日本語自然言語処理における事前学習モデルの公開

Release of Pre-trained Models for Japanese Natural Language Processing

趙 天雨　　沢田 慶 *
Tianyu Zhao　　Kei Sawada

rinna 株式会社
rinna Co., Ltd.

Abstract: We have developed two types of pre-trained models, GPT-2 and RoBERTa, that are trained from a public corpus consisting of about 75-gigabyte texts. The models and its training code have been released under licenses that allow for commercial use. By fine-tuning the released models, users will be able to accomplish a variety of Japanese natural language processing tasks with high task accuracy.

1　はじめに

人間の対話におけるコミュニケーションでは，自分の考えを言語化し，声や表情・身振り手振りによる表現で相手に情報を伝える．また，コンピュータを介したコミュニケーションでは，テキストによる情報伝達が多くを占めている．そのため，言語化された情報やコンピュータとインターネット上に存在するテキストを扱う自然言語処理の研究は重要であり，様々な応用が期待される．

近年では，コンピュータの処理速度・ストレージサイズ・通信速度の向上により，大規模なテキストデータを扱うことが可能となった．さらに，ディープラーニング技術の発展に伴い，大量のパラメータを持つモデルを効率よく学習する手法が提案された．そして自然言語処理の分野において，大規模なテキストデータから学習される大量のパラメータを持った事前学習モデルの登場はブレークスルーをもたらした．

2018 年に提案された Generative Pre-trained Transformer (GPT) [1] や Bidirectional Encoder Representations from Transformers (BERT) [2] は，自己教師あり学習により事前学習される言語モデルであり，事前学習モデルと呼ばれる．事前学習モデルは，大規模なテキストを用いた事前学習により，言語としての一般的な表現をモデルに学習させることができる．そして，目的に合わせたモデル構造の選択や，適切なデータを利用した事前学習モデルの fine-tuning により，自然言語処理の様々なタスクを高い精度で実現する．

*連絡先： rinna 株式会社
東京都渋谷区渋谷 2 丁目 24 － 12
渋谷スクランブルスクエア 39F WeWork
E-mail: keisawada@rinna.co.jp

事前学習モデルを用いた手法は，より大量のパラメータを持つモデルの提案により日々改善されている．しかし，パラメータ数の増加は性能向上をもたらすものの，事前学習のためのコストが膨大に膨れ上がる．そこで我々は，日本語オープンコーパスから学習した GPT-2 [3] と Robustly Optimized BERT Pretraining Approach (RoBERTa) [4] の 2 種類の事前学習モデルを Hugging Face[1]に，事前学習に用いたソースコードを GitHub[2]に商用利用可能なライセンスで公開した．これにより事前学習モデルの利用者は，目的に合わせた fine-tuning を行うことで，事前学習のコストを抑えつつ所望のタスクを実現可能となる．

2　公開した事前学習モデル

オープンコーパスである CC-100 [5] と Wikipedia の日本語テキストを用いて，GPT-2 と RoBERTa の事前学習モデルを学習し，公開した．以下に公開した事前学習モデルについての説明をする．

2.1　モデル構造

GPT-2 と RoBERTa は，主に self-attention 層と feed-forward 層からなる Transformer ブロックの積み重ねから構成される [6]．GPT-2 と RoBERTa には，各ブロック内の feed-foward 層と正規化層の配置などモデル構造に細かな違いはあるが，主要な違いは 2 つのモデルが異なるタスクのために異なったマスキングメカニズムを利用する点である．

[1] https://huggingface.co/rinna
[2] https://github.com/rinnakk/japanese-pretrained-models

Figure 1: The training process of GPT-2 (left) and RoBERTa (right).

Table 1: Pre-trained model configurations.

Pre-trained models	Params	Layers	Emb dim	Epochs	PPL	Time
rinna/japanese-gpt2-medium	336M	24	1024	4	18	45 days
rinna/japanese-gpt2-small	110M	12	768	3	21	15 days
rinna/japanese-gpt2-xsmall	37M	6	512	3	28	4 days
rinna/japanese-roberta-base	110M	12	768	8	3.9	15 days

GPT-2 は自己回帰言語モデルであり，Figure 1 の左側のように，これまでのステップのトークン（単語やサブワード）から次のトークンを予測するように学習される．そのため GPT-2 は，条件付き又は無条件のテキスト生成のためのデコーダとして用いられる．

一方，RoBERTa はマスク言語モデルであり，Figure 1 の右側のように，コンテキストとして前後のトークンが与えられた場合に，マスクされたトークンを予測するように学習される．RoBERTa は，マスクトークンを予測するように学習されるが，意味のあるテキスト表現を取り出すためのエンコーダとして機能するため，テキスト分類や検索など下流タスクのための事前学習モデルとしてよく利用される．

2.2 学習データ

日本語のオープンコーパスで最も大規模であり整備がなされている CC-100 と Wikipedia を用いて，GPT-2 と RoBERTa を学習した．CC-100 コーパスは，様々なトピックとスタイルのテキストを含んでおり，日本語は約 70GB のテキストデータから構成される[3]．Wikipedia コーパスについては，学習時に最新の Wikipedia からダンプされた約 5GB の日本語テキストを利用した[4]．また，トークナイザは語彙サイズ 32000 の SentencePiece [7] とし，テキストをサブワード化した．事前学習にオープンコーパスを用いているため，公開したソースコードから利用者は事前学習を再現することができる．

2.3 学習結果

コンピュータリソースとモデル性能はトレードオフの関係にあり，利用者が利用条件に合ったモデルを柔軟に選択できるように，サイズが異なる事前学習モデル (GPT-2 が 3 サイズと RoBERTa が 1 サイズ) を公開した．Table 1 は，それぞれの事前学習モデルの主要なハイパーパラメータと学習結果である．全ての事前学習モデルは，8 台の NVIDIA V100 (32GB メモリ) が搭載された GPU サーバで学習した．学習時間は最長で 45 日と十分に学習されており，汎用性があるモデルとなっている．

[3] http://data.statmt.org/cc-100/

[4] https://dumps.wikimedia.org/other/cirrussearch/

3 むすび

本稿では，日本語自然言語処理のために公開した GPT-2 と RoBERTa の事前学習モデルについて紹介した．今後は，利用者のモデル選択の幅を広げるために，他のモデル公開について検討する．

参考文献

[1] Radford, A. *et al.*: Improving Language Understanding by Generative Pre-training (2018)

[2] Devlin, J. *et al.*: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *NAACL-HLT 2019*, pp. 4171–4186 (2019)

[3] Radford, A. *et al.*: Language Models are Unsupervised Multitask Learners (2019)

[4] Liu, Y. *et al.*: RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv preprint*, arXiv: 1907.11692 (2019)

[5] Conneau, A. *et al.*: Unsupervised Cross-lingual Representation Learning at Scale, *ACL 2020*, pp. 8440–8451 (2020)

[6] Vaswani, A. *et al.*: Attention Is All You Need, *NeurIPS 2017*, pp. 5998–6008 (2017)

[7] Kudo, T. *et al.*: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, *EMNLP 2018*, pp. 66–71 (2018)