DEIM2022 F43-4

文化財の特徴理解に特化したBERTモデル

三林　亮太[†]　上田　昌輝[†]　川原　敬史[†]　松本　直彰[†]　吉村　拓真[†]

相原　健郎[††,†††]　神門　典子[†††,††††]　荘司　慶行[†††††]　中島　悠太[††††††]

山本　岳洋[†††††††,†††]

山本　祐輔[††††††††]　大島　裕明[†,†††††††,†††]

† 兵庫県立大学 応用情報科学研究科　〒 650–0047 兵庫県神戸市中央区港島南町 7–1–28

†† 東京都立大学 都市環境学部　〒 192–0397 東京都八王子市南大沢 1–1

††† 国立情報学研究所　〒 100–0003 東京都千代田区一ツ橋 2–1–2

†††† 総合研究大学院大学　〒 100–0003 東京都千代田区一ツ橋 2–1–2

††††† 青山学院大学 理工学部　〒 252–5258 神奈川県相模原市中央区淵野辺 5–10–1

†††††† 大阪大学 データビリティフロンティア機構　〒 565–0871 大阪府吹田市山田丘 2–8

††††††† 静岡大学 情報学部　〒 432–8011 静岡県浜松市中区城北 3–5–1

†††††††† 兵庫県立大学 情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1

E-mail: †{aa20r511,aa20t502,aa20m503,aa20y510,aa20w514,ohshima}@ai.u-hyogo.ac.jp,
††kenro.aihara@tmu.ac.jp, †††kando@nii.ac.jp, ††††shoji@it.aoyama.ac.jp, †††††n-yuta@ids.osaka-u.ac.jp,
††††††t.yamamoto@sis.u-hyogo.ac.jp, †††††††yusuke_yamamoto@acm.org

あらまし　本論文では，文化財の特徴理解に特化したBERTモデルを提案する．BERTは，大規模なコーパスで事前学習したモデルを，特定のタスクでファインチューニングすることで，様々なタスクに適応できる汎用言語モデルである．近年，ファインチューニングをおこなう前に，解きたいタスクのドメインで追加の学習をすることで，タスクにおける性能が向上することが知られている．そこで，本研究では，国立民族学博物館の文書データを用いてBERTモデルを追加学習することで，文化財の特徴理解に特化したBERTモデル（みんぱくBERT）を提案する．みんぱくBERTの有効性を検証するために，みんぱくBERTと事前学習済みBERTを文化財に関するタスクでファインチューニングし，性能を比較した．比較するタスクには，国立民族学博物館データに付与されている，用途を表すOCM（Outline of Cultural Materials）ラベルと地域を表すOWC（Outline of World Cultures）ラベルを用いた分類問題を設定した．結果として，OCMラベルを用いた用途分類タスクとOWCラベルを用いた地域分類タスクにおいて，みんぱくBERTの性能が，事前学習済みBERTモデルより高くなることがわかった．さらに，タスクのファインチューニングにおける収束速度も，みんぱくBERTが速い傾向にあった．本研究で作成したみんぱくBERTは，モデル共有サイトに公開し，誰でも使える言語資源として利用可能となった．

キーワード　BERT，博物館，みんぱく，追加学習

1　はじめに

近年，汎用言語モデルであるBERTを特定のタスクでファインチューニングすることで，自然言語処理におけるタスクがいくつか解けることがわかっている．ファインチューニングとは，事前学習済みのモデルに対して，特定のタスクに特化した学習をおこなうことである．たとえば，入力された文書をカテゴリごとに分類する文書分類タスクや，入力された文の品詞を推定する系列ラベリングタスクなどが挙げられる．

ファインチューニングをおこなう前に，解きたいタスクのドメインで追加の学習をすることで，モデルの性能が向上することが知られている [5], [13]．たとえば，Med-BERT [14] では，医療に関するコーパスで追加学習をおこない，疾患の予測タスクについてファインチューニングした結果，事前学習モデルより，予測性能が向上していることを示した．また，LEGAL-BERT [3] においても，法律に関するコーパスで追加学習をおこなうことで，タスクの性能が向上するほか，学習の収束速度の向上やロス値が小さくなる傾向を示した．以上のように，特定のドメインで追加学習をおこなうことで，タスクに対するモデルの性能が向上することが知られている．

現在，文化財のドメインを対象とした追加学習済みモデルは，筆者らの知る限り提案されていない．そこで，本研究では，国立民族学博物館 [1] の文化財に関する文書データを用いて，文化

1：https://www.minpaku.ac.jp/

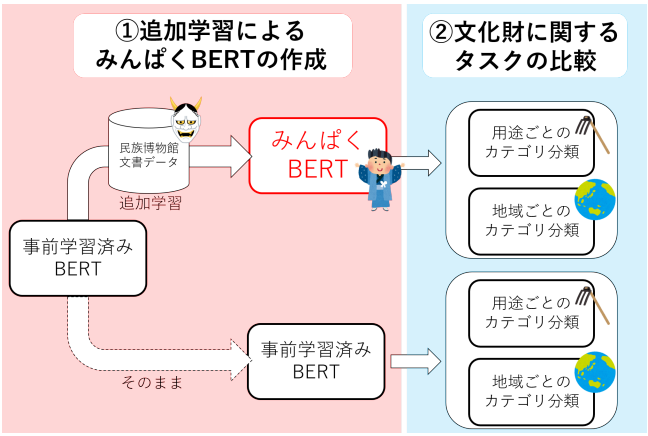


図 1　研究の全体像

財の特徴理解に特化した BERT モデル（みんぱく BERT）を提案する．本研究は，図 1 に示すように，日本語の事前学習済み BERT に対して，国立民族学博物館の文書データを用いた追加学習をおこなうことで，みんぱく BERT を作成する．作成したみんぱく BERT と事前学習済み BERT を，文化財に特化したタスクにおいて，ファインチューニングすることで，みんぱく BERT と事前学習済み BERT の性能を比較する．

本論文では，まず，2 節で追加学習と国立民族学博物館に関する関連研究について述べる．3 節では，モデルの学習に使用する文化財のデータについて述べる．4 節と 5 節では，みんぱく BERT の追加学習方法と，文化財に関するタスクとファインチューニング方法について述べる．6 節で，実験の詳細について述べる．最後に，7 節でまとめと今後の課題について述べる．

2　関 連 研 究

本研究は，BERT における追加学習と国立民族学博物館に関連が深い．よって，本節では，BERT における追加学習と国立民族学博物館に関連する研究を取り上げる．

2.1　BERT

BERT [4] は Devlin らによって提案された，自然言語処理で用いられる汎用言語モデルである．BERT は，Transformer [17] の Encoder のみを使用したモデルであり，事前学習をおこなってから使用することが一般的な使用方法である．事前学習とは，BERT を一度，大規模なコーパスで学習することで，汎用的な言語知識を獲得する学習である．事前学習では，語の予測をおこなう MaskedLM と文の隣接関係を予測する NSP（Next Sentence Prediction）を同時に学習する．しかし，NSP は性能向上にあまり影響しておらず，近年，MakedLM だけを学習した，RoBERTa [11] などの改良モデルが提案されている．このように，事前学習で得られたパラメータをベースに，ファインチューニングをおこなうことが BERT の一般的な使用方法である．

2.2　追 加 学 習

追加学習とは，事前学習済みモデルに対して，解きたいタスクに関連するドメインのコーパスで追加の学習をおこなうことである [2]．追加学習は，Suchin らによって有効性が示されている [5]．Suchin らは，一般的なドメインで学習した事前学習済みモデルに対して，解きたいタスクと同様のドメインで追加の学習をおこなうことで，性能が向上することを示している．また，Han らは，固有表現抽出タスクにおいて，追加学習でよい性能を示している [6]．

追加学習済み BERT モデルはいくつか提案されており，その追加学習のドメインは多岐に渡る．たとえば，beltagy らは化学に関するドメインに特化した追加学習モデルである，SciBERT [1] を提案している．また，LEGAL-BERT [3] では，法律関係のコーパスで追加学習をおこなっている．

また，医療に関する追加学習モデルも提案されており，Lee らは医学論文を用いて追加学習をおこなった BioBERT [10] を提案している．関連したドメインとして，Laila らは，MedBERT [14] と呼ばれる，医療に関するドメインに特化した追加学習モデルを提案している．加えて，ClinicalBERT [8] では，匿名の医療診断書情報を用いて追加学習をおこなっている．さらに，Yao らは生物医学とコンピュータ科学を対象に追加学習をおこなった [19]．

また，インターネットにおけるソーシャルネットワークサービスのドメインに特化した追加学習モデルも提案されている．Tuhin らは，ネット上の意見に関するドメインに特化した追加学習モデルである，IMHO [2] を提案している．関連して，Dat らは，Twitter のコーパスを用いて追加学習した BERTweet [12] を提案している．本モデルは RoBERTa [11] を用いて学習している．また，壹岐ら [21] は事前学習モデルに対して，固有のドメインを持つコーパスを学習させることで，事前学習済みモデルとの MASK の Loss 値の比較をおこなっている．以上のように，多用なドメインのコーパスにおいて，追加学習がおこなわれている．

近年，未知語の追加による，性能の向上も報告されている [7] [15]．未知語とは，モデルで扱わない出現頻度の低い語のことである．一般的に，未知語は専門的なドメインのコーパスにおいて多く存在する．未知語はタスクにおいて重要な特徴であるため，それらの理由から未知語を追加した状態での追加学習が望まれる．

2.3　国立民族学博物館

Wang ら [18] は国立民族学博物館のデータを用いて，データのカテゴリを推定するタスクに取り組んでいる．国立民族学博物館のプロジェクトでは，電子ガイドをベースに，学習支援や鑑賞体験の提案をおこなっている [20] [22]．

3　みんぱくデータ

みんぱくデータとは，国立民族学博物館のデータベース [3] に

2：事前学習の際にドメインに特化したコーパスで学習することを，追加学習と呼ぶ場合もあるが，本研究の追加学習とは異なる．

3：https://htq.minpaku.ac.jp/menu/database.html

標本番号	H0131938
記入情報	記入日：2021-11-11／記入責任者：八杉 佳穂
標本名	舞踏用 仮面（山羊）（ブトウヨウ カメン（ヤギ））
収集日	1985-10-12
OWC	NU24
OCM	532, 535
使用民族	Mixteco;Mixtec
用途・使用法	デホロネスの踊りに使用
製作地	メキシコ合衆国 Oaxaca州 Pinotepa de Don Luis

図 2　みんぱくデータの例

表 1　詳細に収録されている文化財の統計量

文化財の数	73,266
解説文の数	176,943
解説文のユニーク数	49,199
文長平均	59

ある標本資料（文化財）に関する情報を集めたデータである．文化財には，図 2 に示すように,「標本名，OWC ラベル，OCM ラベル，用途・使用法」などの情報が付与されている．

3.1　文　化　財

文化財のデータには「目録」と「詳細」が存在する．目録には，国立民族学博物館が所有する，ほぼすべての文化財 286,221 件についての基本的な情報が収録されている．基本的な情報とは,「標本名，地域，民族」などの情報が含まれる．詳細には，文化財 73,226 件に対して，基本的な情報に加え，用途使用法などの解説文の情報が収録されている．本研究では，追加学習のために，テキストデータが必要であるため，解説文が収録されている詳細のデータを使用する．

詳細の文化財に登録されている解説文の統計量を表 1 に示す．詳細に収録されている文化財の数は 73,266 件であり，すべての解説文の数は 176,943 件である．文化財に付与されている解説文は重複があるため，ユニークを取ると，49,199 件となった．解説文の文長は図 3 に示すように，ほとんどの解説文が短文で構成されており，解説文の文長の平均は 59 文字であった．

3.2　OCM ラベル

OCM（Outline of Cultural Materials）ラベルは，文化財の機能や用途を表したラベルである．たとえば,「音楽」や「印刷」という機能や,「ダンス」や「釣り」といった用途が表されている．OCM ラベルは全部で 735 種類あり，内 1 種類はどれにも属さない「該当なし」のラベルが含まれる．

OCM ラベルは 3 桁の数字で表されており，カテゴリとサブカテゴリを表している．最初の 2 桁はカテゴリを表しており，たとえば,「190」から「198」は「言語」のカテゴリを表す．最後の 1 桁はサブカテゴリを表しており，たとえば,「192」は「ボキャブラリー」,「193」は「文法」のように「言語」というカテゴリをより詳細なサブカテゴリで表している．OCM ラベルはひとつの文化財に複数付与されている場合もある．

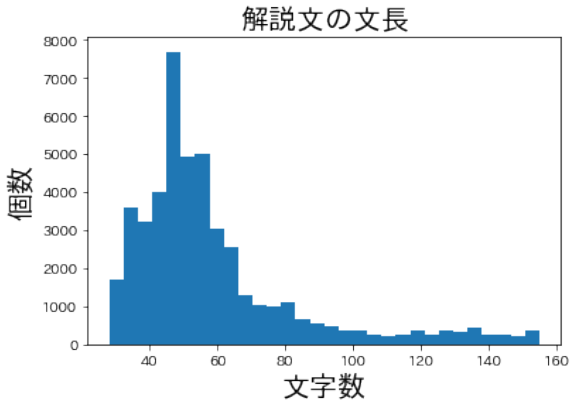


図 3　解説文の統計量

3.3　OWC ラベル

OWC（Outline of World Cultures）ラベルは，文化財の地域を表したラベルである．たとえば,「アジア」や「日本」といったカテゴリや「大阪」や「兵庫」といったサブカテゴリが存在する．これらのラベルには上位概念が設定されており，たとえば,「大阪」のラベルには,「東アジア」という上位概念が設定されており,「イタリア」のラベルには,「南ヨーロッパ」という上位概念が設定されている．これらのラベルが文化財に対して付与されており，たとえば，ある「日本刀」の文化財には,「京都」というラベルがついている．このラベルにより，文化財がどの地域に属するかを判別することができる．

4　みんぱく BERT の追加学習

本節では，みんぱく BERT の追加学習方法について述べる．

4.1　追 加 学 習

みんぱく BERT の追加学習には，日本語事前学習済み BERT をベースに，みんぱくデータを用いて学習する．事前学習済み BERT には，東北大学乾研究室が提供している日本語事前学習済み BERT [4]を用いる．この BERT は，日本語 Wikipedia のデータを用いて事前学習されたモデルである．この事前学習済み BERT に対して，追加学習をおこなう．

追加学習には，MaskedLM タスクをおこなう．BERT の事前学習には MaskedLM タスクと NSP（Next Sentence Prediction）の 2 種類のタスクを同時に解くマルチタスク学習が採用されているが，NSP は事前学習における性能の向上にあまり影響しないことが知られているため，本研究では MaskedLM タスクのみをおこなう [11].

追加学習方法は，まず，図 4 に示すように，入力となる解説文をトークナイザで分割する（トークン化）．使用するトークナイザは，事前学習済み BERT モデルと同じものを用いた [5]．トークン化した語を，事前学習と同様に，ランダムに置き換える処理をおこなう．この時，置き換える方法は 2 つあり，ある

4：https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

5：https://github.com/cl-tohoku/bert-japanese

トークンを別の語に置き換える処理と，あるトークンを特殊トークンである [MASK] に置き換える処理のいずれかをおこなう．置き換えたトークンのみを対象に，そのトークンの元のトークンを予測する．この予測したトークンと元のトークンのロスを計算することで，BERT を追加学習する．

4.2 未知語の追加

本研究では，追加学習の前に，未知語の追加をおこなう．未知語とは，モデルでは扱わない，出現頻度の低い語のことである．たとえば，東北大学の事前学習済み BERT モデルでは，日本語 Wikipedia のコーパスから，BERT で取り扱う語彙のサイズを 32,000 に設定している．よって，特殊なドメインのコーパスに含まれるような，専門的な語彙は未知語と判別される．BERT において未知語は [UNK] という特殊なトークンに置き換えられ，すべての未知語は，同じ [UNK] トークンとして扱われる．しかし，このような専門的な語はドメインにおいて重要な語である可能性があり，タスクに必要な特徴量であるといえる．よって，本研究では，事前学習済み BERT モデルに対して，語彙の追加をおこなうことで，未知語を考慮する．

未知語の追加は，以下の 2 つのステップでおこなう．

（1） みんぱくデータ内の未知語の検知

（2） BERT とトークナイザへの未知語の追加

まず，みんぱくデータから，未知語を検知する．未知語の検知には，事前学習済み BERT モデルで使用したトークナイザ [6] を用いる．みんぱくデータすべてに対して，トークナイザを用いて，トークン化をおこなう．その際，未知語と判定されたトークンをすべて取得する．

次に，未知語と判定されたトークンを，トークナイザの辞書と BERT モデルの embeddings 層に追加する．図 5 に示すように，トークナイザの 32,000 語の語彙に対して，未知語を追加する．これにより，トークン化の際に，未知語と認識されていた語彙が正しく認識される．BERT モデルに対しても，未知語の追加処理をおこなう．事前学習済み BERT モデルの embeddings 層に対して，未知語の数だけ embeddings 層の次元を追加する．事前学習済み BERT モデルの embeddings 層は，32,000 × 768 次元のマトリックスであるが，これを（32,000+追加する未知語の数）× 768 次元のマトリックスに拡張する．この時，追加した次元の要素は，乱数によって初期化される．以上の手順により，未知語を追加したモデルを作成した．

5 文化財に関するタスクとファインチューニング

本節では，みんぱく BERT と事前学習済み BERT の比較のためにおこなう，文化財に関するタスクとファインチューニングについて解説する．本研究でおこなうタスクは，2 つであり，文化財の用途を分類する OCM ラベル分類タスクと文化財の地域を分類する OWC ラベル分類タスクをおこなう．

6：https://github.com/cl-tohoku/bert-japanese

5.1 OCM ラベルを用いた文化財の用途分類タスク

OCM ラベル分類は 3.2 節で述べた，文化財に対して付与されている用途を表したラベルである，OCM ラベルを分類するタスクである．本タスクでは，文化財の解説文を入力に，OCM ラベルを出力するファインチューニングをおこなう．本タスクの問題定義は以下の通りである．

- **入力：解説文**
- **出力：OCM ラベル**

今回，OCM ラベルはカテゴリを表す上位 2 桁のみを使用する．OCM ラベルと解説文のペアデータはすべてで，14,652 件であり，これを 8:1:1 に分割し，訓練データ 11,721 件，検証データ 1,465 件，テストデータ 1,466 件とした．

OCM ラベル分類は，BERT を用いたマルチラベル分類として解く．手法の全体像は，図 6 に示すように，解説文を入力に，ラベルを推定する．まず，解説文をトークナイザを用いて，トークンに分割する．分割したトークンに対して，前後に特殊トークンである，[CLS] と [SEP] トークンを付与し，BERT に入力する．次に，BERT の出力から，[CLS] トークンに対応するベクトルを得る．得たベクトルを FC 層に入力し，sigmoid 関数をかけて予測値を得る．予測値と正解ラベルを BCE（Binary Cross Entropy Loss）にてロス計算をおこない，誤差逆伝播をおこなう．推論時は，予測値が 0.5 以上のものは 1 とし，0.5 未満は 0 として扱った．

5.2 OWC ラベルを用いた文化財の地域分類タスク

本タスクは，OWC ラベルの地域情報を文化財の解説文から分類するタスクである．本タスクでは，OWC ラベルの上位概念を使用する．OWC ラベルにおける上位概念とは，たとえば，「大阪」のラベルには,「東アジア」という上位概念が設定されており,「イタリア」のラベルには,「南ヨーロッパ」という上位概念が設定されている．このような上位概念を分類するタスクが，文化財の地域分類タスクである．本タスクの問題定義は以下の通りである．

- **入力：解説文**
- **出力：OWC ラベル**

具体的には，文化財の解説文を入力に，文化財の上位概念を推定する分類問題をファインチューニングによって解く．データはすべてで，15,320 件であり，これを 8:1:1 に分割し，訓練データ 12,256 件，検証データ 1,532 件，テストデータ 1,532 件とした．

OWC ラベル分類は，BERT を用いたマルチクラス分類として解く．手法の全体像は，図 6 に示すように，解説文を入力に，ラベルを推定する．まず，解説文をトークナイザを用いて，トークンに分割する．分割したトークンに対して，前後に特殊トークンである，[CLS] と [SEP] トークンを付与し，BERT に入力する．次に，BERT の出力から，[CLS] トークンに対応するベクトルを得る．得たベクトルを FC 層に入力し，softmax 関数をかけて予測値を得る．予測値と正解ラベルを CrossEntropyLoss にてロス計算をおこない，誤差逆伝播をおこなう．推論時は，予測値の中で値が最大のものを 1 とし，

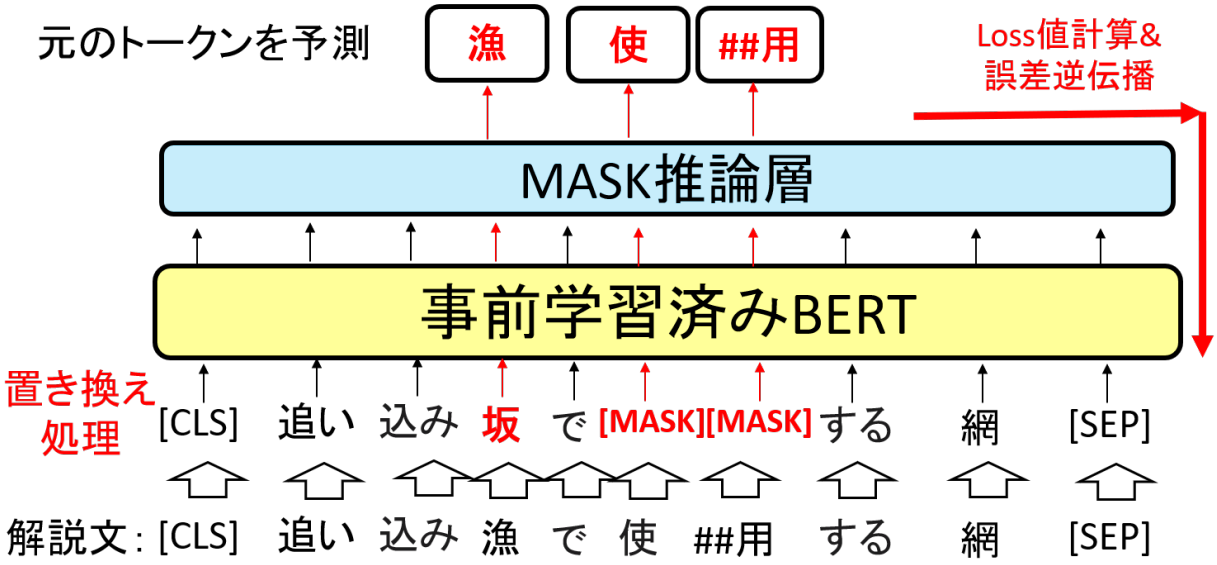


図 4　追加学習の概要図

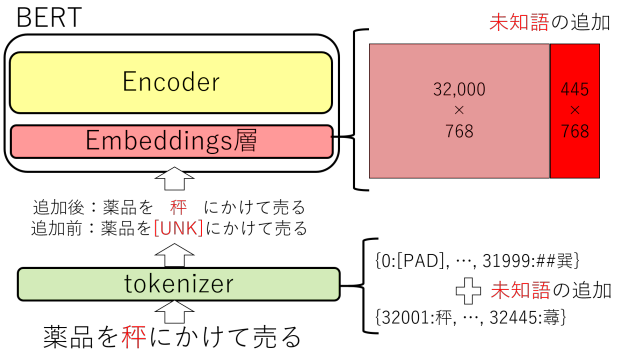


図 5　未知語追加の図

他は 0 として扱った．

6　実　　験

本節では，みんぱく BERT の追加学習の実験条件と，2 つのタスクのファインチューニング方法の実験条件について述べる．

6.1　追 加 学 習

追加学習には，みんぱくデータ内の文化財に付与されている解説文を学習データとして使用した．対象とする解説文は，文化財の「機能・用途」を示すタグに付与されている解説文である．解説文は全部で 58,526 件あり，ユニークをとると，13,366 件であった．本実験では，13,366 件の学習データを追加学習の訓練データとして使用した．

訓練データを用いて，事前学習済みの BERT モデルに対して追加学習をおこなう．事前学習済みモデルには，東北大学が公開している BERT モデル [7]を使用した．東北大学が公開している BERT モデルは，日本語 Wikipedia データを用いて事前学習された BERT モデルである．トークナイザは Mecab+IPA 辞書+WordPiece [16] を使用している．

追加学習は，MaskedLM タスクのみおこなった．MaskedLM タスクは，図 4 に示すように，入力文書に対して，一部の語を MASK トークンや別の語に置き換え，周辺の文脈から，その語を推定するタスクである．今回，以下の設定で MaskedLM タスクをおこなった．

- **80%：MASK トークンに置換**
- **10%：別の語に置換**
- **10%：置換しない**

BERT モデルの追加学習に用いたハイパーパラメータは以下のように設定した．

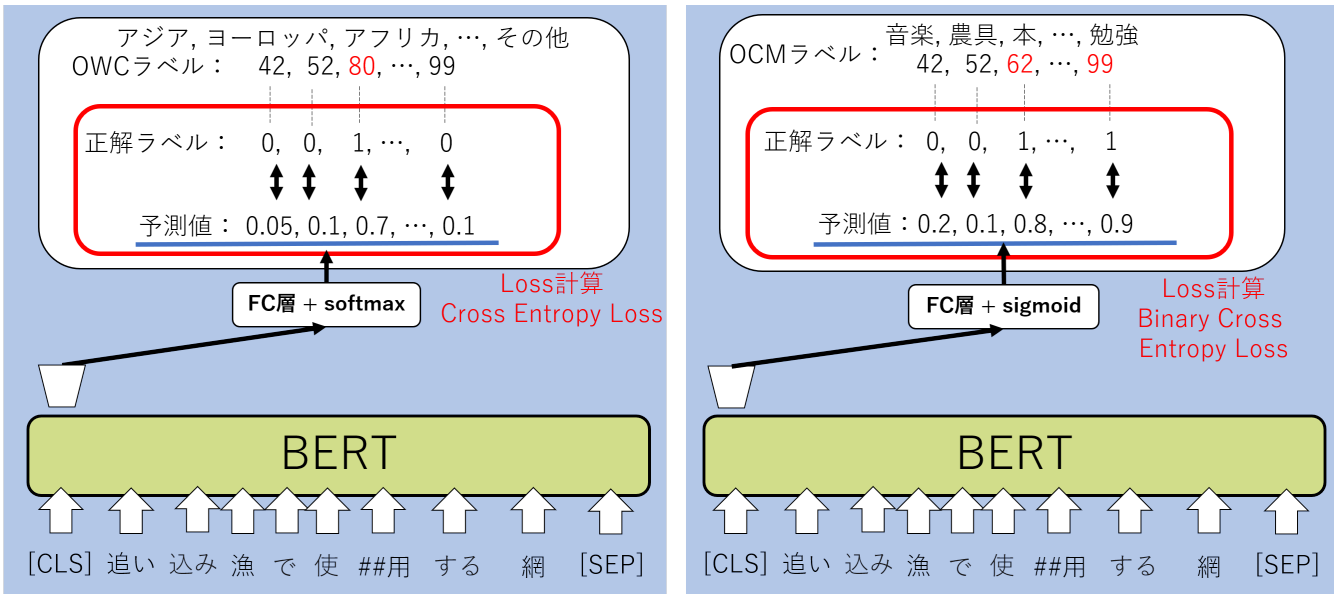
- **batch size：16**
- **optimizer：Adam [9]**
- **ロス関数：Cross-entropy Loss**
- **学習率：2e-5**
- **Dropout：0.1**
- **max length ： 512**

学習図を，図 7 と図 8 に示す．学習開始時点では，MaskedLM の正解率は 0.55 と低いが，学習が進むにつれて上昇している．最終的に，追加学習は 100epoch おこない，正解率は 0.95 程度になった．

6.2　未知語の追加

未知語の追加は，事前学習モデルと同じ東北大学のトークナイザを用いておこなった．まず，訓練データに対して，トークナイザを用いて，トークン化をおこなった．この時，未知語として出力されたトークン 445 語を，トークナイザの語彙に追加し，BERT の embeddings 層の重み行列を新たに 445 次元追加した．追加には huggingface の「resize_token_embeddings」を用いて，モデルの embeddings 層に追加した．

7：https://github.com/cl-tohoku/bert-japanese



（a）OWCラベル分類タスク　　（b）OCMラベル分類タスク

図 6　OCM ラベル分類タスクの学習と OWC ラベル分類タスクのマルチラベル学習

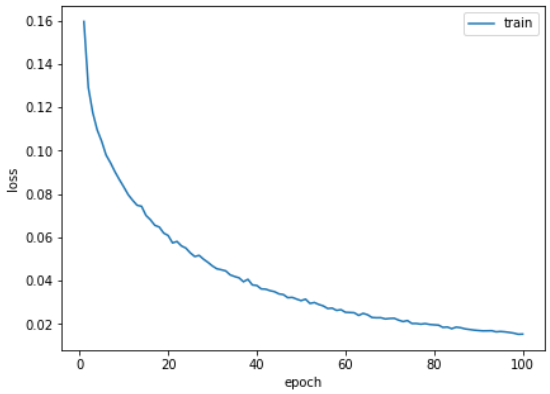


図 7　追加学習の学習曲線：Loss

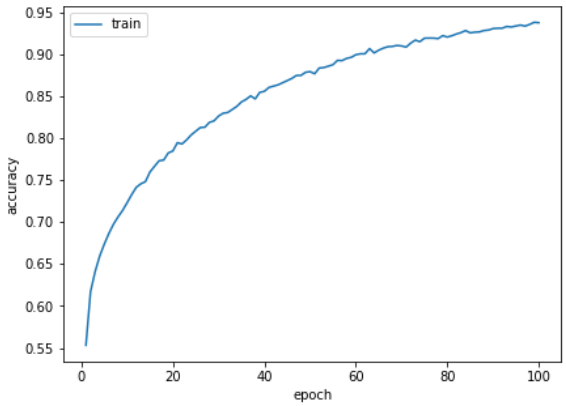


図 8　追加学習の学習曲線：Accuracy

6.3　ファインチューニング

ファインチューニングをおこなうタスクでは，みんぱく BERT と事前学習済み BERT の学習をおこなう．ファインチューニングでは，みんぱくデータ内のラベルを用いて，2 つのタスクに取り組んだ．本節では，それぞれのタスクの実験設定について述べる．

6.3.1　OCM ラベルを用いた展示物の用途推定タスク

OCM ラベルを用いた展示物の用途推定タスクでは，展示物データから OCM ラベルを含む文書のみを取得し，11,720 件のデータを用いてファインチューニングをおこなった．OCM ラベルと解説文のペアデータはすべてで，14,652 件であり，これを 8:1:1 に分割し，訓練データ 11,721 件，検証データ 1,465 件，テストデータ 1,466 件とした．

OCM ラベルは 3 桁の数字で表現されており，上位 2 桁がカテゴリを表し，下位 1 桁がサブカテゴリを表す．本研究では，サブカテゴリの分類はおこなわず，OCM ラベルはカテゴリを表す上位 2 桁のみを使用する．

OCM ラベルは複数付与されている場合がある．そのため，ラベルを同時に複数予測する必要がある．本研究では，OCM ラベルをマルチラベルとして設定し，同時に予測をおこなった．たとえば，図 6 に示すように，62 と 99 がラベルとして付与されている場合は，対応するインデックスに 1 が立つ，マルチホットなベクトルである．みんぱく BERT の学習は 5epoch で終了し，事前学習済み BERT の学習は 7epoch で終了した．ハイパーパラメータは以下の通りである．

- **batch size：32**
- **optimizer：Adam [9]**
- **ロス関数：Binary Cross Entropy Loss**
- **学習率：2e-5**
- **Dropout：0.1**

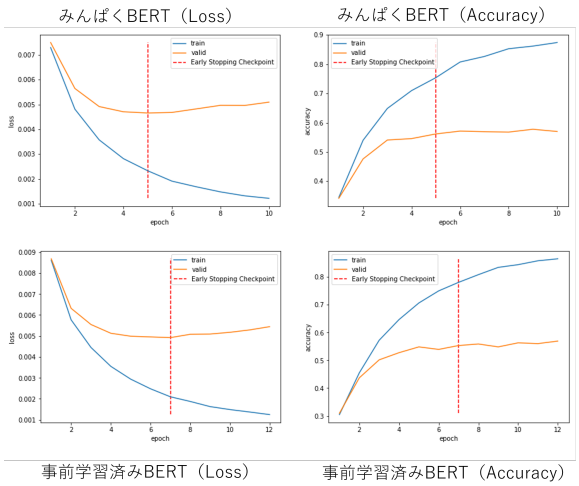


図 9　OCM ラベル分類タスク結果の学習図

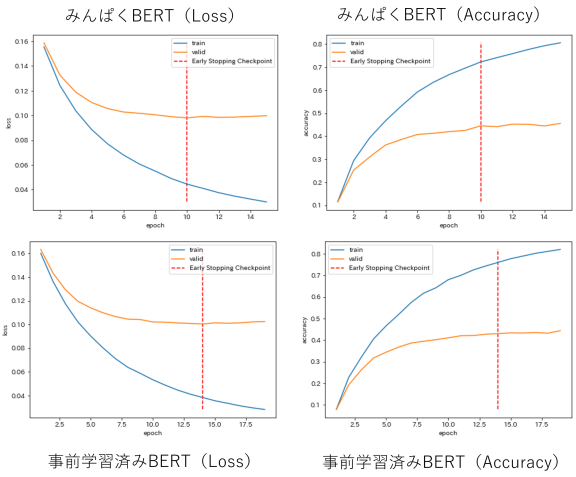


図 10　OWC ラベル分類タスク結果の学習図

- **maxlength：128**
- **early stopping：5 patience**

6.3.2　OWC ラベルを用いた展示物の地域分類タスク

OWC ラベルを用いた展示物の地域分類タスクでは，展示物データから OWC ラベルを含む文書のみを取得し，12,256 件のデータを用いてファインチューニングをおこなった．データはすべてで，15,320 件であり，これを 8:1:1 に分割し，訓練データ 12,256 件，検証データ 1,532 件，テストデータ 1,532 件とした．OWC ラベル分類はマルチラベル分類問題として解いた．ロス関数は CrossEntropyLoss を用いた．事前学習モデルには，東北大学の事前学習済み BERT モデルを用いた．トークナイザは MeCab+WordPiece を用いた．学習結果はみんぱく BERT のファインチューニングは 10epoch で終了し，事前学習済み BERT の学習は 14epoch で終了した．

ハイパーパラメータは以下の通りである．

- **batch size：32**
- **optimizer：Adam** [9]
- **ロス関数：Cross Entropy Loss**
- **学習率：2e-5**
- **Dropout：0.1**
- **maxlength：128**
- **early stopping：5 patience**

6.4　結　　果

文化財に関する 2 つのタスクのファインチューニング結果を表 2 と表 3 に示す．まず，OCM ラベル分類タスクでのみんぱく BERT と事前学習済み BERT の比較をおこなった．評価には，テストデータ 1,466 件を用いた．それぞれの正解率は，表 2 に示すように，事前学習済み BERT の正解率が 0.537 であるのに対して，みんぱく BERT の正解率は 0.547 であった．このことから，みんぱく BERT が良い結果を示したことがわかる．また，表 3 と図 9 に示すように，学習にかかる epoch 数にも違いが見られた．学習にかかる epoch 数は事前学習済み BERT が 7epoch に対して，みんぱく BERT は 5epoch と学習速度に差がでた．結果として，OCM ラベル分類タスクにおいて，みんぱく BERT が正解率と学習速度において良い結果であった．

表 2　みんぱく BERT と事前学習済み BERT の各タスクのファインチューニング結果の正解率

ファインチューニングタスク名	みんぱく BERT	事前学習 BERT
OCM ラベル分類タスク	**0.547**	0.537
OWC ラベル分類タスク	**0.437**	0.433

表 3　みんぱく BERT と事前学習済み BERT の各タスクのファインチューニングの学習が終了したエポック数

ファインチューニングタスク名	みんぱく BERT	事前学習 BERT
OCM ラベル分類タスク	**5**	7
OWC ラベル分類タスク	**10**	14

次に，OWC ラベル分類タスクでの，みんぱく BERT と事前学習済み BERT の比較をおこなった．評価には，テストデータ 1,532 件を用いた．それぞれの正解率は，表 2 に示すように，事前学習済み BERT の正解率が 0.433 であるのに対して，みんぱく BERT の正解率は 0.437 であった．このことから，みんぱく BERT が良い結果を示したことがわかる．また，表 3 と図 10 に示すように，学習にかかる epoch 数にも違いが見られた．学習にかかる epoch 数は事前学習済み BERT が 14epoch に対して，みんぱく BERT は 10epoch と学習速度に差がでた．結果として，OWC ラベル分類タスクにおいて，みんぱく BERT が正解率と学習速度において良い結果であった．これらの結果から，追加学習をおこなったみんぱく BERT は，博物館に関するタスクにおいて，事前学習済み BERT より，タスクの正解率と学習速度において良い性能であるといえる．

6.5　モデルの公開

本研究で追加学習したみんぱく BERT は，誰でも使用可能な言語資源として公開した．みんぱく BERT は，Hugging Face のモデル公開ページ [8]にて公開した．Hugging Face は Transformer を中心とした様々な深層学習モデルを実装し，公開しているサイトである．コード以外にも，データセットや学

8：https://huggingface.co/ohshimalab/bert-base-minpaku

習済みモデルの公開もおこなっており，本サイトから誰でもリソースを入手可能となっている．モデルの公開により，学習済みの，みんぱく BERT を元に，ファインチューニングをおこなうことが誰でも可能である．

7 まとめと今後の課題

本論文では，みんぱくデータで追加学習した BERT モデル（みんぱく BERT）を提案し，事前学習モデルの事前学習済み BERT との比較をおこなった．まず，事前学習済みの BERT モデルを，みんぱくデータで追加学習をおこなった．この時，未知語を考慮した状態で追加学習をおこなうことで，事前学習では扱えなかったみんぱくデータに含まれる語彙を考慮した．次に，文化財に関するファインチューニングタスクを 2 件おこない，それぞれの結果を比較した．比較するタスクは，文化財に付与されている OCM ラベルデータと OWC ラベルデータを用いた分類タスクをおこなった．OCM ラベルデータを用いた用途分類タスクでは，マルチラベル問題として BERT モデルを学習した．結果として，みんぱく BERT が正解率が高く，学習にかかるエポック数も少ない傾向にあった．OWC ラベルデータを用いた地域分類タスクにおいても，みんぱく BERT の方が正解率が高く，学習にかかるエポック数も少ない傾向にあった．以上の結果から，みんぱく BERT は，事前学習済み BERT モデルに対して，分類性能や学習性能において良い性能を示すことがわかった．本研究では，学習したみんぱく BERT モデルを言語資源として，Hugging Face にて公開した.. 今後の予定は，未知語の追加の有無による比較をおこなう予定である．

謝　　辞

本研究は JSPS 科学研究費助成事業 JP21H03775, JP21H03774, JP21H03554, JP18K18161, JP16H01756, ならびに，2021 年度国立情報学研究所公募型共同研究（21S1001, 21S1002）の助成を受けたものです．本研究の実施にあたっては，国立民族学博物館より提供いただいた展示物データベースを利用しました．また，HRAF Association より，OWC，OCM のデータをいただき，独自に翻訳して利用しました．ここに記して謝意を表します．

文　　献

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[2] Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. IMHO fine-tuning improves claim detection. In *Proceedings of NAACL'19*, pp. 558–563, 2019.

[3] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL'19*, pp. 4171–4186, 2019.

[5] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

[6] Xiaochuang Han and Jacob Eisenstein. Unsupervised domain adaptation of contextualized embeddings: A case study in early modern english. *arXiv preprint arXiv:1904.02817*, 2019.

[7] Jimin Hong, TaeHee Kim, Hyesu Lim, and Jaegul Choo. AVocaDo: Strategy for adapting vocabulary to downstream domain. In *Proceedings of EMNLP'21*, pp. 4692–4700, 2021.

[8] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR'15*, 2015.

[10] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

[11] 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.

[12] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English tweets. In *Proceedings of EMNLP'20*, pp. 9–14, 2020.

[13] Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in NLP - A survey. *arXiv preprint arXiv:2006.00632*, 2020.

[14] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *arXiv preprint arXiv:2005.12833*, 2020.

[15] Timo Schick and Hinrich Schütze. BERTRAM: Improved word embeddings have big impact on contextualized model performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3996–4007, Online, July 2020. Association for Computational Linguistics.

[16] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NIPS'17*, pp. 5998–6008, 2017.

[18] Bowen Wang, Liangzhi Li, Yuta Nakashima, Takehiro Yamamoto, Hiroaki Ohshima, Yoshiyuki Shoji, Kenro Aihara, and Noriko Kando. Image retrieval by hierarchy-aware deep hashing based on multi-task learning. In *Proceedings of ICMR'21*, pp. 486–490, 2021.

[19] Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 460–470, 2021.

[20] 神門典子, 大島裕明, 相原健郎, 荘司慶行, 白石晃一, 山本岳洋, 山本祐輔, 楊澤華. 提示型検索モデルに基づくミュージアム鑑賞体験の提案. じんもんこん 2019 論文集, pp. 127–132, 2019.

[21] 壹岐太一, 金沢輝一, 相澤彰子. 学術分野に特化した事前学習済み日本語言語モデルの構築. Technical report, 国立情報学研究所.

[22] 荘司慶行, 大島裕明, 神門典子, 相原健郎, 白石晃一, 瀧平士夫, 中島悠太, 山本岳洋, 山本祐輔, 楊澤華. 提示型検索に基づくミュージアム電子ガイドを中核とした事前・事後学習支援. じんもんこん 2020 論文集, pp. 81–88, 2020.