ニューラル言語モデルの効率的な学習に向けた代表データ集合の獲得

鈴木 潤[1,2*]　全 炳河[1]　賀沢 秀人[1]
[1]Google 合同会社　[2]東北大学
{junsuzuki, heigazen, kazawa}@google.com

概要

本稿では，大規模な学習データから選択した少量の代表データ集合を用いてニューラル言語モデルを学習した際に，元の学習データ全てで学習した場合と同等な性能を達成できるかという研究課題について検証する．実験では，二つの特性の違う言語モデルの尤度差に基づく選別方法により代表データ集合を獲得し，GLEU ベンチマークデータで評価をおこなった．元データの 20 分の 1 から 50 分の 1 程度まで削減した代表データ集合から学習したニューラル言語モデルの GLUE 平均スコアが，全データから学習した場合と同等であることを示す．

1　はじめに

自然言語処理の研究分野では，2017 年頃からニューラル言語モデルの研究が注目され，以降数多くの研究成果が報告されている [1, 2, 3, 4, 5, 6]．これまでに発表された多くの研究において，ニューラル言語モデルを事前学習済みモデルとして解きたいタスクに特化したモデルに組み込むことで，事前学習済みニューラル言語モデルを用いない場合と比較して飛躍的に性能を向上できることを示している．つまり，特定のタスクに依存しない大規模な生テキストから学習した事前学習済みニューラル言語モデルは，多種多様な自然言語処理タスクの汎用特徴量 (universal feature) として効果的に機能することを示している．このことからも，ニューラル言語モデルは，昨今の自然言語処理の成功の根幹を担う必須の基盤技術と言える．

事前学習済みニューラル言語モデルに関しては，最近の様々な研究により，学習データ量，および，モデルサイズを大きくすることが比較的一貫性をもって性能を向上できる二大要因であることが検証され，実験的に立証されている [2, 3, 6]．ただし，その性能向上の効率に関しては，例えばデータ量やモデルサイズに対して概ね対数線形程度で効果が得られる場合が多いことも同時に知られている [7, 8]．つまり，あるデータ量で得られる性能からデータを増やして得られる性能向上と同程度の性能向上をさらに目指す場合は，増やしたデータの 10 倍増やす必要がある．このことから，より性能の高い言語モデルを獲得するには，膨大な量の学習データ，および，その処理に必要な大規模な計算リソースが必要になることを意味する．実際に，性能の高い事前学習済みニューラル言語モデルは，潤沢な計算リソースを持つ大手企業や研究機関からリリースされることがほとんどであり，例えば大学の研究室などの多くの計算リソースや研究資金を持たない場合は，高性能な事前学習済みニューラル言語モデルを構築するのは非常に困難である．このようにな現状から，事前学習済みニューラル言語モデルのような重要な基盤研究を，多数の研究機関で取り組むことができない状況が発生していると考えられる．重要な要素技術の研究に対して広く研究者が参加できないのはあまり適切な状況とは言えない．

そこで，本稿では，ニューラル言語モデルの学習データに着目し，大規模なニューラル言語モデルを学習する際に用いるデータから，同等かそれ以上の性能を持つニューラル言語モデルを学習できる部分集合を抽出できるかについて検証をおこなう．本稿では，便宜上，ある特定のデータセットに対して，そのデータセットの性質を適切に保持する部分集合を「**代表データ集合**」と呼ぶこととする．もし，効果的な代表データ集合を抽出できれば，現実的な計算リソースおよび研究資金で事前学習済みニューラル言語モデルの研究ができるようになり，より多くの研究者が参加し分野の発展がより早く進むことが期待できる．

*Google Visiting Researcher として実施した研究成果

2 関連研究

本研究と同様に，少ない計算資源と研究資金で事前学習済み言語モデルの研究を実現することを主目的とした取り組みが既になされている [9]．この論文では，24 時間で BERT[2] の学習を実現するために必要な学習時の設定やモデルの改良などを示している (通称 24hBERT)．この論文により，計算リソースが少ない環境でもニューラル言語モデルの効率的に学習する知見が数多く示されている．

これまでのニューラル言語モデルの研究では，性能向上の方策の一つとしてデータ量を増やしていった．しかし，最近では，Common Crawl[1] などの現在比較的簡単に取得可能な生テキストのうち最大級のデータ（web テキスト）を使うのが当たり前となり，これ以上は安易にデータが増やせない状況になってきた．このように，データ量増加に関して頭打ちになりつつあることから，次に量から質を高める方向が注目されるようになった．実際に，データの質を制御あるいは向上させることで，最終的なニューラル言語モデルの性能が向上したという報告も散見されるようになった [10]．従来のように方法論としてモデルの改良ではなく，データに焦点をあてデータを改良することで最終的なタスク性能を向上させようという取り組み（例：[11, 12, 12]）を，最近ではデータ中心 AI（Data-centric AI）研究と総称し新たに注目すべき研究カテゴリとなっている[2]．本研究も，こういったデータに着目した研究の一環と捉えることができる．

3 代表データ集合の選別

本来，この代表データ集合という用語を定義するために，データセットの性質と計測方法など上記の説明を定量的かつ明確に計算するための定義が必要であるが，ここでは直接的な定義は設けない．その代わり，目的タスクにおいて，データ全体を用いた時と同等かそれ以上の性能を達成することができる部分集合のデータと定義する．つまり，直接の計測は困難なので，目的タスクの性能をもって，代表データ集合が元のデータ集合の性質を保持していると間接的に計測する方法をとる．本稿においては，言語モデルの性能について評価することが目的であるため，言語モデルの性能評価に一般的に用いられるベンチマークデータの性能を比較することで，効果的な代表データ集合が獲得できたかを判断する．

代表データ集合を選出する方法論には，多くの方法が考えられる．ここでは，簡単な方法論として文献 [13] にて提案された尤度差に基づく方法論を取り上げる．注意点として，本手法は全く何もない状態から代表データを抽出する汎用の方法ではない．また，それを目指しているわけでもない．あくまでも最新のニューラル言語モデルが存在し，その学習が可能な環境があることを前提とし，それを出発点として，いかに有用な代表データ集合を選別できるかを考える．これは，代表データ集合を一度作成することができれば，以降はこのデータを使うことで，少ないデータから高品質な言語モデルを作成できるようになり，言語モデルそのものの研究を限定された計算環境しかない場合でも実行できるようにしたい，という目的を達成するためである．

3.1 尤度差に基づく選別

言語モデルに効果的なデータかどうかを判断する方法として，文献 [13] にて提案された二つの言語モデルの尤度差を利用する方法を，本研究の目的に合わせた形で流用する．この方法は，特定のドメインに特化したデータ (in-domain data) から学習した言語モデルと，特定のドメインに特化しないデータ (non-domain-specific data) から学習した言語モデルにより計算される対象テキストの尤度差により，特定のドメインにより適したデータを選別する方法である．以降ここでは，便宜上，前者を**特化型言語モデル**，後者を**汎用言語モデル**と呼ぶことにする．また，評価対象となる文章を X とする．このとき，特化型言語モデル $\mathscr{M}_I$ による X の尤度を $L_{\mathscr{M}_I}(X)$，汎用言語モデルによる X の尤度を $L_{\mathscr{M}_N}(X)$ とすると，尤度差に基づくスコア S_L は以下の式により計算できる．

$$S_L = L_{\mathscr{M}_I}(X) - L_{\mathscr{M}_N}(X) \tag{1}$$

ただし，ここでの $L_{\mathscr{M}_I}(\cdot)$，$L_{\mathscr{M}_N}(\cdot)$ は．文章中に出現する各単語の対数尤度の和を用いる．

前に述べた通り，本研究では「少ないデータから学習したニューラル言語モデルでも性能が高い」ことが目的である．そこで，特化型言語モデルとして，事前学習済みニューラル言語モデルを用いて，ある自然言語処理タスクを実行した際に，より高い性能が得られると考えられるデータで学習された

1） https://commoncrawl.org
2） 参照例：https://datacentricai.org

ニューラル言語モデルと仮定する．具体的に利用する学習データに関しては，実験（4 節）にて述べる．また，汎用言語モデルは，通常の手順により得られるニューラル言語モデルと仮定する．

3.2 ランキング

本研究において実際に実現したいことは，代表データ集合の獲得であるが，扱えるデータ集合の量は実際には各ユーザの計算機環境や研究資金に依存して決まるので，事前に規定するのは難しい．そこで今回は，代表データ集合として適切かどうかを表すスコアを個々のデータに付与する方法を採用する．これにより，実際に利用したいユーザの環境に応じて，そのスコアの順番に従って上位のデータを取得することで，代表データ集合を利用者が比較的自由に取得できる仕組みとする．

4 実験

本稿では，従来通り大規模なデータを使って学習した通常の事前学習済みニューラル言語モデルと，代表データ集合を使って学習したニューラル言語モデルの性能を比較し，本研究で用いた代表データ集合の選出方法が有効か検証することを目的に実験を実施する．以降，簡単のため，3 節の方法で得られる代表データ集合を RepSet と表記する．

4.1 ベースラインニューラル言語モデル

本実験では，ベースとなるニューラル言語モデルとして Text-to-text Transfer-Transformer (T5)[3] を用いた．ニューラル言語モデルの性能は，モデルサイズにより大きく変わることが知られている．本実験では，バージョン t5.1.1[3] に従って Small (≈77M パラメタ), Base (≈250M パラメタ), Large (≈800M パラメタ), XL(≈3B パラメタ) の 4 モデルを用いた．モデルパラメタや学習時のハイパーパラメタの設定は，基本的に上記文献およびサイトで配布されている設定に従う．

4.2 学習用/評価用データセット

本実験で用いるデータセットは 2 種類ある．一つは，ニューラル言語モデルの学習用データである生テキストの集合であり，もう一つは，言語モデルの性能評価に用いるベンチマークデータである．

3） https://github.com/google-research/text-to-text-transfer-transformer/blob/main/released_checkpoints.md

表 1 C4 データ，および，代表データ集合 (RepSet) に関する統計量．#docs: 文書数，#words: 単語数（トークン区切り未使用），ratio: C4 (default) に対する filesize での比率．

	#docs	#words	filesize	(ratio)
C4 (default)	364.9M	132.0B	745GB	–
RepSet-1	2.8M	1.0B	6GB	(≈ 1/124)
RepSet-2	7.4M	2.7B	16GB	(≈ 1/47)
RepSet-3	16.2M	5.9B	35GB	(≈ 1/21)
RepSet-4	35.2M	13.9B	75GB	(≈ 1/10)

まず，ニューラル言語モデルの学習用データとして，Colossal Clean Crawled Corpus (C4) データ [3][4] を用いた．また，ニューラル言語モデルの評価用ベンチマークデータとしては，GLUE データセット [15] を用いた．本実験の全てのデータは Tensorflow Datasets (tfds) から直接呼び出して利用した．

ここで注意点として，本実験の評価は全て開発用データの結果である．本実験では，データの傾向を調査することが目的であり，設定を変えて多数の実験を実施することになる．このような場合には，評価用データを用いるのは不適切と考えられるため[5]，開発用データにて評価するのは妥当と考えられる．

4.3 実験で用いる設定

RepSet を得るには，式 1 に示した特化型ニューラル言語モデル $\mathcal{M}_I$ と汎用ニューラル言語モデル $\mathcal{M}_N$ が必要となる．本実験では，それぞれの目的に合った学習データを用意することで 2 種類のニューラル言語モデルを区別して構築する．まず，汎用言語モデル $\mathcal{M}_N$ の学習には標準的な C4 データを用いた．次に，特化型ニューラル言語モデル $\mathcal{M}_I$ の学習用には，C4 の部分集合として事前に定義されている c4/realnewslike と c4/webtextlike を選択した．更に，これらのデータに文献 [10] に紹介されている重複削除 (deduplication) の処理を適用した．これらのデータ集合から学習した $\mathcal{M}_I$ および $\mathcal{M}_N$ を用いて式 1 により C4 データ内の各事例のスコア付け，および，ランキングを実施した．

次に RepSet として，4 つのサイズの違う集合 RpeSet-1, 2, 3, 4 を用意した．表 1 に，C4 の全体のデータサイズ，および，作成した RepSet の規模に関する統計量を示す．更に，RepSet と同じデータ量

4） https://www.tensorflow.org/datasets/catalog/c4, また文献 [14] にも詳細情報あり．

5） 多数の実験により評価用データの傾向が判明してしまう点や，評価用データに対してチューニングすることに相当してしまうため．また，GLUE の評価用データはリーダーボードに投稿し，他のシステムと比較するためのデータである．

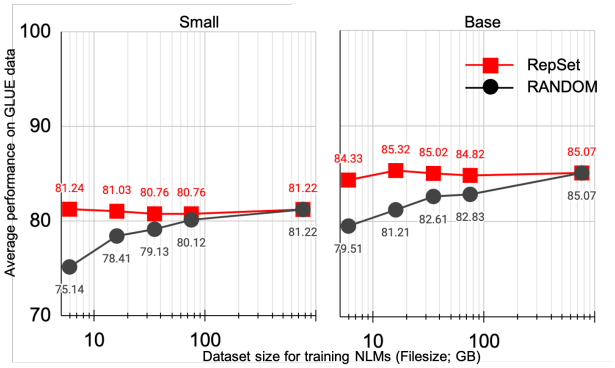

図 1　データ選択基準 RepSet と RANDOM の違いによる平均 GLUE スコアの比較．図中のプロット点は左から RpeSet-1, RpeSet-2, RpeSet-3, RpeSet-4, C4 (default) の値．

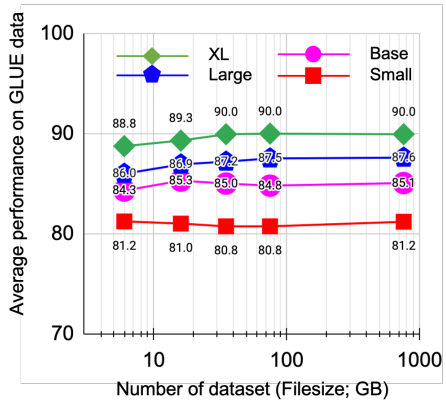

図 2　モデルサイズの違いによる平均 GLUE スコアの比較．図中のプロット点は左から RpeSet-1, RpeSet-2, RpeSet-3, RpeSet-4, C4 (default) の値．

をランダムに抽出したデータ集合（以下，このデータ集合を RANDOM と呼ぶ）を用意した．

RepSet または RANDOM を使い事前学習済みニューラル言語モデルとして T5 をそれぞれ学習した．その後，GLUE の学習データを全て統合し一括でファインチューニングした[6]最終的に得られたファインチューニング済みモデルを GLUE ベンチマークデータを用いて比較することで，間接的に RepSet の有効性を検証する．

4.4　実験結果および検証

図 1 に，データ選択基準 RepSet と RANDOM の違いによる平均 GLUE スコア（縦軸）とデータ量（横軸：対数スケール）の関係および比較を示す．モデルサイズとしては Small と Base の結果である．次に，図 2 にモデルサイズ Small, Base, Large, XL に対する RepSet の結果をプロットした．

6）　タスク毎にファインチューニングしタスク毎の特化モデルを構築するのではなく，全タスクの評価に一つのモデルを使用する設定である．

本実験においては，基本的な性質としてデータ量は多ければ多いほど性能は安定的に高くなるという傾向が見られた．同様に，モデルサイズに関しても，モデルサイズが大きくなればなるほど性能が高くなった (Small < Base < Large < XL)．これらの結果は直感やこれまでの多くの研究成果と同じ傾向の結果と言える．

次に，データを選択する際に RepSet を使うことで RANDOM よりも良い結果が得られることがわかった．つまり，少ないデータ量の設定でも RepSet が効果的なデータを選択できていることを示唆している．例えば，RepSet を用いることで，データ量（ファイルサイズ）を 21 分の 1 程度までなら Small から XL まで全てのモデルサイズで性能を維持できるという結果になった．このことから，T5 の論文と同等の実験をしようと思った際に，21 分の 1，あるいは，Base であれば 47 分の 1 程度のデータ量でも同様の実験ができることが期待できる．ただし，モデルサイズが大きくなると，データ選択でデータ量を減らした際に，性能の劣化が早く訪れることが観測されたので注意が必要である（例：XL の RepSet-3 と 4)．

RANDOM によりデータ量を削減する場合に，事前の予測ほど性能の低下は大きくなかった，これは，元データに冗長性があることが想定される．よって，冗長性を排除する方法論を組み合わせることで，更に性能向上できる可能性が考えられる．

5　おわりに

本稿では，大規模な学習データから代表データ集合を選択しニューラル言語モデルを学習した際に，元の学習データ全てで学習したニューラル言語モデルと同等な性能を達成できるか，という研究課題の検証をおこなった．本稿の実験では，尤度差に基づくランキングにより獲得した代表データ集合は，元データの 21 分の 1 程度のデータ量でも GLUE の平均性能の観点で同等の性能が得られることを示した．また，ランダム選択の場合は，データ量を減らした場合に顕著に性能が低下することも示した．これらの結果から，少ない計算リソースと研究資金しかない研究組織においても最先端の事前学習済み言語モデルの研究をやりやすくする，という本研究の最終目的の実現可能性を示した．代表データ集合を利用することで，事前学習済み言語モデルを改善する研究が更に推進することを期待する．

謝辞

本研究に関して Google 合同会社 澤井裕一郎氏にアドバイスを頂きました．感謝いたします．

参考文献

[1] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[5] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[7] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, 2020.

[8] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling Laws for Autoregressive Generative Modeling, 2020.

[9] Peter Izsak, Moshe Berchansky, and Omer Levy. How to Train BERT with an Academic Budget. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 10644–10652, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[10] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating Training Data Makes Language Models Better, 2021.

[11]

[12] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset Distillation with Infinitely Wide Convolutional Networks, 2021.

[13] Robert C. Moore and William Lewis. Intelligent Selection of Language Model Training Data. In **Proceedings of the ACL 2010 Conference Short Papers**, pp. 220–224, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[14] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus, 2021.

[15] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.