# 段落構造を利用した BERT に基づく事前学習

湯浅 亮也　谷 和樹　田村 晃裕　加藤 恒夫
同志社大学 理工学部
{cguc0095@mail4, cguc1070@mail4, aktamura@mail, tsukato@mail}.doshisha.ac.jp

## 概要

本研究では，文書の段落構造を利用した BERT に基づく事前学習手法を提案する．従来の BERT ベースの事前学習手法では，通常，文単位の連続性を学習する．また，文書における各文の位置づけは考慮しない．一方，Wikipedia 文書などの文書では，文は段落単位でまとめられ，各段落は見出しによって整理されている．そこで本研究では，段落単位のまとまりとつながり及び見出し情報を考慮した事前学習手法を提案する．livedoor ニュースコーパスを用いた文書分類タスクで評価した結果，従来の BERT と比較して，統計的に有意な性能改善（F 値のマクロ平均が+0.77(%)）を確認した．

## 1　はじめに

近年，NLP の分野では，大規模コーパスから単語や文などのテキストの汎用的な分散表現を事前に学習する事前学習手法が数多く提案されている．特に，BERT をベースとした事前学習手法を活用することで，様々な NLP タスクにおいて最高精度が更新されている．

従来の BERT ベースの事前学習では，通常，文単位でトークンのまとまりを捉え，文間の接続関係や文脈を考慮した単語や文の分散表現を事前学習する [1, 3, 4, 8]．一方，Wikipedia 文書を筆頭に，事前学習で使われる文書の中には，文が段落単位でまとめられ，各段落が見出しによって整理されている文書が多い．Wikipedia の記事「機械学習」の概略を図 1 に示す．図 1 の文書では，記事の内容が「1 概要」や「1.1 定義」といった見出しで整理されており，各内容は「段落 A」や「段落 B」のような段落単位のまとまりで記述されている．しかし，従来の BERT ベースの事前学習ではこのような段落単位の情報や見出し情報を活用できない．

そこで本研究では，段落単位のまとまりとつながり及び見出し情報を考慮する，文書の段落構造に基づく BERT による事前学習手法を提案する．具体的には，BERT への入力を段落単位にし，2 つの段落が連続するかを判定する事前学習を行うことで，段落単位のまとまりとつながりを反映した分散表現を学習する．また，段落が見出しに属するかを判定する事前学習を行うことで，各段落の見出しを反映した分散表現を学習する．段落同士の連続判定及び段落の見出しへの所属判定は，Sentence-BERT[5] の Classification Objective Function を用いて分類問題を解くことで実現する．

livedoor ニュースコーパスを用いた文書分類タスクで提案の事前学習手法を評価した結果，従来の BERT を用いた場合よりも F 値のマクロ平均が 0.77(%) 高くなり，統計的に有意な性能改善を確認した．

1 概要
　1.1 定義
　　【段落 A】
　　【段落 B】
　1.2 変数の種類
...
2 教師あり学習
　2.1 概要
　　2.1.1 訓練フェーズと汎化フェーズ
　　【段落 C】

**図 1**　Wikipedia 文書の例（機械学習）

## 2　従来手法

### 2.1　BERT

BERT[1] は Transformer Encoder[6] に基づくモデルであり，様々な NLP タスクにファインチューニング可能で汎用的な分散表現を獲得するための事前学習モデルである．大量のラベルなしテキストから事前学習された分散表現を用いて，目的のタスクに

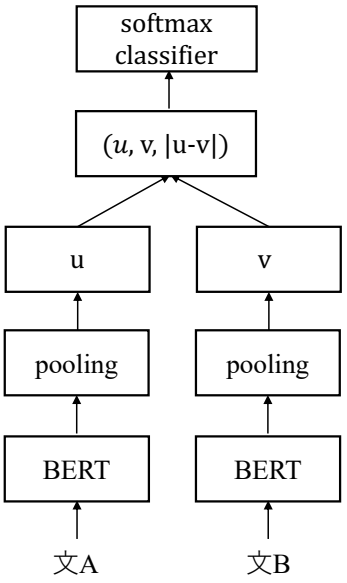


**図 2** Sentence-BERT (Classification Objective Function) モデルの概要図

ファインチューニングすることにより，学習時間の短縮や目的タスクの教師データが少量であっても高精度を達成することが可能である．

事前学習では，ラベルなしテキストデータを用いて，Next Sentence Prediction（NSP）と Masked Language Model（MLM）の 2 つの教師なし学習を行う．NSP では，同時に 2 つの文を入力し，入力された 2 文が連続するか否かの分類問題を解くことで，文の接続関係を考慮した分散表現を得る．MLM では，入力トークンをランダムにマスクし，マスク前のトークンを推定する問題を解くことで，文脈を考慮した分散表現を学習する．これらの事前学習では，入力の先頭に [CLS] トークン，各文の最後に [SEP] トークンを付与し，[MASK] トークンを用いてマスクした入力系列を使用する（例：[CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP] ）．NSP では [CLS] トークンの分散表現に基づき 2 値分類を行う．

この事前学習では，Wikipedia や BookCorpus [9] の文書を学習データとして用いている．これらの文書は，本来，段落単位でまとめられ，各段落が見出しによって整理されている．しかし，従来の事前学習では，見出し情報は陽に利用しておらず，また，文より大きい単位（段落など）のまとまりやつながりを捉えることが困難である．

## 2.2 Sentence-BERT

Sentence-BERT [5] は事前学習済み BERT を Siamese Network で用いてファインチューニングすることで，文の埋め込み表現を獲得する手法である．事前学習済み BERT モデルの出力側に文の分散表現を取り出すための pooling 層を加えたモデルを Siamese Network 内の共有モデルとして使う．先行研究では pooling 層の演算として，CLS，MEAN，MAX の 3 種類が提案されている．CLS は [CLS] トークンの分散表現，MEAN は全トークンの分散表現の平均ベクトル，MAX は全トークンの分散表現に対して各特徴量の最大のものを取る max-over-time 演算を行ったベクトルを文の分散表現とする．この Siamese Network モデルを文間の類似性に基づき学習する．使用する学習データに応じて，分類問題，回帰問題，Triplet Loss で学習する方法が提案されている．

後述の提案手法では，Classification Objective Function を用いた分類問題で学習する Sentence-BERT を利用する．モデルの概要を図 2 に示す．この Sentence-BERT は，2 文 A，B を BERT に個別に入力し，pooling 層で各文の分散表現を取り出す．その後，求めた 2 つの分散表現と要素ごとの差の絶対値ベクトルを結合し，結合したベクトルに基づき softmax classifier で 2 つの入力文が連続するか否かの 2 値分類を行う（式 (1)）．

$$softmax(W_t(\mathbf{u}, \mathbf{v}, |\mathbf{u} - \mathbf{v}|)) \tag{1}$$

ここで $\mathbf{u}$ と $\mathbf{v}$ は，それぞれ，文 A と B の分散表現であり，$W_t$ は学習される重み行列である．

## 2.3 その他の事前学習手法

BERT の派生モデルとして，XLNet [8] や RoBERTa [4]，ALBERT [3] などの多くの事前学習モデルが提案されている．XLNet の事前学習では，NSP を廃止し，[MASK] トークンは使用せずにトークンの予測順序を入れ替える permutation language modeling を学習している．RoBERTa の事前学習でも NSP が廃止されている．また，ALBERT の事前学習では，NSP の代わりに，文章の順序を予測する sentence-order prediction を行っている．

これらの従来モデルにおいても，段落単位での連続性は学習されておらず，また，見出し情報が活用されていない．

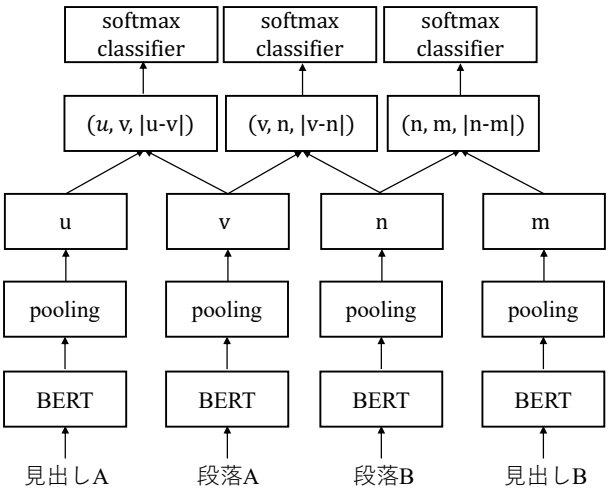


**図 3**　段落構造を利用した事前学習モデルの概要図

## 3　段落構造を利用した事前学習

本節では文書の段落構造を利用した BERT による事前学習手法を提案する．事前学習で用いる文書の多くは，文が段落単位でまとめられ，各段落が見出しによって整理されている．そこで提案手法では，段落単位のまとまりとつながり及び見出し情報を反映した分散表現を事前学習する．具体的には，BERT への入力を段落単位にし，2 つの段落が連続するか否かを推定する連続判定と，段落が見出しに所属するか否かを推定する所属判定を行う．

提案手法の概要を図 3 に示す．モデル構造は 2.2 節で説明した Classification Objective Function を用いた Sentence-BERT を利用する．図 3 の全ての BERT の重みは共有されており，pooling 層は MEAN を用いる．そして，段落同士の連続判定及び段落の見出しへの所属判定は Classification Objective Function を用いて 2 値分類問題として解く．

段落同士の連続判定は，図 3 の中央の softmax classifier で行う．2 つの段落 A と B に対して，BERT と pooling 層で各段落の分散表現を抽出する．その後，式 (1) を用いて，段落 A の次に段落 B が繋がるか否かの 2 値分類を行う．式 (1) における **u** と **v** は，それぞれ，前方段落（段落 A）と後方段落（段落 B）の分散表現となる．この事前学習では，本来の文書内で連続する 2 つの段落の後方段落を，50%の確率で，異なる文書から無作為に取り出した段落に置き換えたデータを学習データとする．

段落同士の連続性を 2.1 節で説明した BERT の NSP ではなく Sentence-BERT の構造で学習した理由は，各段落を個別に BERT の入力にすることで，[CLS] トークンを段落同士の接続関係を表すベクトルではなく，各段落の分散表現として事前学習するためである．

段落の見出しへの所属判定は，図 3 の左右の softmax classifier で行う．2 つの段落 A，B と 2 つの見出し A，B に対して，BERT と pooling 層で分散表現を抽出する．その後，式 (1) を用いて，段落 A が見出し A に，段落 B が見出し B に所属するか否かをそれぞれ 2 値分類する．式 (1) における **u** と **v** は，それぞれ，見出し A または B，段落 A または B の分散表現となる．この事前学習では，本来の文書内に存在する見出しと段落の対に対して，見出しを，50%の確率で，異なる文書にある異なる見出しに置き換えたデータを学習データとする．

ここで，Wikipedia 文書の見出しは予め決められておらず，文書作成者が任意の名前を付けられることに注意されたい．任意の見出しが出現する可能性があるため，提案手法では，予めクラスを仮定した多クラス分類ではなく，段落が見出しに属するか否かの 2 クラス分類で見出し情報を取り込む．

## 4　実験

従来の BERT モデルと提案手法により事前学習された BERT モデルを，それぞれ，livedoor ニュースコーパス[1)]を用いた文書分類タスクでファインチューニングし，その文書分類性能を比較することにより，提案の事前学習手法の有効性を検証する．

### 4.1　実験設定

**事前学習**　本実験では，段落間の連続判定と段落の見出しへの所属判定の両方を導入した提案の事前学習モデルに加えて，段落間の連続性判定のみ導入した事前学習モデル，段落の見出しへの所属判定のみを導入した事前学習モデル，従来の BERT モデルを事前学習モデルとして用いた場合の性能を評価する．

従来の BERT モデルは，東北大学 BERT の bert-base-japanese-whole-word-masking[2)]を使用した．その他の提案手法の事前学習モデルは，従来の BERT モデルを初期値として，Wikipedia の jawiki-latest-pages-articles.xml.bz2[3)]を用いて 3 節で提案した方法で学習した．見出しへの所属判定で用いる見出し

1）https://www.rondhuit.com/download.html#ldcc
2）https://huggingface.co/cl-tohoku
3）https://dumps.wikimedia.org/jawiki/latest/

**表 1** livedoor ニュースコーパス

| クラス（記事サイト） | 文書数（記事数） |
|---|---|
| MOVIE ENTER | 870 |
| 家電チャンネル | 864 |
| 独女通信 | 870 |
| livedoor HOMME | 511 |
| エスマックス | 870 |
| Peachy | 842 |
| IT ライフハック | 870 |
| Sports Watch | 900 |
| トピックニュース | 770 |
| 合計 | 7,367 |

は，段落が属する最上位の見出しを用いた．例えば，図 1 においては，段落 A と段落 B の見出しは「概要」，段落 C の見出しは「教師あり学習」とした．sentence-transformers[4] のライブラリを用いて実装し，BERT への最大入力長は 512 とした．その他のパラメータは，Sentence-BERT[5] の実験設定に倣い，エポック数を 1，学習率を 2e-5 とし，最適化手法には Adam [2] を用いて，学習データ全体の 10%に対して warm-up を行った．

**ファインチューニング**　本実験の評価は，livedoor ニュースから収集されたニュースコーパスを用いた多クラス文書分類で行う．このデータセットは 7,367 件の記事で構成されている．従来研究 [10] に倣い，各記事の収集元である記事サイトをクラスとする．クラス数は 9 である．表 1 に各クラスの記事数を示す．また，各記事は URL，日付，タイトル，本文で構成されているが，本研究では，タイトルと本文を結合したテキストからクラスを分類する．

文書分類モデルは，事前学習した BERT の出力側に，[CLS] トークンの分散表現に基づき softmax classifier でクラスを決める出力層を加えたモデルである．この文書分類モデルを，多クラス文書分類タスクでファインチューニングする．なお，提案手法の事前学習モデルにおいては，Sentence-BERT の構成要素のうち，事前学習した BERT 部分を使う．simple-transformers[5] のライブラリを用いて実装した．最大入力長さは 256 とし，エポック数，学習率，バッチサイズは，それぞれ，100，2e-5，32 とした．その他のパラメータは，simple-transformers ライブラリのデフォルトの値を用いた．

4） https://github.com/UKPLab/sentence-transformers
5） https://github.com/ThilinaRajapakse/simpletransformers

**表 2** 実験結果

| モデル | F 値 [%] |
|---|---|
| 従来の BERT | 93.04 |
| 提案手法（連続判定のみ） | 93.20 |
| 提案手法（所属判定のみ） | 93.60 |
| 提案手法（連続判定＋所属判定） | 93.81 |

### 4.2 実験結果

評価は 5 分割交差検証で行った．交差検証では，1 fold をテストデータとし，残りの 4 fold の内，80%を学習データ，20%を開発データとした．評価指標には F 値のマクロ平均を用いた．

表 2 に実験結果を示す．表 2 より，従来の BERT モデルと比較して，提案の 3 つの事前学習モデルは，いずれも高い分類性能を達成できることが確認できた．また，3 つの提案のモデルの中で，段落同士の連続判定と見出しへの所属判定の両方を導入したモデルが最も高い性能を達成した．最高性能を達成した提案モデルと従来の BERT モデルを用いた場合の文書分類結果の差に対して符号検定を行った結果，有意水準 1%で有意であった．このことから，段落同士の連続性や見出し情報を考慮して分散表現を事前学習することは，文書分類タスクで有効であることが確認できた．また，段落の連続性の情報と見出し情報のそれぞれを考慮するよりも，両方を考慮した方が結果が良いことから，段落同士のまとまりと見出し情報の両方が，文書の特徴を求める際の重要な手がかりになることが実験的に確認できた．

## 5 おわりに

本研究では，事前学習に用いる文書は段落でまとめられており，各段落は見出しによって整理されていることに着目し，これら 2 つの特徴を利用した BERT に基づく事前学習手法を提案した．提案手法では，BERT に段落単位で入力し，2 つの段落の連続判定と段落の見出しへの所属判定による事前学習を行うことで，段落情報をテキストの分散表現の事前学習に取り入れた．提案の事前学習手法を，livedoor ニュースコーパスを用いた文書分類タスクで評価した結果，従来の BERT と比較して，統計的に有意に高い性能が実現できることを確認した．

今後は，NLP のベンチマーク評価である GLEU [7] など，文書分類タスク以外で，提案の事前学習モデルの有効性を確認したい．

## 謝辞

本研究の一部はトランスコスモス株式会社との共同研究により得られたものである．本研究を支援してくださった，トランスコスモス株式会社の伊藤和真氏，及川秀俊氏，濱田充男氏，大林弘明氏に感謝申し上げます．また，本研究を進めるにあたり貴重な助言をしてくださった，愛媛大学の二宮崇先生，梶原智之先生，秋山和輝氏に感謝を申し上げます．

## 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 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

[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. **arXiv preprint arXiv:1909.11942**, 2019.

[4] 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. **arXiv preprint arXiv:1907.11692**, 2019.

[5] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems 30**, pp. 5998–6008. 2017.

[7] 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, 2018.

[8] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. **Advances in neural information processing systems**, Vol. 32, , 2019.

[9] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In **Proceedings of the IEEE international conference on computer vision**, pp. 19–27, 2015.

[10] 菊田尚樹, 新納浩幸. Bert を用いた文書分類タスクへの mix-up 手法の適用. 第 27 回言語処理学会年次大会, pp. 599–602, 2021.