ハンドメイド作品を扱うECサイトに特化したBERTを用いた言語モデル構築に向けた取り組み

酒井 敏彦[1,2,a)]　三宅 悠介[1]　栗林 健太郎[1]

概要：自然言語処理の技術は，EC サイトで扱うテキストデータを対象とする，質問応答や商品の分類などのタスクに活用されている．ハンドメイド作品を扱う EC サイトにおける自然言語処理の課題は (1) 人手でタスクを解くのは困難，(2) ハンドメイド作品を扱う EC サイトの作品が多様，(3) ハンドメイド作品を扱う EC サイトの構造的な変化への追従が困難，の 3 つが挙げられる．本研究では，各課題に対して (1) 機械的にタスクを解くことができる，(2) 扱う作品が多様であっても作品の特徴を捉えられる，(3) 汎用的なモデルから fine-tuning することで構造的な変化へ追従可能，という理由から BERT+fine-tuning のモデルに着眼した．本報告では，ハンドメイド作品を扱う EC サイトの課題を含むタスクのうち，商品分類のタスクにおいて，比較評価を行った．ベースライン手法は従来から一般的に用いられる TF-IDF と分類器を用いた．結果として，上記の課題を解決し，BERT+fine-tuning のモデルが F1-score で良い分類性能であることを示した．今後は他のタスクへの応用を検討していく．

キーワード：BERT, EC サイト, カテゴリ分類, 言語モデル

1. はじめに

自然言語処理の技術は，EC サイトで扱うテキストデータを対象とする，質問応答や商品の分類などのタスクに活用されている．我々の所属する GMO ペパボ株式会社では，ハンドメイド作品を対象とした CtoC の EC サイト「minne」[2] を運営している．以降は，ハンドメイド作品を対象とした CtoC の EC サイトを「ハンドメイド作品を扱う EC サイト」と呼ぶ．また，minne で使われている用語に統一するため「商品」を「作品」,「ユーザ」を「購入者」,「販売者」を「作家」として定義する．ハンドメイド作品を扱う EC サイトにおける自然言語処理の課題として以下の 3 点が挙げられる．

課題 (1) 人手でタスクを解くのは困難

課題 (2) ハンドメイド作品を扱う EC サイトの作品が多様

課題 (3) ハンドメイド作品を扱う EC サイトの構造的な変化への追従が困難

以降は，自然言語処理の課題を含むタスクとして，作品のカテゴリ分類を例として考える．課題 (1) は作品のカテゴリ分類を運営側が人手で行うことは労力の面から現実的ではない．日々大量の作品が出品されるため，運営側で目視で作品を確認し，どのカテゴリに分類するかを判断するのは困難である．このことから，人手で分類するのではなく機械的に分類を行うことが必要となってくる．課題 (2) はハンドメイド作品を扱う EC サイトの作品が多様であることが挙げられる．minne はハンドメイド作品を扱う EC サイトであり，作家が異なる場合，同じ作品は存在しない．そのため，取り扱う作品の種類が多様であることから，適切なカテゴリへ分類するにはハンドメイド作品を扱う EC サイトの商品知識がないと困難である．また，運営側の対応者の商品知識の差によって分類の判断に差が生じる可能性もあり困難である．課題 (3) は運用的な課題として，ハンドメイド作品を扱う EC サイトの構造的な変化への追従が困難であることが挙げられる．例えば，カテゴリの新規作成や削除，統廃合を行う場合はカテゴリの作品集合に共通する特性が変化することから，その都度，手動で分類を行うことは困難である．

本研究では，各課題に対して (1) 機械的にタスクを解くことができる，(2) 扱う作品が多様であっても作品の特徴を捉えられる，(3) 汎用的なモデルから fine-tuning することで

[1] GMO ペパボ株式会社 ペパボ研究所
Pepabo R&D Institute, GMO Pepabo, Inc., Tenjin, Chuo ku, Fukuoka 810-0001 Japan
[2] 九州大学 大学院システム情報科学府 情報知能工学専攻
Department of Advanced Information Technology, Graduate School of ISEE, Kyushu University
[a)] toshihiko.sakai@pepabo.com

構造的な変化へ追従可能，という理由から BERT[1]+fine-tuning のモデルに着眼した．また，fine-tuning 以外にも BERT の事前学習済みモデルを追加学習する方法がある．本研究では，fine-tuning と追加学習を以下のように定義する．fine-tuning は，教師有りデータを用いた学習により，BERT の事前学習済みモデルと分類器のパラメータ調整を行う．追加学習は，BERT の事前学習済みモデルに教師なしのコーパスデータで追加学習することをいう．BERT+fine-tuning のモデルで本研究の課題について考えると，(1) は自明であり，(2) は fine-tuning により BERT の事前学習済みモデルと分類器のパラメータ調整を行うことで効率的に作品の特徴を捉えられる．(3) は一度汎用的なモデルを構築すれば，構造的な変化があったとしても，fine-tuning の部分のみを再度実施すれば良い．また，(3) については，分野に特化したコーパスで事前学習済みの BERT モデルを追加学習する研究がさかんに行われている．BioBERT[4] では BERT の事前学習済みモデルを医学論文のデータで追加学習することで固有表現抽出のタスク性能を向上させた．また，金融文章関連コーパスで追加学習することで金融版 BERT モデルを開発している企業もある [6]．これらの取り組みは，分野に特化したコーパスで BERT の事前学習済みモデルを追加学習することでタスクの性能向上が可能であることを示唆している．そこで，BERT の事前学習済みモデルからハンドメイド作品を扱う EC サイトのデータを追加学習することで，ハンドメイド作品を扱う EC サイトに特化したモデルを獲得できる可能性がある．しかし，今回は追加学習については学習コストやモデル学習時間を考慮し，今後の課題とした．

本研究では，ハンドメイド作品を扱う EC サイトの課題を含むタスクのうち，作品のカテゴリ分類のタスクにおいて，比較評価を行う．作品情報である作品のタイトルと説明文を利用し，作品のカテゴリを分類する．従来から一般的に用いられる TF-IDF により文書をベクトル化し，SVM 及びロジスティック回帰を分類器とする手法をベースライン手法とする．

本研究の貢献は以下の通りである．

(1) ハンドメイド作品を扱う EC サイトにおける自然言語処理のタスクを解くことを目的に作品のカテゴリ分類タスクを複数手法で比較評価した
(2) 実験の結果，ハンドメイド作品を扱う EC サイトの実データに対して BERT+fine-tuning のモデルが F1-score で良い分類性能であることを示した

本論文の構成を述べる．2 章で，実験条件及び BERT+fine-tuning のモデルとベースライン手法について説明する．3 章では結果と考察について述べ，4 章でまとめる．

2. 実験

本章では，実験条件について記載していく．

2.1 実験方法と比較手法

本研究では，ハンドメイド作品を扱う EC サイトの課題を含むタスクのうち，作品のカテゴリ分類のタスクにおいて，複数手法の比較評価を行う．minne では各作品に小カテゴリ及び大カテゴリが付与されており，大カテゴリは小カテゴリを包含している．例えば，大カテゴリ「アクセサリー」の下の階層には「ピアス」,「イヤリング」,「ネックレス」などの小カテゴリが存在する．

本研究では，作品文書から小カテゴリまたは大カテゴリを分類する 2 種類のタスクに取り組む．したがって，モデルが予測するクラス数が異なるタスクに取り組むことでクラス数の違いによる性能比較が可能になる．本研究では，以下の 2 種類の分類タスクに取り組む．

- 作品の小カテゴリを分類するタスク
- 作品の大カテゴリを分類するタスク

作品にはタイトルと説明文が付与されているため，タイトルのみを用いる場合，タイトル及び説明文を用いる場合の 2 通りの実験を行う．したがって，モデルに入力する作品文書の長さが変わることで文書の長さによる性能比較が可能になる．本研究では，モデルへの入力文書は以下の 2 種類とする．

- タイトルのみ
- タイトル及び説明文

ベースライン手法は TF-IDF により作品文書のベクトル化を行い，分類器としては SVM とロジスティック回帰を用いる．その他の分類器については決定木，ランダムフォレスト，AdaBoost を少量のデータを用いて予備実験を行ったが，分類性能が低かったため今回は比較手法から除外した．ベースライン手法との比較手法として BERT+fine-tuning の分類モデルを構築する．本研究では，以下の 3 種類のモデルを比較する．

- TF-IDF，SVM
- TF-IDF，ロジスティック回帰 (LR)
- BERT+fine-tuning

2.2 データセットの概要

各カテゴリにおいて精度の確保が十分と思われる，かつ，計算リソース内で学習可能な件数を仮として 10 万件程度と設定した．ここでいう精度の確保が十分とは，作品全体における小カテゴリ毎の総作品数が 1,000 件以上と設定した．結果として，2021 年 6 月 29 日時点の minne の作品群から 104,161 件を抽出した．

minne の作品にはタイトル，説明文，小カテゴリ，大カテ

表 1　データセットの概要

TF-IDF	サンプル数	語彙数	文の長さ（平均）
タイトル	101,245	39,454	5.44
タイトル+説明文	101,245	39,454	63.9
BERT	サンプル数	語彙数	文の長さ（平均）
タイトル	101,245	13,046	11.27
タイトル+説明文	101,245	21,331	184.45

ゴリが付与されている．今回抽出したデータセットでは，1,000 件未満の作品数である小カテゴリは分類から除いているため，小カテゴリの場合は 239 及び大カテゴリの場合は 19 のクラス数を分類する多値分類問題となった．前処理として，各作品のタイトルと説明文については，括弧記号，全角空白，URL を削除した．タイトルもしくは説明文が存在していない作品は実験対象から除外した．本研究では，作品のデータセットは学習データとテストデータを 9:1 の割合にランダムに分割した．その結果，学習データは 91,120 件，テストデータは 10,125 件となった．データセットの概要を**表 1** に示す．

2.3　BERT による fine-tuning

BERT の事前学習済みモデルは東北大学が公開している日本語学習モデル（以降，Tohoku-BERT と呼ぶ）を使用した．[11] Tohoku-BERT は日本語 Wikipedia をコーパスとして，事前学習を行ったものである．Tohoku-BERT では，tokenizer として MeCab[10] 及び NEologd[9] を使用しており，WordPiece アルゴリズム [7] によりサブワード化を行っている．今回は tokenizer と事前学習済みモデルの両方について Tohoku-BERT を使用した．

学習には Huggingface の Transformers ライブラリ [3] の BertForSequenceClassification を使用した．この関数は事前学習された BERT モデルの [CLS] の出力に線型変換分類器を加えたモデルとなっている．Tohoku-BERT のパラメータを初期値として使うことで，少ない学習データでの fine-tuning により，高い性能を得ることが期待できる．fine-tuning により，学習データの特徴を捉えつつ，BERT と分類器の両方のパラメータを学習できる．今回は多値分類問題のため，交差エントロピー誤差を損失関数として用いている．最適化関数は AdamW[5] を用い，学習率 (lr) は $2e^{-5}$ とした．モデルへ入力する 1 文章のトークンの最大長は 512 とした．タイトルと説明文を合わせた文章において，最大長を超えた場合は [8] にならい head+tail の方式を採用した．head+tail の方式では最大長を超えた場合，文章の先頭と末尾を残す方式である．ただし，タイトルについては重要度が高いと考えたため，タイトルのテキストは全て残し，説明文に対して head+tail の方式を適用した．結果として，説明文の先頭と末尾の部分を残し，中間部分を削除した．BERT での語彙数と文の長さの平均は表 1 に記載している．

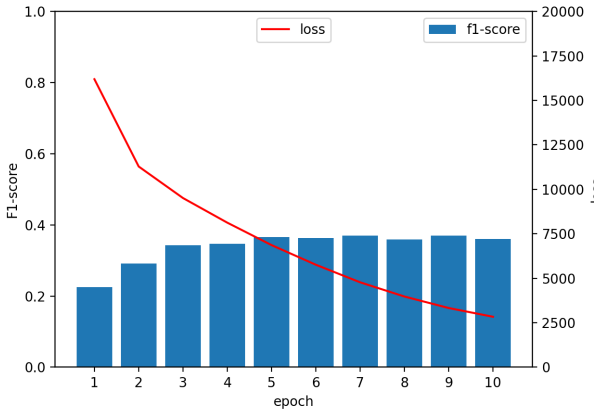

図 1　小カテゴリ分類タスクにおける BERT+fine-tuning の学習状況 (入力はタイトル)

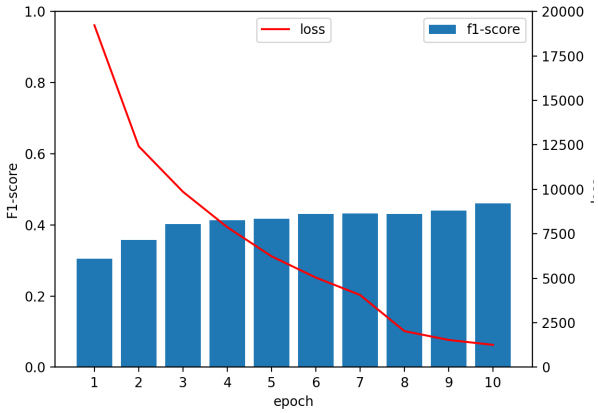

図 2　小カテゴリ分類タスクにおける BERT+fine-tuning の学習状況 (入力はタイトル+説明文)

図 1 と**図 2** に小カテゴリでの fine-tuning を行った際の loss と F1-score の推移を表す．今回は loss の推移から epoch10 までの学習とした．なお，大カテゴリのモデルについては小カテゴリのモデルの出力結果から，大カテゴリを一意に決めることができるため，大カテゴリを分類するモデルは構築していない．

2.4　TF-IDF による実験

この節では，TF-IDF の実験について述べる．TF-IDF は単語の出現頻度 (TF) と逆文書頻度 (IDF) を組み合わせることで表される．2.2 節と同様の前処理を実施後，MeCab-NEologd により形態素解析を行った．今回は品詞が名詞の単語のみを用いた．各文書を scikit-learn の TfidfVectorizer によりベクトル化した．語彙数と文の長さの平均は表 1 に記載している．TF-IDF においては，入力をタイトル+説明文にした場合はメモリ不足を防ぐため，語彙数をタイトルにした場合と同じ 39,454 に合わせた．語彙数が BERT の方が少ない理由としては，Tohoku-BERT の語彙に依存し

表 2　小カテゴリの分類結果

タイトル	Accuracy	Precision	Recall	F1-score
TF-IDF,SVM	0.585	**0.418**	0.320	0.346
TF-IDF,LR	0.558	0.381	0.297	0.322
BERT+fine-tuning	**0.635**	0.405	**0.373**	**0.370**
タイトル+説明文	Accuracy	Precision	Recall	F1-score
TF-IDF,SVM	0.644	**0.518**	0.441	0.458
TF-IDF,LR	0.631	0.454	0.395	0.407
BERT+fine-tuning	**0.708**	0.476	**0.472**	**0.460**

ていると考えられるためである．一方，文の長さは BERT が長い理由としては，BERT ではトークンをそのまま用いているが，TF-IDF では品詞を名詞に限定しているためである．

分類器としては SVM とロジスティック回帰を用いた．それぞれのハイパーパラメータは以下の通りである．SVM は正則化係数 $C = [1, 10, 100]$，rbf カーネル係数 $gamma = [0.01, 0.1, 1]$，カーネルタイプ $kernel = [linear, rbf]$ を候補として用いた．ロジスティック回帰は正則化係数 $C = [0.01, 0.1, 1, 10, 100]$ を候補として用いた．3 分割交差検証を行い，最適なハイパーパラメータを求めた後，求めたハイパーパラメータでテストデータを評価した．

2.5　評価指標

評価指標として文書分類や検索精度の評価を行う際に用いられる Accuracy, Precision, Recall, F1-score を用いた．なお，Precision, Recall, F1-score については macro-f1 を用いた．

3.　実験結果と考察

本章では実験結果と考察について述べる．各評価指標において，最も性能が良い結果を太字にしている．なお，分類結果の表において，BERT+fine-tuning のモデルについては，学習時に F1-score が一番高かった epoch のモデルの結果を記載している．そのため，結果の表には，BERT+fine-tuning のモデルはタイトルのみは epoch が 7，タイトル+説明文は epoch が 10 のモデルでの結果を記載している．

3.1　小カテゴリの分類結果と考察

表 2 に小カテゴリの分類結果を示す．この結果から，Precision 以外の評価指標において，BERT+fine-tuning のモデルが良い結果となった．一方，Precision については TF-IDF，SVM の手法が良い結果となった．

次に BERT+fine-tuning のモデルにおける，タイトルと説明文を学習させた結果を混同行列にて確認した．**表 3** は分類結果の誤り件数が多い順の結果である．誤り件数が 30 件以上の場合を抜粋した．小カテゴリの下の括弧内に小カテゴリが所属する大カテゴリを記載した．この結果は，正解クラスと予測クラスの所属する大カテゴリは一致していることを示している．一方，同じ大カテゴリに所属する小カテゴリを分類する場合において，分類を誤ったことがわかる．この結果から，関連する小カテゴリ間の違いを正しく分類できなかったことが推察される．そこで，大カテゴリを分類した後に，大カテゴリに属するデータのみで個別の小カテゴリを分類するモデルを構築することが考えられる．そうすれば，大カテゴリごとの小カテゴリについて，さらに分類性能を向上できると考えている．

表 3　小カテゴリのカテゴリ誤り分析

誤り件数	正解クラス	予測クラス
71	ヘアアクセサリー (アクセサリー)	ヘアゴム (アクセサリー)
70	イヤリング (アクセサリー)	ピアス (アクセサリー)
47	ヘアアクセサリー (アクセサリー)	バレッタ・ヘアクリップ (アクセサリー)
39	ピアス (アクセサリー)	イヤリング (アクセサリー)
35	ヘアゴム (アクセサリー)	ヘアアクセサリー (アクセサリー)
32	バッグ (バッグ・財布・小物)	トートバッグ (バッグ・財布・小物)

表 4　大カテゴリの分類結果

タイトル	Accuracy	Precision	Recall	F1-score
TF-IDF,SVM	0.772	**0.704**	0.520	0.580
TF-IDF,LR	0.750	0.620	0.511	0.556
BERT+fine-tuning	**0.820**	0.640	**0.630**	**0.630**
タイトル+説明文	Accuracy	Precision	Recall	F1-score
TF-IDF,SVM	0.834	**0.775**	0.640	0.679
TF-IDF,LR	0.824	0.750	0.631	0.665
BERT+fine-tuning	**0.860**	0.730	**0.710**	**0.720**

3.2　大カテゴリの分類結果と考察

表 4 に大カテゴリの分類結果を示す．表 4 の結果から大カテゴリを分類する場合においても小カテゴリの分類結果と同様に Precision は TF-IDF,SVM の手法が良い結果となり，それ以外の評価指標においては BERT+fine-tuning のモデルが良い結果となった．

大カテゴリの分類結果は小カテゴリの分類結果よりも全体的に良い性能となることがわかった．これは分類するクラス数が少ない方が分類性能が良いことを示している．また，大カテゴリにおいても，タイトルだけでなく説明文も用いる方が性能が向上することがわかった．これは入力として用いる文書量は多い方が良いことを示している．

3.3　全体考察

これまでの結果から，小カテゴリの分類結果，大カテゴリの分類結果いずれにおいても，BERT+fine-tuning のモデルが Accuracy，Recall，F1-score で良い結果となった．

入力する文書量の観点では，小カテゴリ，大カテゴリのいずれにおいてもタイトルだけでなく説明文も用いる方が性能が向上することがわかった．これは BERT+fine-tuning のモデルおよびベースライン手法のどちらにおいても同じ傾向であった．このことから入力として用いる文書量は多い方が良いことがわかる．クラス数の観点では，小カテゴリよりも大カテゴリの分類結果の方が全体的に良いことが確認できた．このことから，分類するクラス数が少ない方が良い性能であることがわかる．

全体的な性能について，小カテゴリの分類結果ではタイトル+説明文についての F1-score:0.460 は高い性能とは言えない．小カテゴリの分類においてはクラス数も多く粒度が細かい分類は難しいと考えられる．そのため，3.1 節で考察した大カテゴリ毎に小カテゴリを分類するモデルを構築し，分類する方法を検討していきたい．大カテゴリの分類結果については F1-score:0.720 となっており，こちらも高い性能であるとはいえない．この結果については fine-tuning するデータを増やすことでさらに性能が向上できる可能性があると考えている．また，Precision については TF-IDF,SVM のほうが高い性能であったが，こちらについても fine-tuning するデータを増やすことで，性能が向上できる可能性があると考えている．

4. おわりに

本研究では，ハンドメイド作品を扱う EC サイトの課題を含むタスクのうち，作品のカテゴリ分類のタスクにおいて，比較評価を行った．結果としては，BERT+fine-tuning のモデルが小カテゴリ及び大カテゴリの分類において，Accuracy，Recall，F1-score で良い結果となった．この結果から，BERT+fine-tuning のモデルの汎用性の高さがうかがえ，モデルが作品の書きぶりやニュアンスを捉えることができている可能性がある．

今後の課題としては，作品のカテゴリ分類以外のタスクでも BERT+fine-tuning のモデルの性能評価をしていきたい．また，BERT の事前学習済みモデルからハンドメイド作品を扱う EC サイトのコーパスにより追加学習を行うことで性能が向上するかを検証したいと考えている．ハンドメイド作品を扱う EC サイトに特化したモデルから，カテゴリ分類のタスクや他のタスクに対しての検証や課題 (3) に挙げている構造的な変化に対しても，fine-tuning することで，高い性能を得るか検証していきたい．

参考文献

[1] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019).
[2] GMO ペパボ株式会社：minne, https://minne.com/.
[3] Huggingface: Transformers library, https://huggingface.co/transformers/.
[4] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H. and Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, Vol. 36, No. 4, pp. 1234–1240 (2020).
[5] Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, *arXiv preprint arXiv:1711.05101* (2017).
[6] NTT データ：金融業界向け自然言語処理技術の検証開始〜金融版 BERT モデルの開発〜，https://www.nttdata.com/jp/ja/news/release/2020/071000/.
[7] Schuster, M. and Nakajima, K.: Japanese and Korean Voice Search, *International Conference on Acoustics, Speech and Signal Processing*, pp. 5149–5152 (2012).
[8] Sun, C., Qiu, X., Xu, Y. and Huang, X.: How to fine-tune bert for text classification?, *China National Conference on Chinese Computational Linguistics*, Springer, pp. 194–206 (2019).
[9] Toshinori, S.: Neologism dictionary based on the language resources on the Web for Mecab (2015).
[10] 工藤　拓：MeCab, https://taku910.github.io/mecab/.
[11] 東北大学公開の日本語事前学習済み BERT：https://github.com/cl-tohoku/bert-japanese.