量子計算を用いた文字言語モデル

三輪　拓真[1,2]

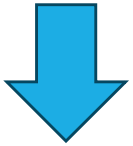
[1]奈良先端科学技術大学院大学

[2]理化学研究所ガーディアンロボットプロジェクト

研究の概要

背景

- Chat-GPT等の大規模言語モデルはモデルの大規模化によって成立
- 訓練時間やコストの観点から、学習の効率化は利益大



研究内容

- 量子計算を用いて言語モデルの効率化ができるか検証
- 小さなモデルを作成し、正常に訓練できることを確認

研究背景

- 近年大規模言語モデル(LLM)により、自然言語処理領域は大きく発展
- モデルの大規模化は精度の向上に寄与[1]
- パラメータ数が一定値を超えることで、モデルが新たな能力を得ることも示唆[2]
- モデルの大規模化は新たな価値提案に直結

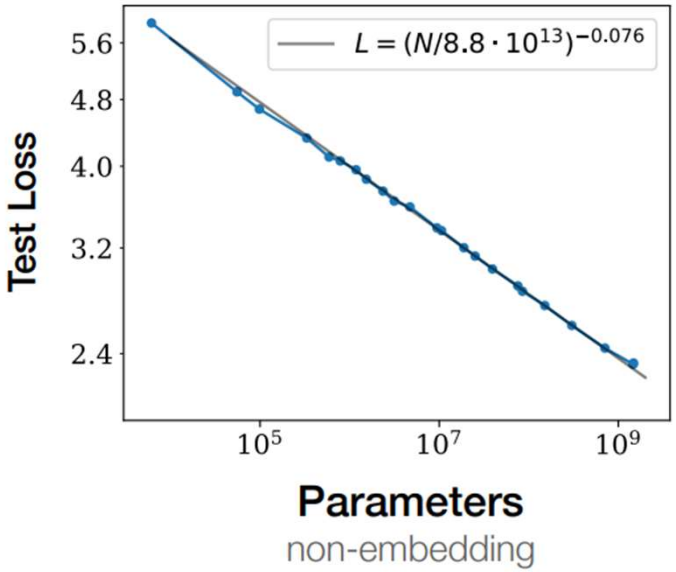


図1: パラメータ数とロス値の関係[1]

[1] Kaplan, Jared, et al. "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361 (2020).
[2] Wei, Jason, et al. "Emergent abilities of large language models." arXiv preprint arXiv:2206.07682 (2022).

研究背景

- 一方でハードウェアの制約等により、大規模化は有限
- 計算効率の良い大規模モデルの必要性大[3]
- 量子計算で言語モデルを実現することでこれらの問題を解決できる可能性に取り組む

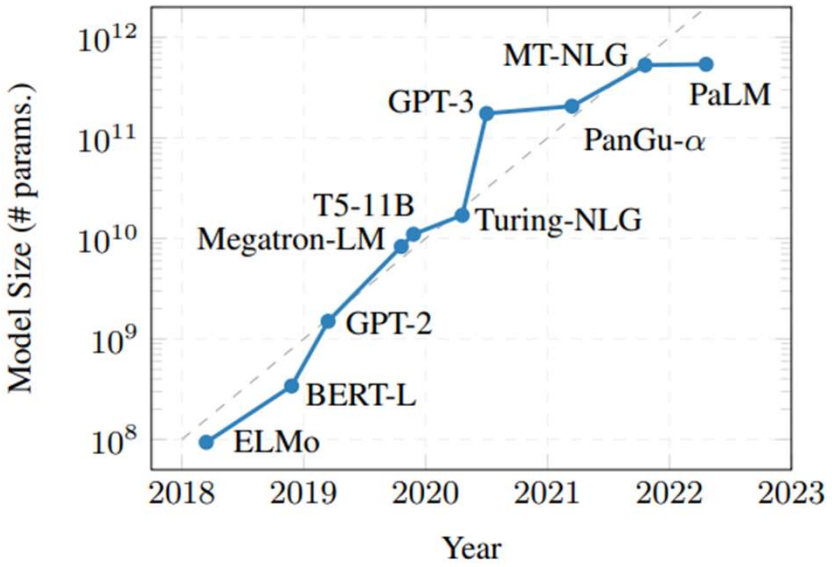


図2: モデルサイズの推移[3]

[3] Treviso, Marcos, et al. "Efficient methods for natural language processing: A survey." Transactions of the Association for Computational Linguistics 11 (2023): 826-860.

量子計算の性質

- 0と1に対応した量子の状態$|0\rangle, |1\rangle$及びその重ね合わせを用いて演算
- 各量子ビットの状態を量子のもつれとして相互に関係付

	長所	短所
普通のコンピュータ	• 安定した出力	• パラメータ数に対して計算時間・メモリが大きく増加
量子コンピュータ	• 高いメモリ効率	• 計算誤差大 • 現状大規模化が困難

図3: 量子計算の特徴

量子アルゴリズムの関数値評価(1/2)

- 各量子ビットは状態$|0\rangle, |1\rangle$の重ね合わせによって存在
- よって量子ビットの状態$|\psi\rangle$は複素数α，βを用いて，

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$$

- 量子ビットの状態は2つの係数で表せるため，2次元ベクトルで書き換え可能

$$|\psi\rangle = \begin{bmatrix}\alpha\\ \beta\end{bmatrix}$$

- 複数ビットの状態は各量子ビットのテンソル積⊗によって表現され，$|\varphi\rangle = \begin{bmatrix}\gamma\\ \delta\end{bmatrix}$とすると，

$$|\psi\rangle\otimes|\varphi\rangle = \begin{bmatrix}\alpha\\ \beta\end{bmatrix}\otimes\begin{bmatrix}\gamma\\ \delta\end{bmatrix} = \begin{bmatrix}\alpha\begin{bmatrix}\gamma\\ \delta\end{bmatrix}\\ \beta\begin{bmatrix}\gamma\\ \delta\end{bmatrix}\end{bmatrix} = \begin{bmatrix}\alpha\gamma\\ \alpha\delta\\ \beta\gamma\\ \beta\delta\end{bmatrix}$$

量子アルゴリズムの関数値評価(2/2)

- 実際の量子計算は図4のような回路図によって実行
- 初期状態$|x\rangle$，$|y\rangle$を量子ゲートU_fによって変更し，各ビットの状態を測定
- ベクトル表現中では，量子ゲートU_fはユニタリ行列として表現

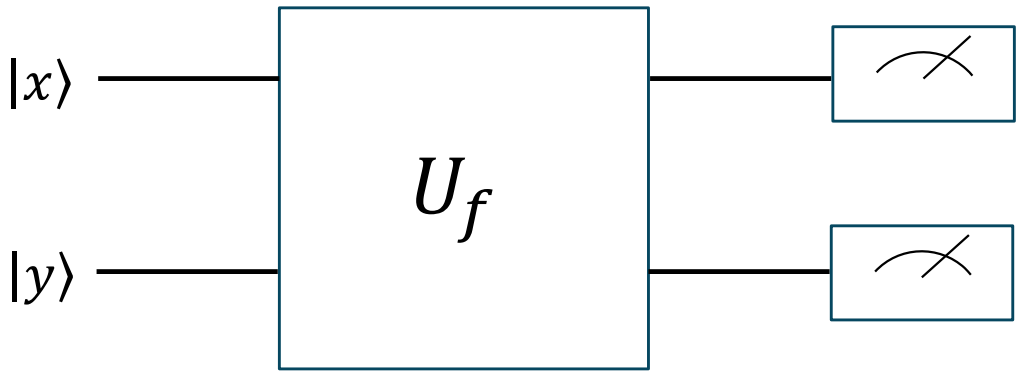


図4:量子ゲートU_fの回路図

量子計算による計算効率の改善

- 量子ビット$n\ bit$に対して、$2^n\ bit$の表現が可能

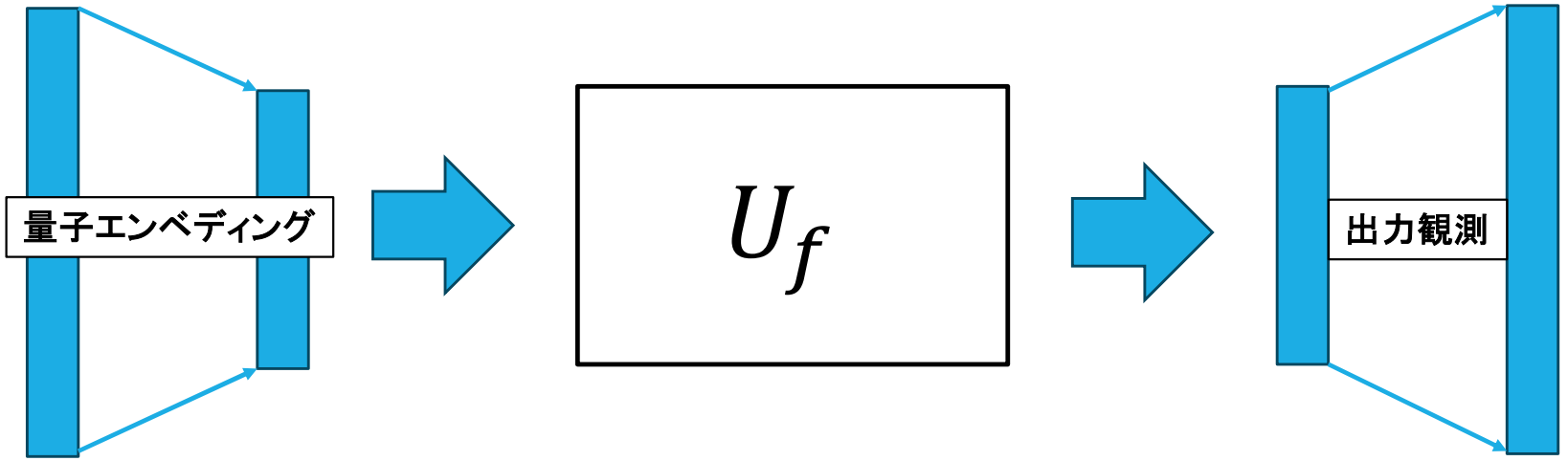


図5:量子計算の適用例

⇒U_f部分の計算に必要なメモリを削減

言語モデルの研究背景

- 近年言語モデルは機械翻訳や質問応答、テキストの要約等にも応用可能
- 文脈に応じて適切に行われるテキスト生成が基盤
- 大規模言語モデルは文章の一部を隠して再度予測する自己教師あり学習により文脈に応じた単語の予測を学習
- 大規模言語モデルの代表であるデコーダモデルはMulti-head Attentionに依存し、計算量大
- 精度を維持しつつ計算効率を向上させる必要性

言語モデルの定式化

- 言語モデルは文の構成要素の並びに対して確率を付与
- 本研究では直前のn文字$c_{t-n}, c_{t-n+1}, \cdots c_{t-1}$から次の文字$c_t$の出現確率$P(c_t|c_{t-n}, c_{t-n+1}, \cdots c_{t-1})$を予測
- モデルの予測性能の指標であるパープレキシティPPを以下のように算出．ただし$|j|$は全文書の文字数の合計であり，Nは文書数

$$H = -\frac{1}{|j|}\sum_{N}\sum_{t} logP(c_t|c_{t-n}, c_{t-n+1}, \cdots c_{t-1})$$

$$PP = 2^H$$

- パープレキシティによって作成したモデルの性能を評価

パラメータ付き量子回路

- パラメータ付き量子回路は，量子状態に与える変化をパラメータとして指定可能
- 出来合いの回路より繊細な演算が可能
- 一方で経験則によるパラメータの最適化は困難
- 効率的な最適化アルゴリズムが必要

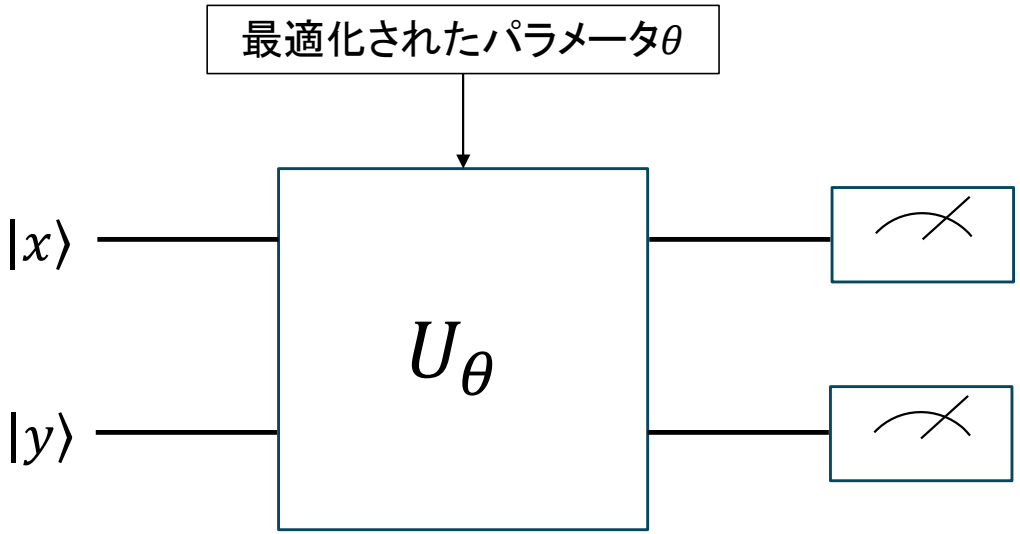


図5:量子ゲートU_θの回路図

量子古典ハイブリッドアルゴリズム

- パラメータ探索には従来のコンピュータのアルゴリズムが応用可能
- この手法は量子古典ハイブリッドアルゴリズムと呼ばれる
- 特に量子ニューラルネットワーク等の従来のコンピュータに近いアーキテクチャを実装する手法を量子機械学習と呼ぶ
- 量子機械学習を用いた言語モデル実装を目指す

量子機械学習の例

- 量子状態の位相をずらす回転ゲートR_x, R_y及び，各量子ビットに依存関係をつくるCNOTゲートを用いて構築

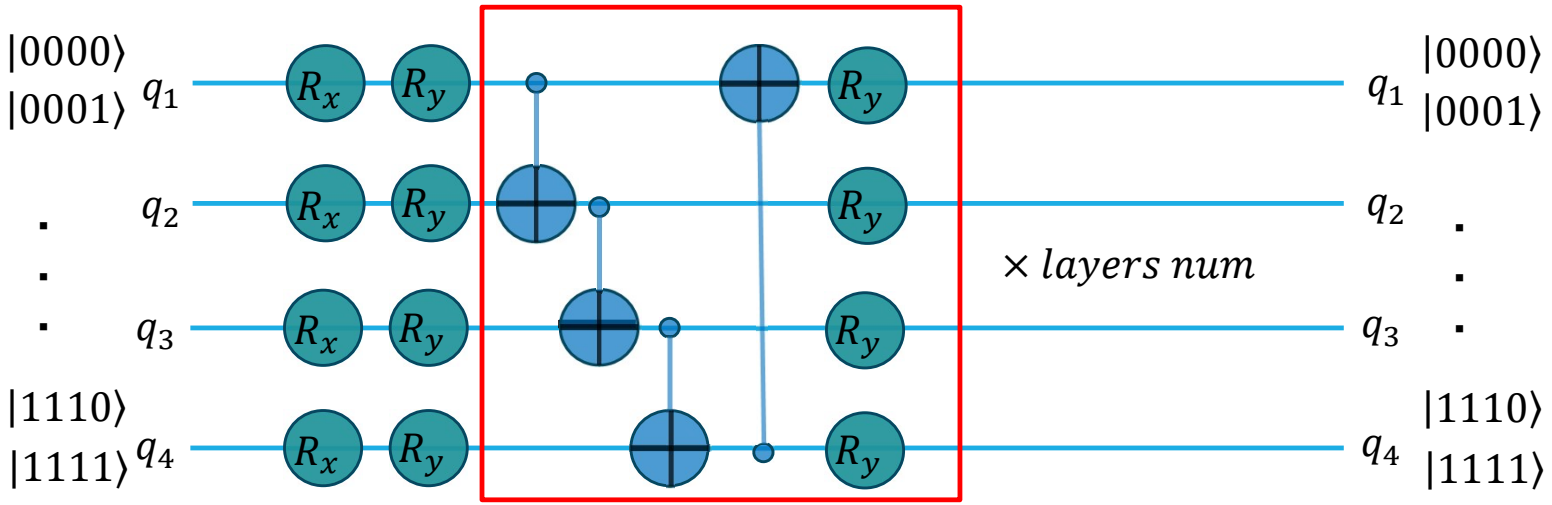


図6: 量子ニューラルネットワーク(QNN)の例[4]

[4] Schuld, Maria, et al. "Circuit-centric quantum classifiers." *Physical Review A* 101.3 (2020): 032308.

量子計算による文字n-gramモデルの実装

- 図7ような文字n-gramモデルを実装
- QNNによる実装
- 入出力はone-hot
- α, βはそれぞれ複素数であり，量子状態$|\psi\rangle$の係数を表現
- $argmax_i|\beta_i|^2$から単語ベクトルを復元し，順次単語予測を実行

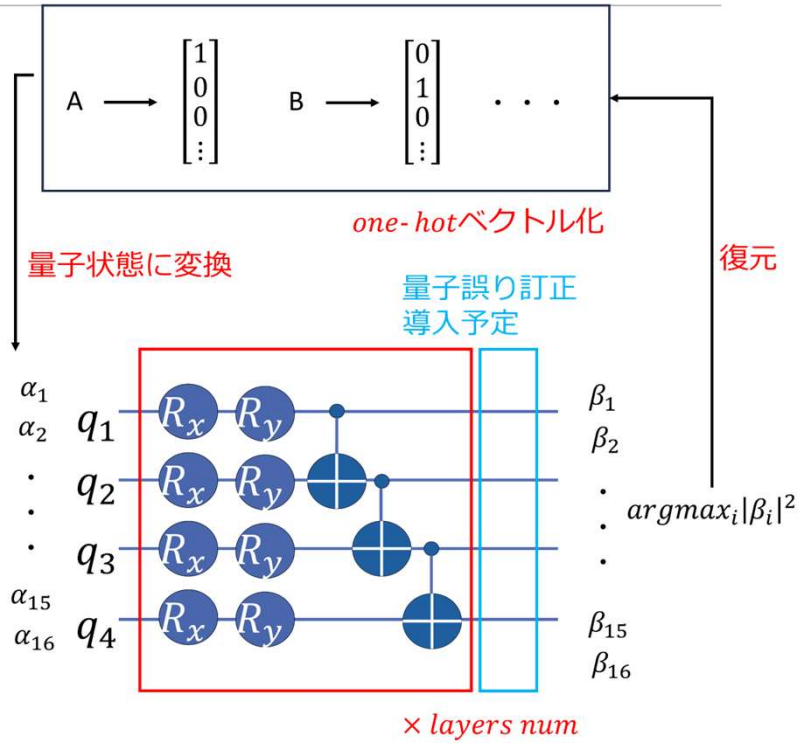


図7: QNNによる言語モデル

テストセットパープレキシティによる評価

【実験結果】

- 文字種類64の辞書を使用
- Wall Street Journalのテキストデータを使用
- テストデータでパープレキシティを算出
- ベースラインとして最尤推定で構築したn-gramモデル使用
- 単純なQNNでもパープレキシティの削減に効果有

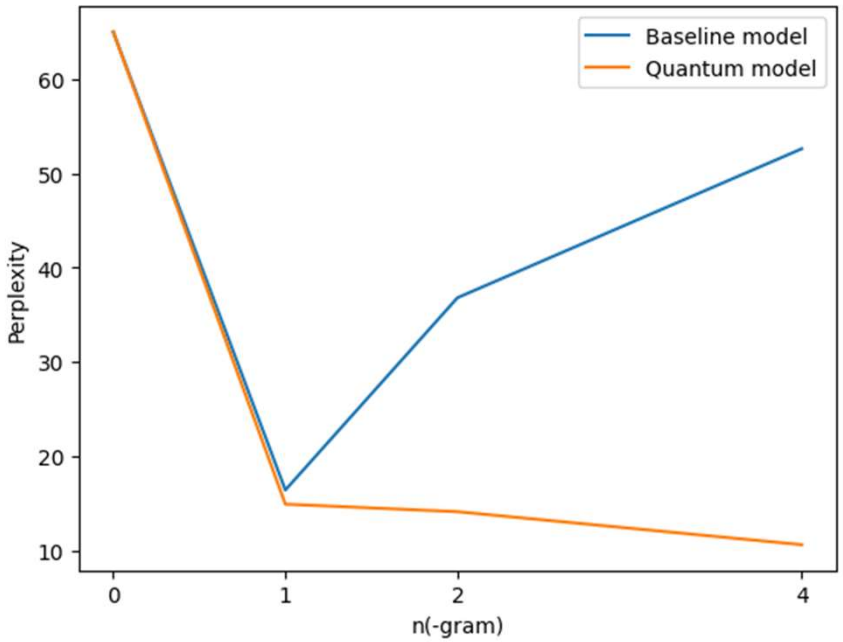


図8: QNNモデルと最尤推定モデルによるPP比較

現状の課題と今後の方針

- 近年のTransformerはAttention機構の文脈の推定により高い精度を発揮
- 一方QNNモデルでは文脈まで意識することができない
- 量子self-attention[5]による文脈の推定

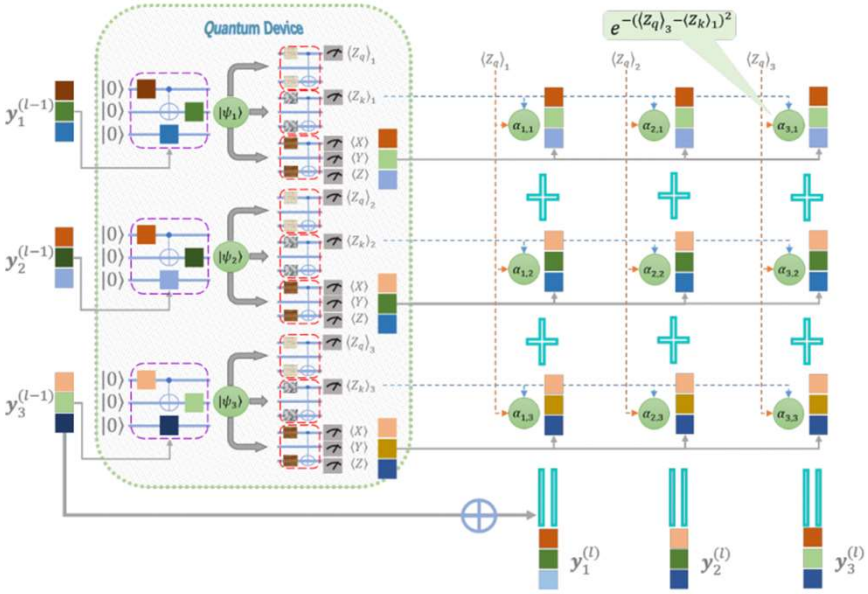


図9: 量子self-attentionの概要

[5] Li, Guangxi, et al. "Quantum self-attention neural networks for text classification." *arXiv preprint arXiv:2205.05625 (2022).*

現状の課題と今後の方針

- 本研究はシュミレータを用いて実行
- ハードウェアの制約から実機を用いた実験には多くの計算誤差が伴う
- アルゴリズムによる誤差抑制へのアプローチ[6]を適用

[6] Acharya, et al. "Suppressing quantum errors by scaling a surface code logical qubit." *Nature, Vol. 614, No. 7949, pp. 676–681 (2023).*

まとめ

- 量子機械学習を用いて文字n-gramモデルを実装
- 最尤推定のn-gramモデルと精度を比較して性能を評価
- 最尤推定モデルでは予測できないケースも正しく予測できることが分かった

インターン活動(学部)

社名： Koozyt, Inc(2021/3 ～ 2023/3)
特筆すべき業務例：
音声認識モデルの後処理実装

- 技術調査から実装方式の提案・実装まで一貫して参加
- メモリ効率や計算時間を重視する要望に応えるため、精度重視の機械学習モデル案を破棄して正規表現を用いて実装
- 正規表現でも精度を最大限上げるため、慎重にデータ分析を実行
- 最終的に精度を数%しか落とさず、計算時間を約3倍短縮

インターン活動(修士)

社名： Fastlabel(2024/1/22 ～ 2024/2/2)
特筆すべき業務例：
大規模言語モデルの継続事前学習

- 技術調査から実装まで一人で実行
- 日本語トークンを用いたStableLMの継続事前学習を実行
- インターン期間では学習が不十分だったため、精度は低かった
- またSwallowモデルと比較すると、学習データの品質も十分に上げることが出来なかった

インターン活動(修士)

社名： 株式会社PKSHA Technology(2024/2/5 ～ 2024/3/1)
特筆すべき業務例：
対話ログからFAQの生成を行うモデルの精度改善

- GPTモデルを用いたFAQ生成アルゴリズムを改善
- GPTモデルへ入力するプロンプト及び、その前後のテキスト処理を改善
- 生成されたFAQを目視にて確認し、約60%のケースで出力が改善していることを確認

その他の発表経験

- Hult Prize岡山大学大会
  - SDGs解決に向けたビジネスコンテストにおいて、チームリーダーを担当
  - 昆虫食のアイデアを家畜の育成に導入し、飼料を作る場所を減らすことによるコストカット案を提案し、準グランプリを獲得
- 技育展：”量子コンピュータシュミレータ”
  - 行列計算を用いて量子計算中の量子状態をシュミレートするモジュールを作成
  - 同モジュールで量子機械モデルを実装し精度を測定
    https://github.com/TakumaMiwa/study/blob/main/%E6%8A%80%E8%82%B2%E5%B1%95_%E9%87%8F%E5%AD%90%E3%82%B3%E3%83%B3%E3%83%94%E3%83%A5%E3%83%BC%E3%82%BF%E3%82%B7%E3%83%A5%E3%83%9F%E3%83%AC%E3%83%BC%E3%82%BF.pdf
- NLP若手の会 ハッカソン："プロンプト当てゲーム“
  - Open APIを用いた画像編集の前後から、編集に用いたプロンプトを当てるゲームを作成
  - 4人チームで開発し、リーダー兼APIに渡す前後の画像処理部を担当