連載

5分で分かる!? 有名論文ナナメ読み

荒瀬由紀（大阪大学）

Tom B. Brown et al. : Language Models are Few-Shot Learners

Advances in Neural Information Processing Systems (NeurIPS), pp.1877-1901 (2020)

大規模言語モデルと転移学習

2018年に登場して以降，BERT（本連載2020年61巻3号で解説）を始めとする事前学習済み大規模言語モデルは自然言語処理分野に大きなインパクトを与えた．BERTを改善した後継モデルがいくつも発表され，現在では「定番」のモデル（たとえば元祖のBERT，さまざまなタスクで高い性能を示すRoBERTa，軽量のALBERT）が定着したように見える．自然言語処理ではこれら定番の事前学習済みモデルを特定タスクのラベル付きデータセットで転移学習（Fine-tuning）するアプローチが，最低限比較すべきベースラインとして定着している（そしてこれらシンプルな転移学習モデルの性能を上回るのは容易ではない）．

本稿で紹介するのは大規模言語モデルの1つ，GPT-3である．GPT-3は1,750億パラメータ（BERTはLargeモデルでも3億4千万パラメータ）を持つ巨大な言語モデルであり，約5千億トークンからなるテキストデータを用いて訓練されている（BERTは33億トークン）☆1．BERTは文中の単語の穴埋めと次の文の予測という事前学習を行っていたが，GPT-3は図-1に示す通り，文中の次の単語を予測する事前学習を行う．この論文が面白いのは「より大きな強いモデル」だからではなく，大規模言語モデルにまつわる多くの疑問に真摯に答えるさまざまな分析を実施し，巨大言語モデルがもたらした新たな社会的課題について論じている点にある．本論文は付録・参考文献を除いても約40ページのボリュームのため，以下の点に絞って紹介したい．

1. 人間は数個の例を見れば新しいタスクを実行できる．大規模言語モデルはどうなの？
2. 大規模言語モデルはすごいというけれど，事前学習で見たテキストを丸覚えしているだけじゃないの？
3. 大規模言語モデルがもたらす社会的課題とは？

転移学習の先へ

先述の通り事前学習済みモデルを使うには，対象タスクのラベル付きデータで転移学習を行うのが常套手段である．転移学習に必要なラベル付きデータの量はタスクによってさまざまであるが，文分類問題であれば数千〜数十万文のラベル付きデータが用いられることが多い．しかし現実的に大規模言語モデルを利用することを考えると，数千とはいえ，すべてのタスクに対しラベル付きデータを準備するのはなかなか大変である．また実用においてはラベル

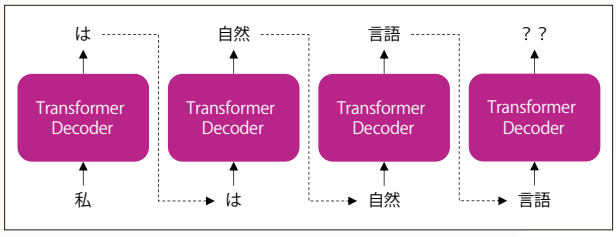

図-1　GPT-3の事前学習では次の単語の予測を行う．

☆1　2021年4月現在で最大の事前学習済みモデルは1.6兆パラメータを持つSwitch Transformer

の種類が定まっていない，ラベルの更新が頻繁に起こるということも考えられる．一方で人間は，初めて見る言語処理タスクであっても，多くの場合数千の例は必要としない．翻訳をしたことがない人であっても，「私は自然言語処理を勉強しています．=> I study natural language processing.」というような例を数例見れば「翻訳」というタスクが何を意味するのか理解し，翻訳を実際に行うことができるだろう．

タイトルにもある通り，本論文は大規模言語モデルがラベル付きデータを前提とする転移学習から脱却できる可能性を示している．図 -2 に示すように少数の例（10 ～ 100）のみ与える Few-shot 設定（パラメータ更新は行わない）により所望のタスクを達成できることが，質問応答から算術演算，機械翻訳など 24 を越えるデータセットを用いた網羅的な実験により示されている．

一例として機械翻訳では，フランス語・ドイツ語・ルーマニア語から英語への翻訳において，Few-shot 設定を用いた GPT-3 が，既存の（対訳データで訓練された）機械翻訳モデルと同等もしくはそれを上回る翻訳性能を達成している．GPT-3 が学習しているのは文中における次の単語の予測で，翻訳を学習しているわけではない．ではなぜ Few-shot での翻訳が可能なのか？　GPT-3 の訓練データの 93% は英語であるが，残り 7% は多言語データである．また日本語とその英訳など，対訳文を記載した Web ページは比較的多くみられる．そのため，Few-shot 設定での入力とは大きく異なるが，事前訓練において対訳文の生成を学習している可能性が高いことを著者らは議論している．フォーマットが異なるが事前訓練で目撃した翻訳という作業を，GPT-3 は Few-shot で与えられる例に基づき再現しているというのだ．

一方で，人工的に作成された意味のない単語を特定するタスクにおいても GPT-3 は Few-shot 設定で高い性能を達成してるが，このようなデータは事前訓練には存在しない．Few-shot 設定でなぜタスクを達成できるのか，詳細な分析が今後の重要な課題として述べられている．

大規模言語モデルは「記憶している」だけ?

多くの言語処理タスクの評価用データセットは Wikipedia など Web から収集したテキストを元に作られている．GPT-3 のように Web から収集した大規模テキストコーパスで事前学習した言語モデルは，タスクを実行する能力を獲得しているのではなく，事前学習で目撃したテキストを記憶しているだけなのではないか，という懸念は長らく議論されてきた．

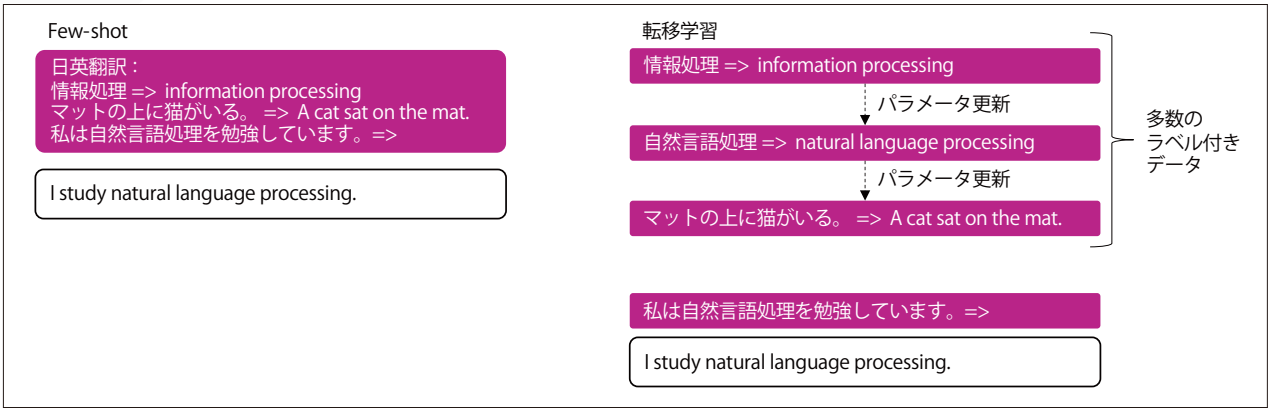

図 -2　Few-Shot 設定による予測と転移学習の違い：Few-shot 設定では少数の例を入力として与え予測を行うが，パラメータの更新を要する学習は行わない．転移学習ではラベル付きデータによる学習によりパラメータを更新することで，特定のタスクに適応したモデルを獲得する．

本論文では事前学習データから評価データセットとオーバラップしたテキストを除く処理と，残ったオーバラップテキスト☆2を評価データセットから除く処理により，事前学習データにオーバラップするテキストが存在しないクリーンな評価データセットを構築している．クリーンな評価セットでの性能と元の評価セットでの性能を比較したところ，クリーンな評価セットによる性能低下は無視できるレベルであったことが示されている．このことから，大規模言語モデルは事前学習中のテキストを正解として記憶している，という懸念は棄却されている．

巨大言語モデルがもたらした社会的課題

GPT-3がもたらした社会的課題として，(1) テキスト自動生成技術の悪用，(2) 生成テキストに現れるバイアス，(3) 環境への影響，が議論されている．

(1) について，GPT-3のように巨大な言語モデルは，すべての出力がそうでないとはいえ，人間が記述したものと区別するのが非常に難しいレベルのテキストを生成する．GPT-3にニュースのタイトルとサブタイトルを与え，生成した人工のニュース記事を人間が評価したところ，人工的に生成されたものであると見破れた精度は52%と報告されている．技術が進み，より高品質なテキストを自動生成できるようになると，このようなテキスト生成技術が悪用される懸念がある．本論文では人工的に生成したテキストを高い精度で識別する技術の必要性が議論されている．

(2) について，Webテキストを中心とした巨大なコーパスを学習した言語モデルは，現実に存在するバイアスを反映する．本論文では一例として，性別，人種，宗教に関するバイアスがGPT-3の生成するテキストに存在することを示しており，このほかにもさまざまなバイアスが存在するであろうことを示唆している．大規模言語モデルに含まれるバイアスを防ぐには，任意の評価指標を最適化するような画一的なアプローチではなく，まずバイアス問題を理解すること，そして問題全体を見据えた包括的なアプローチを取ることが必要であると述べられている．

最後に (3) について，巨大な言語モデルを訓練することの環境への影響についても議論されている．University of Massachusetts at Amherstの発表によると，大きな深層学習モデルの訓練は約284トンのCO2（車5台分の総CO2排出量に相当）を排出すると言われている☆3．本論文では言語モデルの訓練時だけでなく，応用時における計算効率の重要さを説き，GPT-3のような巨大なモデルからドメインやタスクに特化した軽量モデルを取り出すDistillation技術の重要性を議論している．

多くの研究者がこれら新たな社会的課題に懸念を示しており，今後解決に向けての研究が進展することが期待される．

(2021年4月20日受付)

☆2 余談であるが，本来すべてのオーバラップテキストを事前学習データから除くはずだったが，バグにより部分的にしか除けていなかったという記述があり，苦労が伺える（この規模のモデルになると気軽に再訓練は不可能なのである）．

☆3 https://news.mit.edu/2020/shrinking-deep-learning-carbon-footprint-0807

荒瀬由紀（正会員）
arase@ist.osaka-u.ac.jp
2010年大阪大学大学院情報科学研究科博士後期課程修了．博士（情報科学）．同年，北京のMicrosoft Research Asiaに入社，自然言語処理に関する研究開発に従事． 2014年より大阪大学大学院情報科学研究科准教授，現在に至る． 言い換え表現抽出，言語学習支援技術，対話システムに興味を持つ．