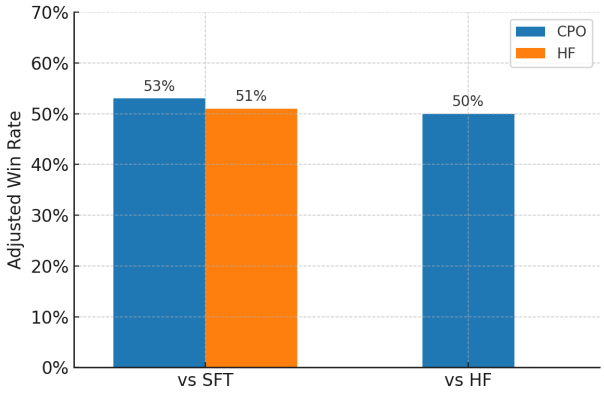

図 4: 認知フィードバックとヒューマンフィードバックの性能比較

Rate を算出することで、ポリシーモデルの評価を行った。

4 結果

脳波デコーダの性能 モデルサイズごとの脳波デコーダの性能の結果の一部を図 3 に示した。適切なモデルサイズにおいては Ground Truth に対する正解率がチャンスレベルを大きく上回っており、脳波の持つ情報からある程度適切に選好情報がデコードできていることを示している。

認知フィードバック vs. ヒューマンフィードバック 認知フィードバックと人手で作成されたフィードバックとで、それぞれ作成されたポリシーモデルの性能比較を図 4 に示した。二つのポリシーモデルを直接比較した場合だと差が出なかったが、ベースラインである SFT モデルとの比較では、CPO により作成したモデルがわずかに上回った。それぞれのフィードバックタイプにおいてどのように異なるのかを、今後の分析で明らかにしたい。

提案手法のスケール性の検証 今回、ポリシーモデルの訓練に使用できる脳波から作成した選好ペアは合計で 3640 件であり、比較的小規模である。学習に使用するデータの件数が多い

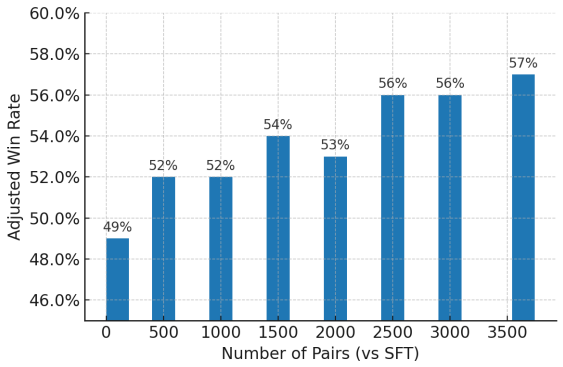

図 5: ポリシーモデルの学習に用いる選好ペアの数と、そのペア数で学習したモデルの性能の推移

ほど最終的なポリシーモデルの性能も良くなるのであれば、今後さらに多くの脳波データを収集することにより、さらに性能の高いモデルが作成できると期待される。図 5 ではペア数をコントロールして作成した複数のポリシーモデルの性能を比較しており、提案手法のスケール性の検証を行なった。今回の検証の範囲内ではより多くのペアを用いるほどポリシーモデルの性能も改善しており、より多くの脳波データがさらなる改善をもたらす可能性があることを示唆する結果となった。

5 おわりに

本研究では、脳波データを用いて選好情報を収集し、大規模言語モデルをポリシー最適化するための新しい手法を提案した。具体的には、脳波データを基に選好情報を推定する脳波デコーダを構築し、これを活用して認知フィードバックを収集し、DPO (Direct Preference Optimization) に基づくモデルの最適化を行った。実験の結果、提案手法を用いたモデルは、ベースラインモデルと比較してよりポジティブな出力を生成するポリシー最適化が可能であることが示された。また、手動で作成したフィードバックを用いた場合と比べても、提案手法の性能は同等程度であることが確認された。今後の課題として、生成された出力例のさらなる分析を行い、提案手法がどのようにモデルの選好に影響を与えるのかを詳細に検討する必要がある。また、適用範囲の拡大や、より大規模なデータを用いた実験も行い、手法の妥当性をより深く検証していきたい。

謝辞

本研究は，JST さきがけ JPMJPR21C2 および JSPS 科研費 24H00087 の支援を受けたものです.

参考文献

[1] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

[2] Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Heylar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. Deliberative alignment: Reasoning enables safer language models. **arXiv preprint arXiv:2412.16339**, 2024.

[3] 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. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 3008–3021, 2020.

[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 e-prints**, pp. arXiv–2309, 2023.

[5] 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.

[6] Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez. Towards making a dependency parser see, 2019.

[7] Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. Sequence classification with human attention. In Anna Korhonen and Ivan Titov, editors, **Proceedings of the 22nd Conference on Computational Natural Language Learning**, pp. 302–312, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[8] Nora Hollenstein and Ce Zhang. Entity recognition at first sight: Improving NER with eye movement information. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 1–10, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Yuqi Ren and Deyi Xiong. CogAlign: Learning to align textual neural representations to cognitive language processing signals. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 3758–3769, Online, August 2021. Association for Computational Linguistics.

[10] Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. Advancing nlp with cognitive language processing signals, 2019.

[11] Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. Improving natural language processing tasks with human gaze-guided neural attention. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 6327–6341. Curran Associates, Inc., 2020.

[12] Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. **Scientific data**, Vol. 5, No. 1, pp. 1–13, 2018.

[13] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[14] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.