トピックエントロピーに基づく学習データ選択による事前学習言語モデルの訓練安定性向上

永塚光一　渥美雅保
創価大学大学院 理工学研究科
e20d5301@soka-u.jp　matsumi@soka.ac.jp

概要

本稿では，事前学習言語モデルの訓練の安定化を目的として，新たな学習データ選択手法である Topic Entropy-based Data Selection (TEDS) を提案する．TEDS では，テキストに含まれるトピックの多様性指標であるトピックエントロピーを定義した上で，(i) トピックエントロピーの高い一部の学習データによる訓練，(ii) 全学習データによる訓練という 2 段階の事前学習を行う．提案手法の効果を検証するために，検証損失に基づく訓練安定性と GLUE ベンチマークを用いた下流タスクの性能を評価した．実験により，TEDS を用いたモデルはベースラインと比較して検証損失が安定的に減少し，より高い汎化性能を獲得することを示した．

1　はじめに

近年，事前学習言語モデル[1)] (Pre-trained Language Model，PLM) は自然言語処理分野において大きな成果を上げている [1, 2]．PLM は，小規模なラベル付きデータセットを用いてファインチューニングを行うことにより，様々な自然言語処理タスクにおいて最高精度を達成している．一方で，大規模なコーパスを用いた事前学習は計算コストが非常に高く，特に訓練が不安定である場合には，計算コストが更に増加するという問題がある．

PLM の学習効率を向上させるために，これまでに多くの研究が行われている．代表的な例として，よりパラメータ数の少ないモデルアーキテクチャを設計するアーキテクチャベースの手法 [3, 4, 5, 6, 7]，効率的な事前学習を行うために新たな目的関数を定義するオブジェクティブベースの手法 [8, 9]，更に．効果的な学習データ選択と学習スケジュールの実行を通して，学習効率を向上させるデータセットベースの手法がある [10, 11, 12]．アーキテクチャベースやオブジェクティブベースの手法は既に PLM の学習効率向上に大きく貢献している一方で，データセットベースの手法はまだ比較的研究が少ない．

代表的なデータセットベースの手法として，カリキュラム学習 [12] が挙げられる．カリキュラム学習は，定義された難易度の指標に基づき，簡単な学習サンプルからより難易度の高い学習サンプルへと訓練を徐々に移行させることにより，モデルの学習効率を高める手法である．PLM にカリキュラム学習を適用するためには，学習データとなるテキストに対して，難易度の指標を設計する必要がある．テキストの難易度を定義することは，言語学の分野では Readability の研究として知られており，テキストの難易度を決める要因は，(1) 語彙レベル，文法構造の複雑さ，単語や文の長さなどのミクロ要因，(2) テキストのトピックや一貫性などのマクロ要因に分類される [13]．ここで，(1) ミクロ要因に関しては，語彙頻度 [12] やテキストの長さ [10, 11] を活用した手法が提案されているが，(2) マクロ要因については十分に研究が行われていない．

本稿では，学習サンプルの難易度指標としてマクロ要因の一つであるトピックに着目し，新たな学習データ選択手法である Topic Entropy-based Data Selection (TEDS）を提案する．TEDS では，学習効率が高い (難易度が低い) テキストは多様なトピックから構成されるという仮説に基づき，学習サンプルのトピックの多様性を計測する新たな指標として，トピックエントロピーを定義する．テキストの潜在的なトピックを推定する手法には，トピックモデルの一つである Latent Dirichlet Allocation (LDA) を用いる．実験により，TEDS を用いたモデルでは，ベースラインと比較して訓練の安定性及び下流タスクにおける汎化性能が向上することを示す．

1）本稿では，GPT のような自己回帰型言語モデルではなく，BERT に代表される自己符号化型言語モデルを指すこととする．

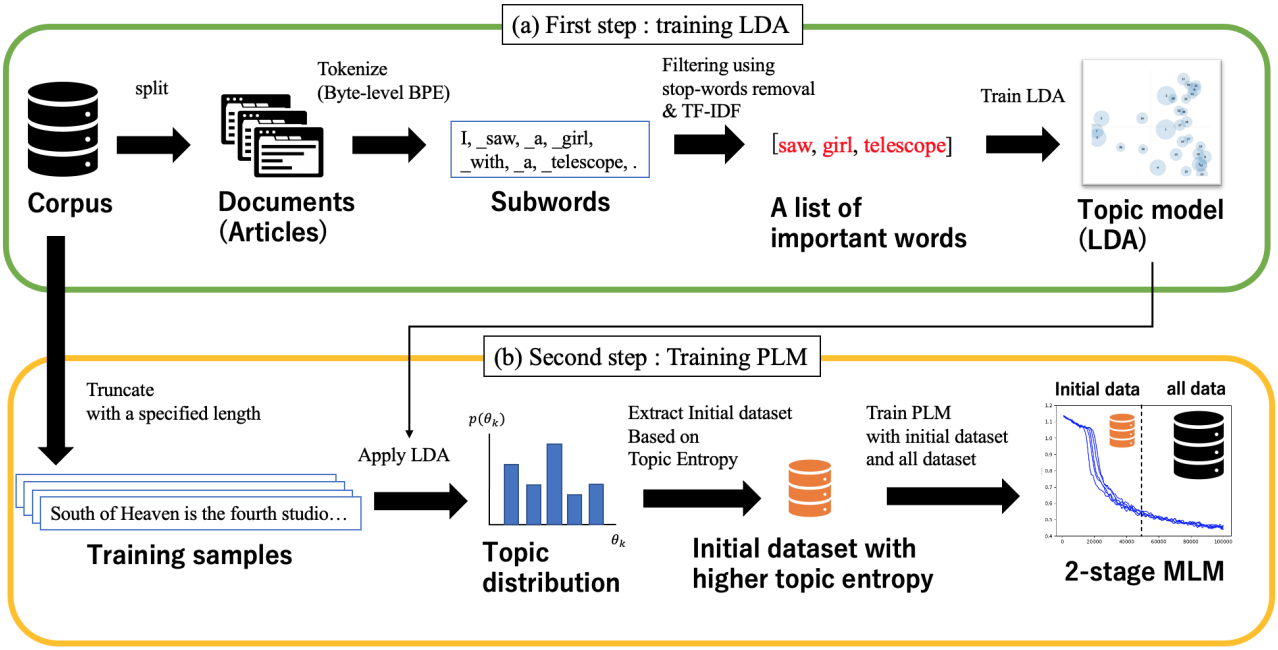


図 1: TEDS の全体像

2 関連研究

近年の研究では，大規模な言語モデルの学習効率化にデータセットベースの手法が用いられている．例えば，GPT-3[14] や T5[15] などの高コストな事前学習を伴うモデルでは，複数のコーパスから指定されたサンプリング比率で学習データを収集することでデータセットの品質を高めるデータ選択手法を採用している．また，テキストの長さを難易度の指標としてカリキュラム学習を GPT-2 や BERT の事前学習に適用する研究が行われている [10, 11, 16]．ここで，標準的なカリキュラム学習では難易度の指標に加えて，学習データを徐々に増加させるためのスケジューラを設計するが，我々の TEDS は，二段階のみで構成されるよりシンプルな手法であり，こうしたスケジューラが提案手法に与える影響については検証しない．したがって，提案手法をカリキュラム学習とは呼称しないこととする．

3 提案手法

図 1 に示すように，TEDS は 2 つのステップから構成される．1 つ目のステップでは，コーパスに含まれる文書集合を用いて，LDA の訓練を行う (図 1a)．2 つ目のステップでは，訓練した LDA を用いて各学習サンプルのトピックエントロピーを計算する．続いて，トピックエントロピーの高い一部の学習サンプルを抽出したのち，抽出データと全学習データを用いて，2 段階の事前学習を行う (図 1b)．LDA 及び PLM の訓練には，Wikipedia の英語記事から構成されるコーパスである WikiText-103[17] を用いる．

3.1 LDA の訓練

LDA を訓練するために，WikiText-103 から各記事をあらかじめ定義されたフォーマットに従って抽出し，記事集合を得る．続いて，Byte-level BPE トークナイザ [18] を用いて各記事をサブワードに分解したのち，ストップワード[2)]除去を行う．また，TF-IDF によるフィルタリングにより，語彙数を約 50%削減する．ここで，LDA は単語ではなくサブワードに基づいて学習される．既存のトピックモデルでは，単語による訓練が一般的であるが，提案手法では Transformer ベースの言語モデルの入力形式との一貫性を保つために，サブワードを用いる．

3.2 PLM の訓練

3.2.1 トピックエントロピー

トピックエントロピーは，与えられたテキストに含まれるトピックの多様性を測定するために我々が提案した新たな指標である．TEDS では，少数のト

2) NLTK ライブラリ (https://www.nltk.org/) を使用

ピックのみが存在する学習サンプルよりも多様なトピックを含む学習サンプルを選択することが，言語モデルの初期の訓練安定化に有効であると仮定する．学習サンプルを構成するトークン系列 $\boldsymbol{x}$ が与えられた時，トピックエントロピーは以下の式で計算される．

$$TE(\boldsymbol{x}) = -\sum_{k=1}^{K} p(k \mid \boldsymbol{x}) \log p(k \mid \boldsymbol{x}) \tag{1}$$

ここで，K はトピックサイズ，$p(k \mid \boldsymbol{x})$ は LDA によって計算された事後確率分布におけるトピック k の条件付き確率である．各学習サンプルに対して，トピックエントロピーを計算し，トピックエントロピーが高い上位の学習サンプルを抽出することで，初期の事前学習に用いる学習データセットを構築する．

3.2.2 2 段階事前学習

学習する言語モデルとして，本研究では，RoBERTa[2] を採用する．TEDS の事前学習は 2 つのサブステージから構成される．まず，トピックエントロピーの高い学習サンプルを用いて，総トレーニングステップ数の半分まで RoBERTa を訓練する．次に，全ての学習サンプルを用いて，トレーニングステップ数の終了まで RoBERTa の訓練を継続する．学習データには，LDA の訓練と同様に WikiText-103 を用いる．また，学習サンプルのトークン数は 128 に設定する．

4 実験

4.1 実験設定

提案手法を評価するために，訓練安定性と汎化性能という 2 つの観点から，TEDS とベースラインの比較を行う．ベースラインとして，通常の事前学習と同様にはじめから全ての学習データを用いる All Data Selection (ADS) 及びランダムに学習データを抽出する Random Data Selection (RDS)，語彙頻度に基づくカリキュラム学習である Vocab CL[12] を設定する．Vocab CL は，カリキュラム学習の研究の初期において提案された手法であり，単純な FFN 層から構成されるニューラルネットワークに基づく N-gram 言語モデルの訓練効率化に有効であることが報告されている．Vocab CL ではコーパス中での頻度を基に語彙をランキングしたのち，最初のステージで低頻度語彙 (希少語彙) を含まない学習データを抽出してモデルに与える．

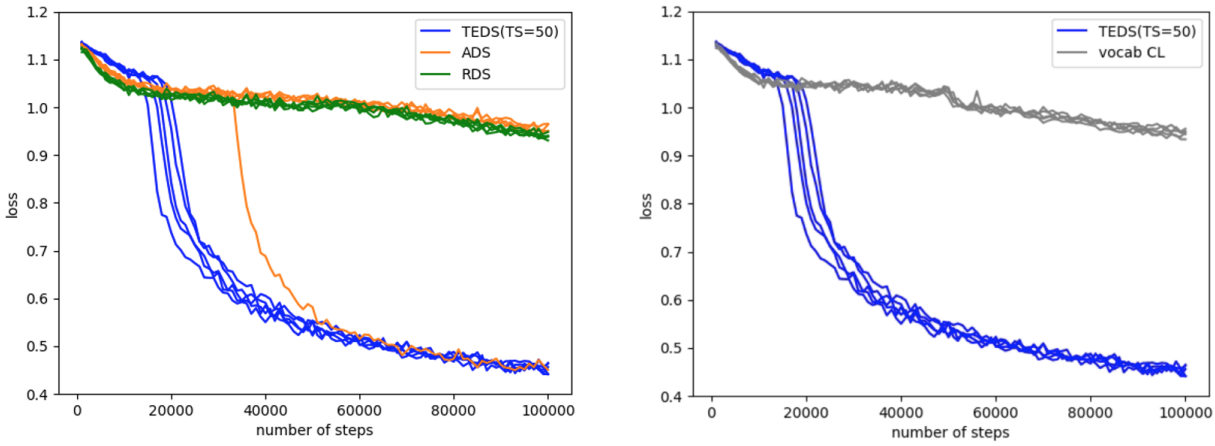
LDA のトピックサイズは 50 に設定し，TEDS 及び RDS において抽出する学習データサイズは，全データの 25%とする．また，Vocab CL では，低頻度語彙の定義を下位 10%の頻度の語彙とすることで，TEDS 及び RDS と同等程度のデータサイズである約 27%の学習データを抽出する．訓練安定性を評価するために，各モデルに対して異なるランダムシードを用いて 5 回ずつ事前学習を行う．総トレーニングステップ数は 10 万回とし，5 万トレーニングステップで全学習データに訓練を移行させる．事前学習を行った各モデルの汎化性能の計測には，GLUE ベンチマーク [19] における 9 つの下流タスクを採用する．ここで，TEDS の後半の訓練において，全学習データを追加する効果を検証するために，全学習データに訓練を移行させないモデルの汎化性能も評価する．PLM のファインチューニングに関しても，異なる 3 つのランダムシードで訓練を行ったのち，各下流タスクの平均スコアを評価する．全てのモデルに共通のモデルアーキテクチャとして，12 層と 12 個の注意ヘッドから構成される RoBERTa-base[3] を用いる．トークナイザの訓練により獲得された語彙数は 29,833 であり，モデルの入力トークン数は 128，ミニバッチ数は 64 とする．マスキング率は 15%に設定し，最適化手法として，AdamW[20] を使用する．

4.2 実験結果

4.2.1 訓練安定性の評価

図 2 に TEDS とベースラインの学習曲線を示す．図 2a からわかるように，ADS と RDS の検証損失は少しずつ減少し続けているものの，ADS における一度の訓練を除いて，全ての訓練で検証損失が高止まりした．また，Vocab CL を用いたモデルについても，他のベースラインと同様に TEDS のような急激な検証損失の減少は確認されなかった (図 2b)．一方で，TEDS を用いたモデルでは，どの訓練においても，検証ロスが 2 万トレーニングステップ付近で安定的に減少した．これらの結果は，TEDS により，PLM の訓練安定性が通常の学習よりも大幅に向上することを示している．更に，RDS を用いたモデルの訓練安定性が向上していないことから，ランダム

3） https://huggingface.co/roberta-base



(a) TEDS, ADS and RDS　　(b) TEDS and vocab CL

図 2: 提案手法とベースラインの学習曲線の比較

表 1: 各モデルの GLUE ベンチマークスコアの比較．太字は各タスクの最高値を示す．QQP では F 値を，CoLA についてはマシューズ相関係数を，それ以外のタスクについてはアキュラシーをそれぞれ報告している．

Model	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST2	STSB	WNLI	Avg.
ADS	16.58	71.38	71.02	76.74	79.34	54.36	83.96	14.00	54.92	58.06
RDS	16.75	70.79	71.26	79.46	78.79	54.00	84.03	7.643	54.64	57.48
Vocab CL	17.98	70.66	70.24	74.47	78.39	54.21	83.38	10.64	**55.76**	57.41
TEDS w/ all data	**25.82**	**73.93**	**74.06**	**82.61**	**83.01**	55.80	**86.87**	**23.08**	55.48	**62.30**
TEDS w/o all data	23.88	73.45	72.49	82.32	82.87	**55.95**	86.19	21.62	54.07	61.43

にデータサイズを削減するだけでは不十分であり，トピックエントロピーの高い学習サンプルを選択することが重要であることがわかった．

4.2.2　汎化性能の評価

表 1 に各モデルの GLUE スコアの比較を示す．TEDS を用いたモデルは WNLI を除く全ての下流タスクにおいて最高値を達成し，全タスクの平均値において，ベースラインを 3 ポイント以上上回った．このことから，TEDS により安定した事前学習を行ったモデルは汎化性能も向上することが確認された．ベースラインでは，ADS が最も良いスコア (58.06) を達成し，RDS と Vocab CL をわずかに上回る結果となった．更に，TEDS では，全学習データを追加学習するモデル (TEDS w/ all data) の方が，全学習データを追加学習しないモデル (TEDS w/o all data) と比較して，7 つの下流タスクにおいて高いスコアを獲得し，平均スコアで約 0.9 ポイント上回った．訓練安定性の評価では，はじめから全データを用いることは初期の学習の安定性を損なうことが確認された．一方で，今回の結果から，トピックエントロピーが高い学習サンプルを選択することで初期の訓練を安定化させた場合，後半の訓練で全ての学習データを用いて追加学習することが汎化性能の向上にとって好ましいことが確かめられた．

5　まとめ

本稿では，事前学習言語モデルの訓練安定性を向上させる新しいデータ選択手法である TEDS を提案した．TEDS では，テキストに含まれるトピックの多様性指標であるトピックエントロピーに基づいて学習データを抽出し，2 段階の事前学習を行う．WikiText-103 を用いた実験の結果，TEDS を用いたモデルはベースラインよりも訓練安定性が大幅に高くなることがわかった．また，GLUE ベンチマークを用いた汎化性能の評価においても，TEDS に基づくモデルが平均スコアでベースラインを大きく上回ることが示された．今後の課題として，本手法をより大規模なコーパスを用いた訓練に応用することが期待される．

謝辞

本研究は，国立研究開発法人科学技術振興機構 (JST) による次世代研究者挑戦的研究プログラムの支援を受けて実施されたものです.

参考文献

[1] 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, pages 4171–4186, 2019.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. [arXiv:1907.11692].

[3] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations, 2020. [arXiv:1909.11942].

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter, 2019. [arXiv:1910.01108].

[5] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pages 5797–5808, 2019.

[6] Adrian de Wynter and Daniel J. Perry. Optimal subarchitecture extraction for bert, 2020. [arXiv:2010.10499].

[7] Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. Adaptive attention span in transformers. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pages 331–335, 2019.

[8] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020. [arXiv:2003.10555].

[9] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In **Advances in Neural Information Processing Systems**, volume 32, 2019.

[10] Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. Pre-training a BERT with curriculum learning by increasing block-size of input text. In **Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)**, pages 989–996, 2021.

[11] Ameeta Agrawal, Suresh Singh, Lauren Schneider, and Michael Samuels. On the role of corpus ordering in language modeling. In **Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing**, pages 142–154, 2021.

[12] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In **Proceedings of the 26th Annual International Conference on Machine Learning**, page 41 –48, 2009.

[13] Jae-Ho Lee and Yoichiro Hasebe. Readability measurement of japanese texts based on levelled corpora. **The Japanese Language from an Empirical Perspective**, pages 143–168, 2020.

[14] 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 **Advances in Neural Information Processing Systems**, volume 33, pages 1877–1901, 2020.

[15] 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**, 21(140):1–67, 2020.

[16] Conglong Li, Minjia Zhang, and Yuxiong He. Curriculum learning: A regularization method for efficient and stable billion-scale gpt model pre-training, 2021. [arXiv:2108.06084].

[17] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. [arXiv:1609.07843].

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

[19] 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**, pages 353–355, 2018.

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. [arXiv:1711.05101].