国立大学法人
電気通信大学

AIXセミナー
AIを使って小説を書いてみよう！

実習編 2日目 （14:15～）

大規模言語モデルを使って小説を書いてみよう

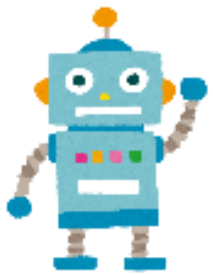
電気通信大学 人工知能先端研究センター
准教授 稲葉 通将

今回行うこと

- 言語生成用深層学習モデルGPT-2 (Generative Pre-trained Transformer) を用いて，文を生成
- さらに，GPT-2を小説っぽい文体が生成できるようにFine-tuning(微調整)
- GPT-2の作った文同士をBERTを使って一貫性を判定し，文をつなげていく
  - BERT：分類・抽出用深層学習モデル

今回行うこと

- シードとなる1文を人が与えるとGPT-2はそれに続く文を複数パターン生成する



①:吾輩のモルモットは死んでいる。

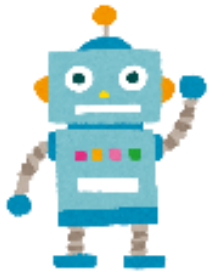
②:我輩の人生は間違いであった。

③:私にはそれがわからない。

④:野良犬なので散歩する場所はない。

今回行うこと

- GPT-2が入力文の続きとなる文章の候補を**人が選ぶ**ことによって自然な文章を作成



①:吾輩のモルモットは死んでいる。

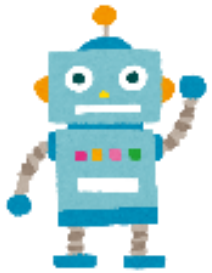
②:我輩の人生は間違いであった。

③:私にはそれがわからない。

④:野良犬なので散歩する場所はない。

じゃあ③で

今回行うこと



①:吾輩のモルモットは死んでいる。

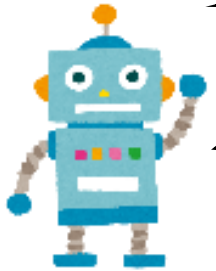
②:我輩の人生は間違いであった。

③:私にはそれがわからない。

④:野良犬なので散歩する場所はない。

吾輩はモルモットである。

じゃあ③で



①:モルモットのように動いてはいられぬ。

②:私がモルモットであるということは、この私のことを知っているということを意味するのだ。

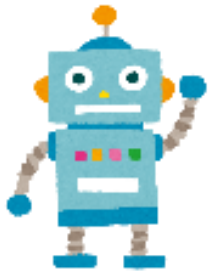
③:私の知る限りモルモットは人間を喰らうとの話は聞いたことがない。

④:俺はモルモットだ、と言うのは、モルモットのようではないと言う事だ。

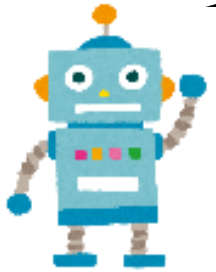
次は①で

今回行うこと

吾輩はモルモットである。



③:吾輩にはそれがわからない。



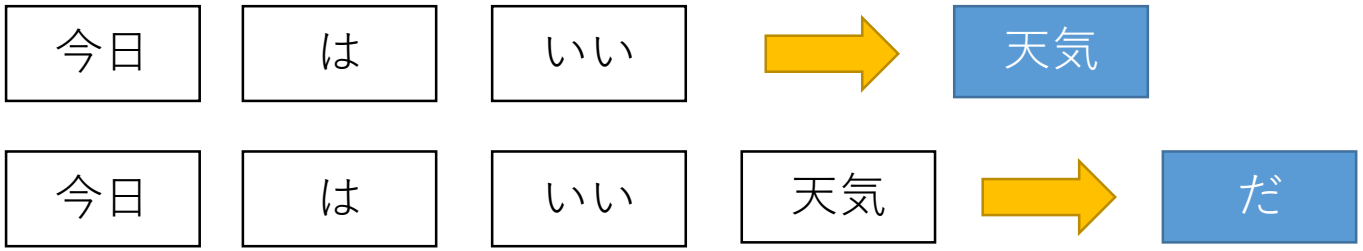
①:モルモットのように動いてはいられぬ。

作成された文章

吾輩はモルモットである。
吾輩にはそれがわからない。
モルモットのように動いて
はいられぬ。

GPT-2とは

- 2019年にOpenAIによって発表された深層学習言語モデル
- 言語モデル
  - それまでの単語列をもとに次の単語を予測するモデル
  - 連続的に次の単語を予測することで文章が生成できる



- 数億という規模の大規模な内部パラメータを持つモデルを,
  大規模なテキストデータ(>100GB)で言語モデルを学習したもの

Fine-tuning

- GPT-2はWebから収集したテキストで学習
  - ニュース，掲示板，Wikipedia, ブログなど様々な文体を含む
- なので，生成される文体はあまり統一されていない
  - 我々としては小説を作りたいので，小説っぽい文体がほしい

Fine-tuning

- 学習済みのモデルを別のデータで追加学習すること
- 今回は小説のデータでFine-tuningすることで小説っぽい文体で文章が生成されるようにする

以降はGoogle Colab上で作業

- まず以下にアクセス
  - https://github.com/1never/UEC_AIX_seminar2021/blob/main/UEC_AIX_seminar2021.ipynb
  - 一番上のOpen with Colabをクリック
- Google Colab上で左上の「ドライブにコピー」をクリック
- プログラムが自分のGoogle Driveにコピーされる

ここから後半

深層学習の適用

- これまではGPT-2の生成文をそのまま使用

問題点

- 意味的に自然ではない文も選択肢に上がってくる

解決策

- 深層学習を使用し，つながりが適切な文のみを選択肢とする
- BERTを使用

BERTとは

- 2018年10月にGoogleが発表した自然言語処理のための深層学習モデル
- それまでのモデルは特定のタスクのデータだけで学習していた
  - 感情推定のタスクには感情推定用のデータ，対話には対話データなど
- BERTは最初に大量の非構造化テキストデータで学習(事前学習)し，その後に特定のタスクのデータで学習(Fine-tuning)する
  - 事前学習でタスクに依存しない言語的な情報を学習
  - Fine-tuningでタスクに依存した情報を学習
- 上記の２ステップにより，複数のタスクで当時最高性能を達成

BERTの事前学習

- 単語穴埋めタスク (Masked Language Model)
  - テキストからランダムに単語を欠落させ，欠落した単語を当てるタスクで学習

    例：「今日 は いい [MASK] です」の[MASK]に入る単語は何か？

- 隣接文推定 (Next Sentence Prediction)
  - 与えられた2文が連続する文か否かを当てるタスクで学習

Input = `[CLS] the man [MASK] to the store [SEP]`
`penguin [MASK] are flight ##less birds [SEP]`
Label = `NotNext`

Input = `[CLS] the man went to [MASK] store [SEP]`
`he bought a gallon [MASK] milk [SEP]`
Label = `IsNext`

今回やること

- 事前学習済みのBERTを使用し，文と文のペアデータでFine-tuning
- GPT-2で多めに文を生成
- BERTを用いて文の間の適切さを判定し，適切さが高い用例を選択肢として表示