仮タイトル

1　はじめに

自然言語分野では, 文章生成など創作に関わるものも人間と遜色なく生成できるようになり, 大きく注目されている.

人の考えた文章と計算機が生成した文章との違いに注目した.

本研究ではオープンソース化されていない GPT-3 に代わって, 前身の GPT-2 を使って, 文章の破綻推定をした. 人工知能に人の文章と自動生成の文章が識別できるのかと, GPT-2 の文章生成能力の確認実験をした.

2　要素技術

2.1　Transformer

Transformer[1] は再帰的ニューラルネットワーク (Recurrent Neural Network: RNN)[2] を使わずに, Self-Attention を使って並列計算を可能にするモデルである.

RNN は時系列に沿って順番に計算する構造であるため並列計算ができず, GPU などを使っても計算時間が長いという欠点がある. GPU の性能を十分に活用し, 計算速度を向上させるため, RNN を使用しないことが必要となる.

Self-Attention は, シーケンス内の単語間の関係性に注目する. Transformer は Self-Attention の計算を並列化するとこで効率的な計算ができる.

2.2　GPT-2

Generative Pretrained Transformer 2 (GPT-2)[3] は

特定のタスクに特化した数万の教師ありデータでファインチューニングが必要だった GPT, BERT をはじめとしたモデルと異なり, Transformer のデコーダ部分を利用し, 大規模な言語コーパスで様々なタスクを学習する汎用的なモデルとして設計された.

任意の長さの文章をシンボルにしたシーケンスを $(s_1, s_2, ..., s_n)$ とする. 同時確率は条件付確率の積に

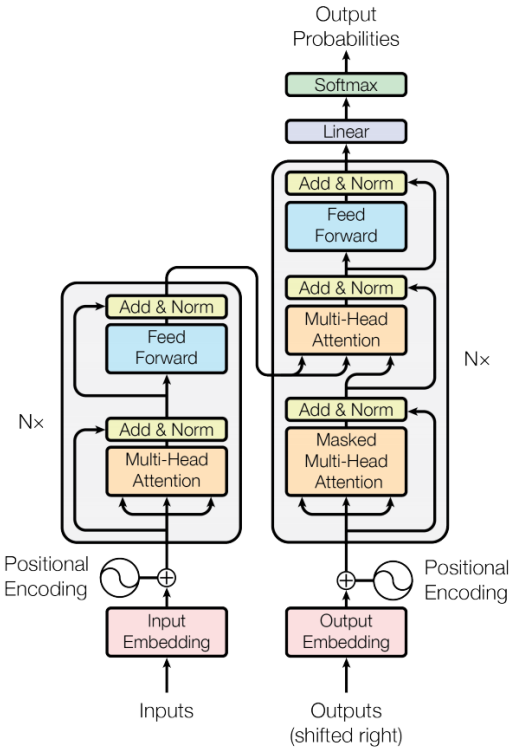

図 1: Transformer

分解される.

$$p(x) = \prod_{i=1}^{n} p(s_n|s_1, ..., s_{n-1}) \tag{1}$$

この条件付確率は, Transformer[1] によって効率的に計算できる.

GPT-2 では, $p(\text{output}|\text{input})$ を一般化して, $p(\text{output}|\text{input}, \text{task})$ と表しタスクの種類も学習することで, 特定の問題のためのファインチューニングなしで, 1 つのモデルでも問題を解けるようにしている.

2.3　Byte Pair Encoding

Byte Pair Encoding (BPE) は高頻度の単語は単語全体を辞書に登録し、低頻度の単語は文字単位に分割する言語モデルにおける未知語処理手法である.

言語モデルは語彙サイズをハイパーパラメータとして設定するため, ニューラルモデルに扱われない未知語

表 1: GPT-2 の解けるタスク

AutomaticSpeechRecognition
Conversational
FeatureExtraction
FillMask
ImageClassification
QuestionAnswering
Summarization
TextClassification
TextGeneration
TokenClassification
Translation
ZeroShotClassification

が存在する. 未知語処理として, $< unk >$ などの特殊トークンに置き換える方法と, 単語をより細かく分割しサブワードにすることで語彙を少なくする方法がある.

BPE はデータ圧縮手法を言語モデルに応用した. 文字単位に単語を分解し, 2 文字のペアの中で高頻度の要素を結合して 1 つのサブワードとする手順を繰り返すことで, 単語を接頭辞や接尾辞などの意味のある単位に分解できる.

2.4 Manga 109

Manga 109 [4] は, 漫画の学術研究への使用を目的として作成された研究用コミックデータセットである. 1970 年代から 2010 年代に公開された, 日本のプロ漫画家によって描かれた漫画 109 冊で構成されている.

2.5 BERT

BERT [5] は, 2018 年に Google が発表した汎用言語モデルである. 複数の双方向 Transformer を用いることで文脈を考慮することができるモデルとされている. 各タスクに対してファインチューニングをすることでさまざまなタスクに柔軟に対応することができる.

事前学習には入力の一部の単語を“[MASK]”に置き換えてその元単語を予測するように訓練するタスク (Masked Language Model) と 2 文を入力としてその連続性を識別するように訓練するタスク (Next Sentence Prediction) が用意されている. 本研究では, 入力する文章の分類器として選んだ.

表 2: GPT-2 の設定

parameter	value
max length	50
top p	0.95
top k	60

表 3: BERT の設定

parameter	value
Optimizer	AdamW
lr	2e-5
loss	Cross Entropy
batch size	32
train data	12502
valid data	1389
test data	1543
epoch	20

3 提案手法

人間が作成した文章と, GPT-2 が作成した文章のデータセットを作成し, BERT で作成者を推定する問題を解く.

3.1 データセット

人間が作成した文章として Manga 109 中のセリフを利用した.

として, 連続する 2 文をつかう.

文章間の意味が繋がっている必要があるが, 意味を考慮した単位に文章を分割するのは難しい. そのため決められたフォーマットで, 繋がりを持った文章として仮定しやすい 4 コマ漫画を対象として Manga109 中の 5 作品のデータを用いた.

Manga109 のセリフデータを S_1, S_2, S_n として, 連続する 2 文を実際のデータ, GPT-2 で次の文章を推測した S’ を含んだ 2 文を生成データとして, データセットを作成した. 文と文の間は [SEP] トークンで結合しました.

GPT-2 による生成では, 生成文の句点までを 1 文とし, [SEP], [EOS] などのトークンは削除したものを S’ とした.

表 4: GPT-2 の生成した文章の例

title		sentence
Akuhamu	S_i	我 梢つつじの名において命じる！
	S_{i+1}	いでよ！ 悪魔っ
	S'_{i+1}	あなたは、我 の の は を どうする ら れ ば いいのですか
YouchienBoueigumi	S_i	最近世間がぶっそうだわ
	S_{i+1}	いつまでも人任せに安穏としていていいのかしら…
	S'_{i+1}	・・・・・なんでこういうことになったのか知らないが、今から思えば大問題 ・・・ こういうことをやっている人には何を言っても無駄なのか

ⓒあくはむ 新居 さとし, ⓒ幼稚園ぼうえい組 テンヤ

3.2 分類問題

実データのラベルを 1, 生成データのラベルを 0 として BERT を利用して 2 値分類問題を解く.

高いと BERT で分類可能性. 低いと GPT-2 の文章生成能力が高い.

表 5: 分類実験の結果

	loss	accuracy
train	0.010	0.997
test	0.123	0.959
baseline	—	0.50

表 6: 混合行列

		予測	
		1	0
ラベル	1	753	39
	0	25	726

4 実験

表 2, 表 3 は GPT-2 と BERT の実験パラメータである. GPT-2 の事前学習パラメータには rinna/japanese-gpt2-medium を, BERT の事前学習パラメータには東北大学が公開するモデルを使用した.

4.1 生成の結果

表 4 に GPT-2 による生成結果の例を示す. この結果より, BERT の学習するデータセットを作成した.

生成されたデータには漫画的特徴よりも, 事前学習に含まれていたと思われる ブログ, 掲示板, 小説, SNS などの表現が多く生成されていた. 表 4 のように漫画のセリフとしては, コマに入らないような比較的長い文章が生成される傾向にあった.

本実験では入力するプロンプトとして 1 文のみを入力したが, より前の文章も与えることで漫画の特徴を掴んだ文章生成ができることが考えられる.

4.2 分類の結果

表 5 は BERT による分類の学習結果を示す. 2 値分類であるため正答率のベースラインは 0.5 であるが, テストデータで 0.959 という高い精度で分類できたと言える. この結果は BERT の文章の識別問題への有効性を示すとともに, 本実験設定での GPT-2 と人間による文章との間には大きな違いがあることを示唆している.

表 6 は分類結果の混合行列, 表 7 は BERT の分類の誤りの例である. 表 6 の結果から, 人間の文章を GPT-2 の文章と誤っている回数が多いことが分かる.

表 7 の label=1, pred=0 となっている部分を見ると, [SEP] トークンを挟んで文章が続いているパターンがあった. GPT-2 の生成した文章に頻繁にあるパターンであるため, 誤って識別したと考えられる. また 2 文のうち 1 方が, 呼びかけなどで短すぎる場合もあり, この情報だけで判断するのは人間とっても困難であると思われる.

反対に GPT-2 の文章を人間の文章と誤っている例を見ると, 文脈の短さも相まって一見自然な文章が生成されていると感じられた. 特に最初の例では, 時間的空白を表す文字が挟まっているにも関わらず, 文体の特徴を捉えながら意味のつながった文章が出力されている. この例から考えると, GPT-2 でも一部自然な文章が生成できているということが確認できた.

表 7: BERT の誤識別例

input	label	pred
[CLS] 極道をやっとりましたが足を洗い [SEP] ホストクラブや寿司屋で働いてましたが [EOS]	1	0
[CLS] というふうに去年は笑ってられたけど [SEP] 今年はそうもいきませんね [EOS]	1	0
[CLS] いつもどったんですか !? [SEP] よっ [EOS]	1	0
[CLS] ふふん [SEP] 正しいんだけど何かちがう気がする……[EOS]	1	0
[CLS] 大き ‥ さ ‥ など関 … 係 … ない [SEP] ‥ 事 … だと言う事 … [EOS]	0	1
[CLS] 立ちどまってもくれない [SEP] いや 立とうともしない [EOS]	0	1
[CLS] 極道の生態に親しむこのツアー [SEP] も 4 年目になる [EOS]	0	1
[CLS] 子供の頃から普通の男にはなるなと [SEP] 先生に言われていた [EOS]	0	1

ⓒ徹さん 川口 憲吾, ⓒOL ランチ さんり ようこ, ⓒ高校の人達 葛原 兄

5 まとめと今後の課題

本研究では人の考えた文章と GPT-2 が生成した文章との違いに注目し, 人間と比較して GPT-2 が生成した文章の自然さの確認を実験した. 文章の比較に BERT を利用して識別タスクを学習した.

結果 BERT は高い精度で人間と GPT-2 の文章を見分けることができるということがわかった. 一方でこの識別に関してはデータセットに使用した漫画という媒体が持つ特有の文章形式と GPT-2 の学習したデータの違いが識別に少なくない影響があったと考えられる. また GPT-2 の文章を人間の文章と誤っている例から GPT-2 が一部自然な文章生成ができることも確認できた.

今後の課題としては, GPT-3 を利用して同様の実験をし性能を比較することが挙げられる. また GPT-2 に与える入力文章を長くすることによって, GPT-2 がより自然な文章を出力できるようになる可能性もあるため, GPT-2 の設定を見直したい.

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. CoRR, abs/1706.03762, 2017.

[2] Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. Recurrent neural network based language model. In Eleventh annual conference of the international speech communication association, 2010.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[4] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications, 76(20):21811–21838, 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.