

卒業研究報告題目

GPT-2モデルを用いたキーワードからの文章生成
Sentence Generation from Keywords using GPT-2 Model

指導教員 出口利憲 教授

岐阜工業高等専門学校 電気情報工学科

2018E40 森 駿

令和05年(2023年) 2月17日提出

Abstract

In this study, we generated headlines from keywords using GPT-2, a pre-trained natural language processing model that is publicly available. By using a model that can run and be trained on the Google Colaboratory environment, anyone can easily generate headlines that suit them. The generated headline sentences were evaluated by the author himself, a human being who actually uses the headlines, to scrutinize them. Using keywords, the author examined whether the headline text was grammatically correct and attractive, and evaluated its behavior. In this study, we confirmed that even models that do not learn context can generate headlines by combining keywords with each other using only fine tuning and shot learning with a small amount of data. It will be possible to perform larger tasks with a larger model and more data.

目次

Abstract	i
第1章 序論	1
1.1 序論	1
第2章 基本知識	2
2.1 自然言語	2
2.2 自然言語処理	2
2.3 学習手法	2
2.3.1 ファインチューニング	2
2.3.2 ショット学習	3
第3章 関連技術	4
3.1 言語モデル	4
3.1.1 GPT-2	4
3.1.2 Bert	4
3.1.3 Transformer	6
第4章 実験	7
4.1 実験内容	7
4.2 事前準備	7
4.2.1 Colab の設定	7
4.2.2 Google ドライブのマウント	7
4.2.3 GPT-2 のクローン	8
4.2.4 学習データの用意	8
4.3 japanese-gpt2-medium を用いた実験	8
4.3.1 必要ライブラリのインストール	8
4.3.2 ゼロショットでの見出しの生成	10
4.3.3 フューショットでの見出しの生成	10
4.4 ファインチューニングを行った japanese-gpt2-medium での実験	16
4.4.1 ファインチューニング	16
4.4.2 ゼロショットでの見出しの生成	17

4.4.3	フューショットでの見出しの生成	17
4.5	japanese-gpt-1b での実験	22
4.5.1	モデルの読み込み	22
4.5.2	ゼロショットでの見出しの生成	22
4.5.3	フューショットでの見出しの生成	22
4.6	結果考察	27
第 5 章	結論	29
5.1	結論	29
	参考文献	31

第1章 序論

1.1 序論

近年のデジタル機器の発達・普及の勢いは目覚ましく、我が国における情報通信機器の世帯保有率は、モバイル端末全体で97.3%にまでなっている。¹⁾ もはやデジタル機器は生活のあらゆる面においてなくてはならないものとなっており、今後この特徴は世界中に広がっていくであろう。また、生活の大部分がインターネットに接続されたことで、世界中から膨大なデータ蓄積されるようになり、人工知能技術は飛躍的な進歩を遂げるようになった。自然言語処理の分野でも、文章の要約やチャットボット、小説の生成など、多岐にわたる分野で人工知能の研究が行われており。これらの研究はいずれ文章を扱うタスクにおいて大きな役割を果たすことであろう。さて、誰もがデジタル機器を用いてインターネットを利用することができるようになったことで、企業や団体、個人がほぼ同じ条件で情報を発信することもできるようになった。これによってあらゆるコンテンツで供給が爆発的に増加し、現代の社会は多くの分野で買い手市場に大きく傾いている。そのため、コンテンツ1つ1つに見出しなどを設定し顧客の視線を引き寄せることは、自身のコンテンツを人の目に触れさせるためには必須となっている。しかし、現状公開されている文章生成サービスはある分野に特化しているものが多く、詳細な設定や文章の傾向の調整を行うことができず、個人が自身の目的に合わせて文章の生成を行うことが困難となっている。本研究では、一般に公開されている言語モデルである GPT-2 を用いて、ネット上に無数に存在する見出しをもとにキーワードから見出しの例文を生成する。GPT-2 は Google Colaboratory 上で学習、実行が可能であり、文章の続きの逐次的に予測することのみで文章を生成するため、複雑なラベル付けを行っていないデータでの学習が可能である。これにより個人が目的に合わせてモデルを調整し、必要な文章の生成が可能になると考えられる。

第2章 基本知識

2.1 自然言語

自然言語とは日本語や英語、中国語などの、人間が日常的に意思疎通のツールとして使用する自然発生的な言語である。構文や単語の用法に厳格な規則が存在せず、社会的文脈に沿った曖昧な規則が存在する。対となる概念として人工言語、形式言語という、プログラミング言語、論理式などの構文や意味が厳格に定義された言語が存在する。これら二つの違いは人工言語がコンピュータを対象としているためコンピュータに解釈しやすいような構造になっており厳格に解釈が定められているのに対し、自然言語は類義語や多義語など解釈の幅が広く、使用する単語を入れ替えたり、単語の順番を入れ替えたりすることで自由な表現が可能となっている点である。

2.2 自然言語処理¹⁾

自然言語処理とは、自然言語の処理をコンピュータに行わせ、データと自然言語の変換や自然言語のより形式的な表現への変換などを行う技術のことである。機械翻訳や予測変換などにも用いられており、一般に広く浸透した技術となっている。前述の通り自然言語には曖昧な規則しか存在しないため、単語の区別や語義、構文の曖昧性の解析が非常に困難である。そのため、処理には言語学や確率論、統計学、機械学習などを使って形態素解析や構文解析、意味解析、文脈解析等を行うことで、意味を含めた解析が可能となっている。

2.3 学習手法

2.3.1 ファインチューニング²⁾

ファインチューニングとは「微調整」を意味し、Figure 2.1 に示すようにすでに学習済みのニューラルネットワークモデルの1部もしくは全体を別のデータセットを用いて再学習することである。既存のモデルを目的に合わせて微調整することで、1からモデルを学習するよりも少ないデータの領域へ、少ない学習時間でアプローチすることが可能である。

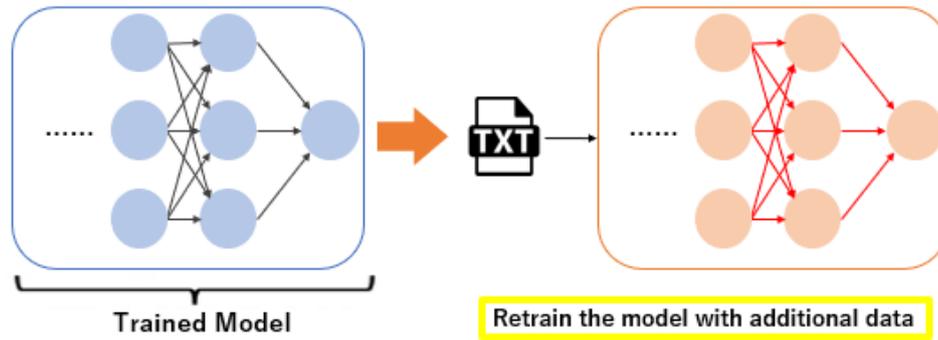


Figure 2.1 Fine-Tuning.

2.3.2 ショット学習³⁾

ショット学習とは、人間が数点の例を見て傾向を学習するように、ごく少ないデータをもとに未知のタスクを解くことを目標とする学習方針である。本研究ではGPT-2に文章の始まりを渡すプロンプトにていくつかの教師データを与え、特定の方法で指定されたキーワードから見出しを生成するというタスクを解決することを目標とする。また、このような学習手法はプロンプト学習とも呼ぶ。

第3章 関連技術

3.1 言語モデル

3.1.1 GPT-2²⁾³⁾⁴⁾⁵⁾

GPT-2 (Generative Pre-trained Transformer 2) とは、OpenAIが開発した大規模教師なし言語モデルである。800万 Web ページ分以上の膨大なデータから成る大規模な言語コーパスを用いた教師なし学習が行われており、特定の分野に関する固有のデータセットを使用していない。より大きな言語コーパスを使用し、より大きなモデルを事前学習させることによって、大量のデータを用いたファインチューニングを必要とせず、ゼロショットもしくはフューショットでも高い精度を出せる汎用的な言語モデルを目標として開発されている。

Figure 3.1 が示すようなその時点までの単語列から逐次的に次に続く単語を予測するという単純な学習手法を膨大なテキストデータに対して実行することで非常に高い精度を出すことが可能であり、文章生成の際も同様に予測をする処理を行っている。

GPT-2 は、OpenAI というアメリカの団体が開発しているため対応言語は英語となっているが、GPT2-japanese という GPT-2 のモデルを日本語で事前学習したものもあり、Rinna 社が公開している。4種類の規模のモデルを公開しており、最大のモデルのパラメーター数は13億以上となっている。

3.1.2 Bert²⁾⁵⁾⁶⁾

Bert (Bidirectional Encoder Representations from Transformers) とは、Transformer をベースに Google が開発した自然言語処理モデルである。Bert も大規模言語コーパスによる教師なしデータによって学習が行われたモデルであるが、ファインチューニングにはラベル付けを行ったデータを用いる必要があるという特徴を持つ。Bert の最大の長所は文章を双方向から学習することで文脈を理解することができる点であり、Figure 3.2 のような学習を行うことで、質疑応答や文の比較といった前後の意味が重要なタスクにおいて高い性能を発揮することが可能である。

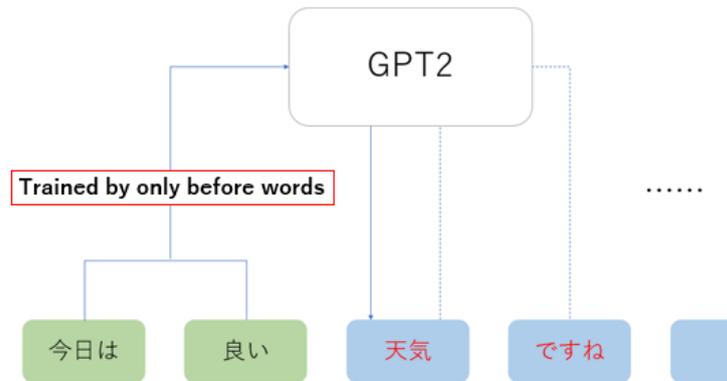


Figure 3.1 GPT-2 algorithm.

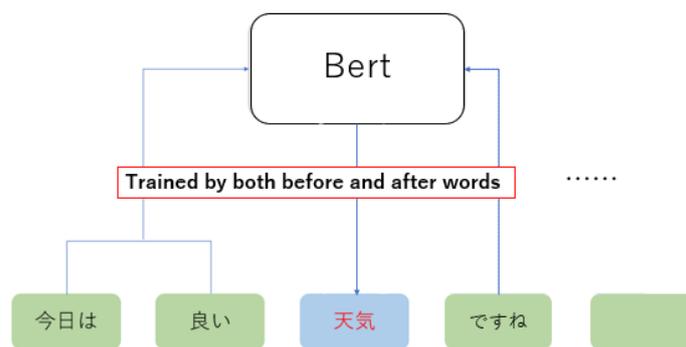


Figure 3.2 Bert algorithm.

3.1.3 Transformer²⁾⁶⁾

Transformer とは、2017 年に発表された”Attention Is All You Need”という自然言語処理に関する論文の中で初めて登場した深層学習モデルである。それまで主流だった CNN、RNN を用いたエンコーダ・デコーダモデルとは異なり、エンコーダとデコーダを Attention というモデルのみで結んだネットワークアーキテクチャである。Attention とは、文中の単語の意味を理解するのにどの単語に注目すればいいのかを表すスコアを決定するモデルのことであり、入力されたデータに重み付けをし、重要性を考慮したベクトル量として出力する。自然言語処理において Transformer と RNN 等を比較した際に最も大きな違いは処理の速度である。従来の自然言語処理では再帰構造を用いて言語構造を学習していたため処理の並列化が難しく、学習に非常に長い時間がかかった。しかし、Attention で入力文字列自体を対象とすることによって、入力文字列の言語的構造を学習することができる。学習された言語的構造を用いることで、さまざまな言語タスクが実行できるようになる。

第4章 実験⁷⁾

4.1 実験内容

Colab を用いて、性能の異なる 2 種類の日本語 GPT-2 モデルに対して学習の度合いの度合いを変化させながら見出しの生成を行った。Table 4.1 は使用した 2 種類のモデルにおける各評価項目のスコアである。表中のパラメーター数はモデル内の重みのパターン、層数はモデル全体のネットワーク層数、次元数は扱う単語ベクトルの次元、エポック数は学習の繰り返し回数、perplexity はモデルが文章生成時に絞り込める予測単語数を示している。これらのモデルを用い、各モデル、各学習方法ごとに【小惑星】と【観測】、【パソコン】と【価格高騰】という 2 組のキーワードに対する見出し生成をそれぞれ 10 回行い、その結果を記録し、比較した。前者のキーワードはファインチューニングに用いるデータに関連したもので、後者はあまり関連のない物となっている。また、japanese-gpt-1b は規模が大きく Colab 上でファインチューニングを実行できないため、ショット学習のみで実験を行った。

4.2 事前準備

4.2.1 Colab の設定

ドライブに十分な空きのある任意の Google アカウントにログインし、Colab を起動する。次に、画面上部のツールバーの「ランタイム」から「ランタイムのタイプを変更」を選択しハードウェアアクセラレータを GPU にした。

4.2.2 Google ドライブのマウント

モデルやデータの保存や呼び出しをするために、Colab 上で以下のコードを実行して Google ドライブをマウントした。また、今回の実験の関連ファイルをまとめる GPT というフォルダを作成した。

Table 4.1 Model summary.⁴⁾

モデル名	パラメーター量	層数	次元数	エポック数	perplexity
japanese-gpt-1b	1.3B	24	2048	10+	13.9
japanese-gpt2-medium	336M	24	1024	4	18

```
from google.colab import drive
drive.mount('/content/drive')
```

4.2.3 GPT-2のクローン

作成したフォルダへ移動し、Colab上で以下のコマンドを実行して日本語学習済みのGPT 2をクローンした。実行後にドライブでフォルダを確認し、クローンができていることを確認した。

```
cd /content/drive/MyDrive/GPT
!git clone https://github.com/tanreinama/gpt2-japanese
!git clone https://github.com/rinnakk/japanese-pretrained
    -models
```

4.2.4 学習データの用意⁸⁾

ファインチューニングを行う際に用いるテキストデータとしてWebサイト「Science Portal」の記事の見出しを100件収集し、各見出しから2つずつキーワードとなるフレーズを抜き出した。収集する記事には同サイトの「地球」ページ上のものを用い、著しく限定的な用語が見出し文の主題となっているものは用いないようにした。キーワードと見出しのデータは一度csvファイルでそれぞれを切り離した状態で保存し、Pythonを用いて任意の形の文章としてテキストファイル出力できるようにした。Table 4.2とTable 4.3は、それぞれ見出しを保存したデータと実際に学習に使用する際に出力したテキストの一例である。また、テキストファイルの文字コードは文字化けを防ぐためにUTF-8とした。このファイルを実験に使用するGoogleアカウントのドライブへアップロードした。

4.3 japanese-gpt2-mediumを用いた実験

4.3.1 必要ライブラリのインストール

必要なライブラリをインストールするため、クローンされたモデルと同じフォルダへ移動してColab上で以下のコマンドを実行した。また、requirement.txtの内容もFigure 4.1に示す。

Table 4.2 Data on csv.⁸⁾

キーワード1	キーワード2	見出し文
大地震	犠牲者 3.5 万人	プレート境界で長大な活断層が動いたトルコ大地震 もろい建物を直撃し、犠牲者 3.5 万人を超える
CO2 排出量	過去最高	世界の CO2 排出量は 過去最高の 331 億トン IEA 報告書
世界各地	極端な気象現象	世界各地で熱波や大雨、干ばつなどの「極端な気象現象」 温暖化が影響と国際機関や専門家
海水面積	最小記録	南極で海水面積の最小記録を更新、 極地研など観測 今後の予測は海中データに期待
マンモス	地球 2 周	マンモス、生涯で合計地球 2 周分も 歩いていたことが判明

Table 4.3 Data on txt.⁸⁾

出力した文章
見出しの生成。【大地震】【犠牲者 3.5 万人】：プレート境界で長大な活断層が動いたトルコ大地震 もろい建物を直撃し、犠牲者 3.5 万人を超える。
見出しの生成。【CO2 排出量】【過去最高】：世界の CO2 排出量は昨年も増えて過去最高の 331 億トン IEA 報告書。
見出しの生成。【世界各地】【極端な気象現象】：世界各地で熱波や大雨、干ばつなどの「極端な気象現象」 温暖化が影響と国際機関や専門家。
見出しの生成。【海水面積】【最少記録】：南極で海水面積の最小記録を更新、極地研など観測 今後の予測は海中データに期待。
見出しの生成。【マンモス】【地球 2 周】：マンモス、生涯で合計地球 2 周分も歩いていたことが判明。

```
tqdm
numpy
jaconv
tensorflow
```

Figure 4.1 requirement.txt (japanese-gpt2-medium).

```
cd /content/drive/MyDrive/GPT/japanese-pretrained-models/
!pip install -r requirements.txt
```

4.3.2 ゼロショットでの見出しの生成

japanese-gpt2-mediumを読み込み、ファインチューニング用データの形式でキーワードを与えて見出しを生成した。Table 4.4は【小惑星】と【観測】に対する出力で、Table 4.5は【パソコン】と【価格高騰】に対する出力である。結果を見ると、キーワードを用いた見出しといえるものはほとんど存在せず、関係のない文章や、意味の通らない文章が多く生成されていることが分かった。また、ファインチューニングをしていないため当然ではあるが、それぞれのキーワードごとに大きな差は見られず、どちらも類似した出力であった。

4.3.3 フューショットでの見出しの生成

同様の条件で、プロンプトへの入力に見出しの正解となる文章を3つ渡して同じキーワードに対して見出しの生成を行った。渡した文章をFigure 4.2に示す。Table 4.6が【小惑星】と【観測】、Table 4.7は【パソコン】と【価格高騰】に対する出力である。結果を見ると、キーワードに類似した分野の別の文章が生成されてしまうことも多いが、ゼロショットの時と比較して明らかに文章の質が上がり、見出しとしての文脈を持ったものが出力されていることがわかる。また、今回もそれぞれのキーワードごとに大きな差は見られず、どちらも類似した出力であった。

Table 4.4 Results of Keywords1 (0shot/japanese-gpt2-medium).

出力された文章
見出しの生成。【小惑星】【観測】：【小天体】【天球儀】【惑星】【海王星の輪】 【冥王星の輪】【小天体】【流星】【気象】。
み見出しの生成。【小惑星】【観測】：小惑星の軌道に近づくと、小惑星の重心 が地球に近づくと考えられ、地球に衝突することが考えられます。
見出しの生成。【小惑星】【観測】：全自動タイプ:.h1 の生成例。【画像】：縮小 画像ファイル(要 18mb).h1 はスーパースペクトルである。
見出しの生成。【小惑星】【観測】：【受信】【更新】 見出しは、常に情報の正確 さを保証するものではありません。
見出しの生成。【小惑星】【観測】：小澤啓一「(小惑星探査機はやぶさ)～心が 躍るようなとっておきの旅物語～」【地球】【旅】→【旅行】
見出しの生成。【小惑星】【観測】：彗星は8種類存在する。この表記のほかにも、 一般名詞の語順を決めるものもある。)
見出しの生成。【小惑星】【観測】：地球の軌道周回軌道をとる惑星のリスト情報 などを提供する。【一般】の概要。【映画】5万年 猿の惑星〈新世紀〉(2012) フルメイキング版
見出しの生成。【小惑星】【観測】：ライトカーブ、コロナグラフの画像、画像、 光源の分解などを行う。【インターネット】【テレビ】【音楽】【映画の字幕化】
見出しの生成。【小惑星】【観測】：4月4日05時26分(日本時間)、小惑星探査 機「はやぶさ2」が東京都内で、地球に最も接近した状態にある
見出しの生成。【小惑星】【観測】：小惑星の一覧、小惑星の一覧(30001-14000) 小惑星の一覧(36001-37000)

Table 4.5 Results of Keywords2 (0shot/japanese-gpt2-medium).

出力された文章
見出しの生成。【パソコン】【価格高騰】:価格高騰の理由は、価値が上がらないから。
見出しの生成。【パソコン】【価格高騰】:価格は当社の標準です。複数の販売サイトの共通仕様に基づいてお選びいただくか、標準と異なる価格でご購入いただけます。
見出しの生成。【パソコン】【価格高騰】:wordpress の設定ファイルを編集して「更新」して、ページ数や記事数を取得して反映する方法-
見出しの生成。【パソコン】【価格高騰】:合計:1,300 円 (税込・送料別) の商品 【パソコン】【安】:合計:990 円 (税込・送料別) の商品
見出しの生成。【パソコン】【価格高騰】:無印 『ニコニコ動画』 の動画が見やすくなりました。 6,980 円。 ※購入者が獲得したポイントの一部は次回以降の購入額にポイントが反映されます。
見出しの生成。【パソコン】【価格高騰】:を最初に選択してから「見出しの生成」をクリックする。 【パソコン】【価格高騰】:にチェックすると、見出しを作成する画面が表示されます。
見出しの生成。【パソコン】【価格高騰】:1 万円 (税別)～【機種のご紹介】:1 万円 (税別)～【機種のご紹介】:1 万円 (税別)～
見出しの生成。【パソコン】【価格高騰】:f-032 価格競争時代の脱却.vps 転送速度測定結果.vps 転送速度と利用状況の調査結果
見出しの生成。【パソコン】【価格高騰】:【タブレット】【書籍】【デジタル】 【電子書籍
見出しの生成。【パソコン】【価格高騰】:価格の新しい安いの商品の検索と、新しい安い最新のを検索します。

見出しの生成。【積乱雲】【連続発生】：積乱雲が連続発生する「バックビルディング現象」が起きていた 西日本豪雨で防災科研が解析。見出しの生成。【温室ガス】【減少】：17年度の日本の温室ガスは4年連続減少 パリ協定の目標には一層の削減努力必要。見出しの生成。【位置情報】【本格運用】：日本版GPS体制の本格運用始まる 高精度位置情報の活用を期待。

Figure 4.2 Few-shot data.⁸⁾

Table 4.6 Results of Keywords1 (few-shot/japanese-gpt2-medium).

出力された文章
見出しの生成。【小惑星】【観測】：「7月2日に375億kmを1周する小惑星探査機はやぶさ2の打ち上げも控え」と11日朝、宇宙航空研究開発機構(jaxa)の関係者が情報公開。
見出しの生成。【小惑星】【観測】：2018年10月3日に木星探査機「ガリレオ」が木星に衝突 衛星での観測で「木星の太陽面通過」が確認された
見出しの生成。【小惑星】【観測】：小型天体から小型衛星までを識別可能。最新の観測観測結果から見える新たな発見を紹介。
見出しの生成。【小惑星】【観測】：宇宙科学に学ぼう：国際宇宙ステーションから約100億円が盗まれる事件。日本の軍事秘密が暴かれるか!?
見出しの生成。【小惑星】【観測】：共同観測サイト「gcomm」は今年9月に運用開始。観測機器の準備やデータの収集を行っている。
見出しの生成。【小惑星】【観測】：小惑星探査機「はやぶさ2」が運用開始 衛星からの微量の物質サンプルを地球大気に戻す実験に成功、注目 液体窒素を常時7倍にする水処理装置を開発。
見出しの生成。【小惑星】【観測】：スマホの写真から天気予報が確認できる「ファインモーション」など追加。
見出しの生成。【小惑星】【観測】：2018年8月6日、彗星が小惑星の一覧表に追加 1月1日(金)深夜1時55分：「日本の国会で8日の辞意表明演説中に北朝鮮が核実験」
見出しの生成。【小惑星】【観測】：探査機「はやぶさ」から5枚の画像が届いた 福島県、鹿島神社に桜田門から入る。
見出しの生成。【小惑星】【観測】：はやぶさ2の帰還を予測 大規模な噴火 警戒レベルを1にして緊急地震速報が作動する確立も

Table 4.7 Results of Keywords2 (few-shot/japanese-gpt2-medium).

出力された文章
見出しの生成。【パソコン】【価格高騰】：価格が41万6000円に迫る勢い。東京新聞は2013年の1人当たり1万円台のパソコンも販売している。
見出しの生成。【パソコン】【価格高騰】：3年の総額が90億円に迫りました。台風16号の被害地域の被害を軽減へ。
見出しの生成。【パソコン】【価格高騰】：windows xp のサポートがなくなって、14年に購入したpcはサポート対象外になる。
見出しの生成。【パソコン】【価格高騰】：デスクトップパソコン大手のドスパラが「タブレット」の注文数を30万本以上に引き上げる。→2015年11月に発売のノートパソコン販売数が前年比9.4%減
見出しの生成。【パソコン】【価格高騰】：本日のコピックは円安の影響が指摘されている 宇宙航空研究開発機構(jaxa)と isas(isas) が共同開発。
見出しの生成。【パソコン】【価格高騰】：価格を安定させる難度の高い「office 365」向け管理サービスの月額利用料金が「10万円」になる
見出しの生成。【パソコン】【価格高騰】：「値段が高騰している」 コニカミノルタが独自に開発した高精度位置情報の活用に期待。
【パソコン】【価格高騰】：17年度の日本の パソコン利用は16年度比、17年度の価格は8割減 東日本大震災の発生時刻の時刻ずれを修正。
見出しの生成。【パソコン】【価格高騰】：19年度半ばまでに普及目標値の1,000万円超超を突破 フランスの国民投票結果を速報するニュースの見出しの生成。
見出しの生成。【パソコン】【価格高騰】：パソコンの低価格化でパソコンの値段が高騰。販売価格が月々2万円を切る販売商品も出始めた

```
accelerate >= 0.12.0
torch >= 1.3
datasets >= 1.8.0
sentencepiece != 0.1.92
protobuf
evaluate
scikit-learn
```

Figure 4.3 requirement.txt (Fine Tuning).

4.4 ファインチューニングを行ったjapanese-gpt2-mediumでの実験

4.4.1 ファインチューニング

一度GPTフォルダへ戻ってから、以下のコマンドを実行し、ファインチューニングの実行に必要な追加のライブラリをインストールした。Figure 4.3は、requirement.txtの内容である。その後ドライブにアップロードしたテキストファイルを用いてファインチューニングを行った。この時、学習のエポック数は3回とした。

```
cd /content/drive/MyDrive/GPT
!git clone https://github.com/huggingface/transformers
pip install -r ./transformers/examples/pytorch/language -
    modeling/requirements.txt
```

4.4.2 ゼロショットでの見出しの生成

ファインチューニングを行った japanese-gpt2-medium を読み込み、ファインチューニング前と同様にキーワードを与え、見出しの生成を行った。Table 4.8 と Table 4.9 がそれぞれ【小惑星】と【観測】、【パソコン】と【価格高騰】に対する出力である。結果を見ると、ファインチューニングを行う前と比較して、キーワード指定の後ろに意味のある文章が生成されるようになっていることがわかった。また、【小惑星】【観測】に対する生成結果では【パソコン】【価格高騰】に対する出力と比較して、見出し然とした文章が多く生成されている。一方で文中に「小惑星」や「はやぶさ2」といった単語が過剰に使用される傾向があることもわかった。

4.4.3 フューショットでの見出しの生成

ここまでと同様に、ファインチューニング前と同じ見出しの教師データを与え、同じキーワードから見出しの生成を行った。Table 4.10 は【小惑星】と【観測】、Table 4.11 は【パソコン】と【価格高騰】に対する出力である。結果を見ると、主述関係やキーワードの意味が逆転してしまうことはあるものの、ほぼすべての文章が見出しとして成立するものとなっていることがわかった。【小惑星】【観測】側では相変わらず同じ単語が多く表れる特徴がみられた一方、【パソコン】【価格高騰】側ではそのような特徴は現れなかった。

Table 4.8 Results of Keywords1 (0shot/ft).

出力された文章
見出しの生成。【小惑星】【観測】：「browser culture retailers」が見つめる「見える宇宙」 地球の生命の謎解明へ「...全文へ」
見出しの生成。【小惑星】【観測】：マレビト 1pbm (nasa 提供) 【発見】：しらせ 2a (nasa 提供) 【発見】：しらせ 2a (nasa 提供)
見出しの生成。【小惑星】【観測】：小惑星探査機はやぶさ 2、小惑星「イトカワ」を観測する 観測期間 6 年超。
見出しの生成。【小惑星】【観測】：小惑星探査機「はやぶさ 2」に捕獲されたはやぶさ 2(jaxa) 「はやぶさ 2」の捕獲を成功させた jaxa
見出しの生成。【小惑星】【観測】：地球史の中のとてつもない役割 -朝日新聞 デジタル 平成 28 年度 (2016 年度) の、1 次観測「海図解読」の成果
見出しの生成。【小惑星】【観測】：m89 まで 210 億年前の小惑星「イトカワ」が初めて観測。宇宙の全プロセスを解明する。
見出しの生成。【小惑星】【観測】：彗星の軌道に近づいた小惑星は、「小惑星」を生成し、近づいた星の観測を始める。
見出しの生成。【小惑星】【観測】：2018 年 6 月 3 日に天文学者の jake a. thorai により発見された太陽系近傍の小惑星。
見出しの生成。【小惑星】【観測】：小惑星、21 年ぶり小惑星、超新星の発生に期待が膨らむ：科学ドットコム
見出しの生成。【小惑星】【観測】：「日本科学未来館」から提供する小惑星探査機「はやぶさ 2」が撮影した、小惑星「イトカワ」の画像

Table 4.9 Results of Keywords2 (0shot/ft).

出力された文章
見出しの生成。【パソコン】【価格高騰】:google、10月には1ドル90円以上も。株価も9日に933ドルまで高騰、さらに自動車価格との連動が効いている？
見出しの生成。【パソコン】【価格高騰】:2000年6月1日 さて、2年前の2000年6月1日にamazonで発表されたタイトル
見出しの生成。【パソコン】【価格高騰】:2000年6月1日 さて、2年前の2000年6月1日にamazonで発表されたタイトルですが、現在もどんどん値上がりしつつあります。
見出しの生成。【パソコン】【価格高騰】:更新日(12月28日) 【株価】:2461円 【業績】:2兆円を超える黒字 (the page)
見出しの生成。【パソコン】【価格高騰】:インターネットを使った、毎日のお買い得情報をお届け! itmedia internet で家電系ニュースを毎日配信中。
見出しの生成。【パソコン】【価格高騰】: 【価格高騰】に関するサイトを抽出して順位付け。まとめサイトのリンク先にある【価格高騰】に点数をつけ加え、高得点順に並べ替えてランキング表示。
見出しの生成。【パソコン】【価格高騰】:パソコンなどの価格が気になる…。もうちょっとだけ気になる情報をまとめてみよう。
見出しの生成。【パソコン】【価格高騰】:9割以上が20年後も買い続ける…。テーマ「カメラとノート」:複数の焦点距離を持つ3つのレンズが高機能デジカメに!
見出しの生成。【パソコン】【価格高騰】:2015年9月1日にマイクロソフトによる価格改定が行われました。市場全体の価格上昇を受け、従来価格3,240円(税込)が2,500円(税込)に価格改定。
見出しの生成。【パソコン】【価格高騰】:日本ibmが「パソコンの商品は、従来機種より高いです」。販売価格を調べてみました。

Table 4.10 Results of Keywords1 (few-shot/ft).

出力された文章
見出しの生成。【小惑星】【観測】：小惑星「リュウグウ」に1月10日に初検出 探査期間に間に合わずか。
見出しの生成。【小惑星】【観測】：巨大小惑星「イトカワ」が10月28日16時 ごろ東北東の日本海で観測。1ヵ月ぶりに小惑星を観測できる。
見出しの生成。【小惑星】【観測】：「小惑星探査機はやぶさ2」に期待、来る21 日小惑星「はやぶさ2」を追跡。
見出しの生成。【小惑星】【観測】：小惑星探査機「はやぶさ2」の動き「はや ぶさ2」、前回と違う動き方を解明。
見出しの生成。【小惑星】【観測】：日食観測で小惑星を発見する動き。
見出しの生成。【小惑星】【観測】：東京大学天文学会が小惑星24「リュウグウ」 の観測を再開 小惑星「リュウグウ」への重要探査も始動。
見出しの生成。【小惑星】【観測】：東海で観測 小惑星「(2018)イトカワ」をと らえる新たな指標を発見 国際天文学連合で発見。
見出しの生成。【小惑星】【観測】：小惑星は地球の自転と軌道の間係を探る jaxa、 イトカワと、月探査など 日本人初の小惑星で月探査など 小惑星の探査に期待。
見出しの生成。【小惑星】【観測】：今夜小惑星探査機「はやぶさ2」がイトカ ワに衝突 高イトカワ孔を発見。
見出しの生成。【小惑星】【観測】：小惑星・【観測】 明日に地球に大接近 南極 の衛星「リュウグウ」観測結果の連続観測開始を明日に。

Table 4.11 Results of Keywords2 (few-shot/ft).

出力された文章
見出しの生成。【パソコン】【価格高騰】：「パソコン」価格が高騰 1万円を超える。
見出しの生成。【パソコン】【価格高騰】：「ピーク時のパソコン価格が43%増。」pcの安値がピークを迎えたため。
見出しの生成。【パソコン】【価格高騰】：価格が上昇している アマゾンやソニーは新たなパソコンを順次投入 家電量販店で購入可能に。
見出しの生成。【パソコン】【価格高騰】：欧州のpc販売価格高騰が影響 大手電機大手のマルク社、全商品の新価格値上げを要請 東京五輪用のパソコン4倍の価格設定が。
見出しの生成。【パソコン】【価格高騰】：necパソコンが2万3千円から2万5千円に メモリ64g搭載「g-bookpc」の最上位モデルなど。
見出しの生成。【パソコン】【価格高騰】：パソコン販売価格が上昇 インテルが早期価格据え置き見通し。見出しの生成。
見出しの生成。【パソコン】【価格高騰】：4割引の中古パソコン 東芝、最新のモデルが20%オフ。
見出しの生成。【パソコン】【価格高騰】：通信容量が少なくて採算が合わない？普及拡大の陰で価格が高騰 国産パソコン人気で価格が高止まり。
見出しの生成。【パソコン】【価格高騰】：シェアウェアで人気1万円台のパソコンは、今年1月現在の価格が1万円を割り込み【パソコンの値引き幅が大きかった】。
見出しの生成。【パソコン】【価格高騰】：通販サイトの「お得でお得」パソコン購入。価格が急速にアップ、amazonのクラシックモデルが一番高く

```
torch >=1.7.1
transformers >=4.5.0
tensorflow
tensorboard
sentencepiece
tqdm
fugashi
unidic
```

Figure 4.4 requirement.txt (japanese-gpt-1b).

4.5 japanese-gpt-1b での実験

4.5.1 モデルの読み込み

中規模モデルの時と同様に、クローンされたモデルと同じフォルダへ移動して Colab 上でコマンドを実行し、必要なライブラリのインストールした。Figure 4.4 は requirement.txt の内容である。

4.5.2 ゼロショットでの見出しの生成

japanese-gpt-1b を読み込み、同じの形式で同じキーワードから見出しの生成を行った。Table 4.12、Table 4.13 は【小惑星】【観測】と【パソコン】【価格高騰】それぞれのキーワードに対しての出力である。結果を見ると、japanese-gpt2-medium と比較しても全体を通して文章が崩れておらず、見出しといえる文脈を持つ物もいくつか生成されていることがわかった。

4.5.3 フューショットでの見出しの生成

ここまでの実験と同様に、見出し生成の教師データをプロンプトに渡して【小惑星】【観測】と【パソコン】【価格高騰】を用いた文章生成を行った。Table 4.14 は【小惑星】と【観測】、Table 4.15 は【パソコン】と【価格高騰】に対する出力である。結果を見ると、ほぼすべての生成文が見出しの文脈を取っており、キーワードの要素を持ったものが生成されていることがわかった。

Table 4.12 Results of Keywords1 (0shot/japanese-gpt-1b).

出力された文章
見出しの生成。【小惑星】【観測】：小惑星探査機「はやぶさ2」が捉えた地球に急接近する小惑星「リュウグウ」の巨大なクレーター
見出しの生成。【小惑星】【観測】：大小のクレーターには、太陽からの電子と星との衝突により電子が飛散し、その部分にくぼみなどが生まれ、小惑星のでき方がわかるようになった。
見出しの生成。【小惑星】【観測】：NASA【画像】：米軍が撮影 - GIGAZINE【追記】：小惑星は地球に衝突する。
見出しの生成。【小惑星】【観測】：Karenium は、アンドロメダ A の大マゼラン雲に 24 日ごろに衝突しました。
見出しの生成。【小惑星】【観測】：このプロットを含むイメージは 1 枚しかありません。作成者は予告なしにこのイメージを変更する場合があります。
見出しの生成。【小惑星】【観測】：このプロットを含むイメージは 1 枚しかありません。作成者は予告なしにこのイメージを変更する場合があります。
見出しの生成。【小惑星】【観測】：リュウグウ表面の写真や月面ショット。小惑星探査機「はやぶさ2」の詳細な軌道。
見出しの生成。【小惑星】【観測】：小惑星探査機「はやぶさ」が撮影... 地球から 1 億キロ離れた小惑星イトカワに 2010 年 7 月 9 日到着
見出しの生成。【小惑星】【観測】：小惑星(ひてんしょう、英語: planet ball, 固有名詞: polaris) とは、微小な天体。
見出しの生成。【小惑星】【観測】：地球から 700 万キロ以上も遠くに存在するため、小惑星観測で小惑星まで近づいたとしても、肉眼では小惑星内部の構造までは見通せない。

Table 4.13 Results of Keywords2 (0shot/japanese-gpt-1b).

出力された文章
見出しの生成。【パソコン】【価格高騰】:WordPress を使ってみよう! WordPress でスマホサイト構築 初心者向け
見出しの生成。【パソコン】【価格高騰】:シャープ製 PC が不正コピーで売れる時代がやってきた。
見出しの生成。【パソコン】【価格高騰】:レイトン教授の買取価格一覧とおおすすめの買取業者を紹介!
見出しの生成。【パソコン】【価格高騰】:パソコンがいくら高くなっても、ある程度の相場観をもった人が購入するので値上がりするという。
見出しの生成。【パソコン】【価格高騰】:価格高騰の度合をランキング表示します。価格表の最上部右端がランキングのスタート地点。
見出しの生成。【パソコン】【価格高騰】:【パソコン】が激安特価!。以上、もしあなたが楽を約束してパソコンをお薦めするならこう言おう。
見出しの生成。【パソコン】【価格高騰】:Winで『XZ Premium』をダウンロードするデスクトップアプリの最新版リリース!
見出しの生成。【パソコン】【価格高騰】:Android、iPhone、タブレット向けのAppleが作っているブラウザゲームが突然高騰してます。
見出しの生成。【パソコン】【価格高騰】:Windows 側で、OS 標準の拡張機能拡張「拡張子検索」を用いて、見出しを生成しよう!
見出しの生成。【パソコン】【価格高騰】:こんにちは、少し前に出ていた記事で、今話題になっているのがこの「見出し生成」ソフト「N-you-in」です。

Table 4.14 Results of Keywords1 (few-shot/japanese-gpt-1b).

出力された文章
見出しの生成。【小惑星】【観測】：初めての「あかつき」到着 宇宙科学研究所で初めての天体観測 14日正午に撮影。
見出しの生成。【小惑星】【観測】：深宇宙で大きな小惑星が観測される。小惑星の観測体制充実: 準天頂衛星初号機搭載のすばる望遠鏡で小惑星を探る。
見出しの生成。【小惑星】【観測】：小惑星リュウグウの土壤中の放射性元素の分布を確認 宇宙科学研究所が観測 これまで発見されていなかった小惑星の核の組成を明らかにする。
見出しの生成。【小惑星】【観測】：約2年ぶりの小惑星の発見。
見出しの生成。【小惑星】【観測】：小惑星発見!本格運用は27年ぶり。観測成功、データの公開も。
見出しの生成。【小惑星】【観測】：平成29年に小惑星リュウグウに到着した地球観測衛星 (ALOS)、太陽光の少ない場所の観測に成功、表面温度の観測や水に関する情報などの情報。
見出しの生成。【小惑星】【観測】：小惑星リュウグウを35年掛けて観測。
見出しの生成。【小惑星】【観測】：小惑星探査機「はやぶさ2」初の岩石採取に成功。
見出しの生成。【小惑星】【観測】：小惑星探査機「はやぶさ2」が地球と火星の間の小惑星イトカワから持ち帰った物質の観測を開始。
見出しの生成。【小惑星】【観測】：小惑星が捉えた物体は計33個 打ち上げからの最初の1週間で初観測 数年後に宇宙の謎が少しずつ解明される。

Table 4.15 Results of Keywords2 (few-shot/japanese-gpt-1b).

出力された文章
見出しの生成。【パソコン】【価格高騰】：パソコンの価格も上昇し、複数台購入を躊躇う 値上がりの原因に迫る。
見出しの生成。【パソコン】【価格高騰】：パソコンの高値つり上げ問題、政府は詳細調査を指示。見出しの生成。
見出しの生成。【パソコン】【価格高騰】：パソコンやデジタル家電の価格が2～5%高い 昨今の世界的な需要拡大を受け価格高騰。
見出しの生成。【パソコン】【価格高騰】：パソコン価格は高騰中。見出しの生成。【海洋汚染】：有害廃棄物の海洋汚染問題が深刻化。
見出しの生成。【パソコン】【価格高騰】：パソコンが高い 高い 3～4万円も... 国内パソコンメーカー、値上げ傾向続く。
見出しの生成。【パソコン】【価格高騰】：値上がり続く ハードウェアで100万円以上高騰。見出しの生成。
見出しの生成。【パソコン】【価格高騰】：PCの購入に新たな打撃。【IoTでつながる世界】：IoT(モノのインターネット)が世の中を豊かにする。
見出しの生成。【パソコン】【価格高騰】：パソコンの価格高騰続く。価格高騰で家計は迫。
見出しの生成。【パソコン】【価格高騰】：スマホ発売急落でPCの値下がり歯止め。見出しの生成。
見出しの生成。【パソコン】【価格高騰】：中国製パソコンの価格高騰について 貿易摩擦を懸念する声強い 経済への悪影響を懸念する声強い。

4.6 結果考察

今回の実験では、使用するモデルの規模と学習の度合いを変化させて、同じキーワードの入力に対しての出力結果の変化を調べた。まず、japanese-gpt2-medium のファインチューニングの有無による出力結果を比較すると、ファインチューニング後のゼロショットでは「見出しの生成。【】【】:」というフォーマットの入力に対する出力で正常な文章を生成できるようになっていることがわかる。これはファインチューニングによって入力の意図を学習したことを示している。その一方で生成した文章には見出しの文脈以外の物も多く含まれており、学習があまり進んでいないことがわかる。これは、キーワード指定の入力形式が使用されにくい文体であることでもともとのデータが少ないからであると推測する。GPT は文章の次に続く単語を逐次的に予測するという手法で文章を生成する。そのため、もともとの使用例が少ない文型のキーワード指定の部分に続く要素はファインチューニングによる学習の影響が現れやすく、その後の通常の文章形式で表現されている続く見出し部分には学習の影響が現れにくいと推測できるからである。次に、japanese-gpt-1b の出力結果に注目すると、japanese-gpt-1b のゼロショット時のキーワード指定に対する出力がファインチューニングを行っていないにもかかわらず正常な文章になっていることがわかる。これは、モデルの規模が大きくより多くの文章を学習する上で、キーワード指定のフォーマットに類似する形式も学習しているからであると考えられる。

ところで、フューショットの出力に注目してみると、ファインチューニングの有無、モデルの規模にかかわらずキーワードと見出しという関係がある程度成立していることがわかる。これは、ファインチューニングがあくまでデータの一つとして学習を行うだけであるのに対し、フューショットで与える教師データは出力と同じ文章中に入っていることが要因であると考えられる。ファインチューニングでは無数にある文章の構成の一つとして学習が行われるためサンプル数の多い文型にはあまり大きな影響は与えられない。しかし、フューショットでは、出力する文章は与えられたデータに続く形になるため、教師データの文章校正の影響が強く出るからである。

また、ファインチューニングを行った際に【小惑星】【観測】をキーワードとして出力した文章に注目すると、明らかに高い頻度で「小惑星」や「はやぶさ」「リュウグウ」などの単語が用いられていることがわかる。【パソコン】【価格高騰】をキーワードとした際にはこの現象が発生していないことから、ファインチューニングに用いたデータに類似

するキーワードを使用したからであると考えられる。これを回避するためには、ファインチューニングのエポック数を小さくすることで与える影響を小さくするか、学習データを幅広く大量に集めることで相対的に影響を小さくすることが有効であると推察できる。

これらの結果、考察から、典型的なデータを用いたフューショットと、幅広い多量のデータを用いたファインチューニングを併用する手法を用いることで、中規模な GPT-2 モデルを用いて見出しの生成が可能であると推測できる。また、ニュース記事の見出し生成以外を目的とした場合でも、同様の手法を用いることで目的に合わせた学習が可能であり、十分な成果を上げることができると考えられる。

第5章 結論

5.1 結論

本研究では GPT-2 の日本語学習済みモデルと Google Colaboratory を用いて、ファインチューニングやショット学習、あるいはその両方を用いてキーワードからの見出し生成を行った。その結果、ファインチューニングとフューショット学習を組み合わせることで、GPT-2 モデルの単純な逐次予測でも自然な見出しの生成が可能であることがわかった。

また、使用するモデルが大規模であれば少ない学習で自然な文章を生成することが可能であった。しかし、中規模なモデルを使用した場合でも、与える教師データの形式を工夫し、ファインチューニングやショット学習を組み合わせることでその差を小さくし、十分な成果を挙げられることがわかった。

謝辞

最後に、本研究を進めるにあたり、ご多忙中にもかかわらず多大なご指導をしていただきました出口利憲先生、また、ともに勉学に励んだ同研究室のメンバーに厚く御礼を申し上げます。

参考文献

- 1) 総務省 情報通信機器の保有状況

<https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r04/html/nd238110.html#:~:text=%E3%82%A2%20%E6%83%85%E5%A0%B1%E9%80%9A%E4%BF%A1%E6%A9%9F%E5%99%A8%E3%81%AE,8%2D1%2D1%EF%BC%89%E3%80%82>

- 2) 新納 浩幸, 古宮 嘉那子, ”書類分類から始める自然言語処理入門”

- 3) 日本語で学習させた GPT2 で GPT3 っぽいことを試してみた

<https://www.geekfeed.co.jp/geekblog/gpt2/>

- 4) 自然言語処理初心者が「GPT2-japanese」で遊んでみた

<https://developer.mamezou-tech.com/blogs/2022/07/08/gpt-2-japanese/>

- 5) 杉山 聡, ”本質をとらえたデータ分析のための分析モデル”

- 6) Huggingface Transformers 入門 (1) - 事始め

<https://note.com/npaka/n/n5bb043191cc9>

- 7) 我妻 幸長, ”あたらしい 人工知能技術の教科書”

- 8) SciencePotal 地球 記事一覧

<https://scienceportal.jst.go.jp/earth/page/1/index.html>