# BERT の転移学習と Mis-leading データの削除による識別精度の改善

岩本昇太
茨城大学工学部
情報工学科
18t4012t@vc.ibaraki.ac.jp

新納浩幸
茨城大学大学院理工学研究科
情報科学領域
hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

## 概要

自然言語処理のタスクを機械学習で解決するとき，ラベル付き訓練データの不足に対処する必要がある．転移学習はターゲット領域のラベル付き訓練データが全く無い場合にも有効な手法であるが，「負の転移」と呼ばれる問題に対処する必要がある．本稿では BERT を用いた文書分類タスクで転移学習を行うときに負の転移に対処するため Mis-leading データの削除と BERT の転移学習を行い，文書分類タスクで識別精度を改善できることを確認した．

## 1 はじめに

自然言語処理のタスクの多くは，機械学習の手法により解決できる．ただし，そのためには大量のラベル付き訓練データが必要となる．BERT [1] のような事前学習済みモデルを用いる場合も，fine-tuning のためのラベル付き訓練データは必要となる．転移学習を用いれば，ターゲット領域の訓練データが全く無い場合にも対応できる．しかし，転移学習を用いる場合は負の転移 [2] と呼ばれる問題に対処する必要がある．負の転移は，ソース領域のデータとターゲット領域のデータの性質が著しく異なるときにソース領域のデータを訓練に用いると識別精度が悪化する現象である．

本稿では，BERT を用いた文書分類タスクで転移学習を行うときに Mis-leading データの削除と BERT の追加学習を行い，識別精度の改善を試みる．Mis-leading データは，ソース領域の訓練データのうちターゲット領域での識別精度に悪影響を及ぼすものである．事前に Mis-leading データを削除することで，ターゲット領域での識別精度の向上が期待できる．また，ターゲット領域のラベルなしデータを用いて BERT の追加学習を行うことで，fine-tuning 後の識別精度の向上が期待できる [3]．Mis-leading データの削除に加えて BERT の追加学習も行い，識別精度のさらなる改善を図る．

## 2 関連研究

### 2.1 BERT

Bidirectional Encoder Representations from Transformers(BERT) [1] は 2018 年に Google より公開された事前学習済みモデルである．BERT は入力としてトークン列（単語もしくはサブワード）を受け取り，それに対応する埋め込み表現列を出力する．ここで，出力される埋め込み表現は文脈を考慮したものとなっている．大規模コーパスにより事前学習したモデルを下流タスクのラベル付きデータで fine-tuning することで，文書分類・固有表現抽出・質問応答など様々なタスクに対応できる．日本語版の事前学習済みモデルとして，東北大学で公開されているモデル[1)]などがある．

### 2.2 タスク適応型事前学習

Gururangan ら [3] は BERT の派生モデルである RoBERTa [4] の領域適応のために タスク適応型事前学習 (Task-Adaptive Pretraining; TAPT) を提案した．タスク適応型事前学習は，事前学習済みモデルに対して解きたいタスクのテキストデータを用いた追加の事前学習を行う手法である．

### 2.3 負の転移

負の転移 (Negative Transfer) [2] は，ソース領域のデータとターゲット領域のデータの性質が著しく異なるときにソース領域のデータを訓練に用いると識別精度が悪化する現象である．転移学習では，ど

1) https://github.com/cl-tohoku/bert-japanese

のようにして負の転移を検知・回避するかが問題となる．

## 3 提案手法

タスク適応型事前学習を行えば，BERT を用いた文書分類タスクにおける識別精度の改善は実現できる．しかし，転移学習を用いる場合に負の転移が生じうるという問題は解消されていない．本稿では，BERT を用いた文書分類タスクで転移学習を行うときに Mis-leading データの削除と BERT の追加学習を行うことで識別精度を改善する手法を提案する．

Mis-leading データを削除する方法は以下の通りである．

- ターゲット領域のラベルなしデータを離散確率分布 $P$ で表現する
- ソース領域の個々の訓練データを離散確率分布 $Q_k$ で表現する（$k$ 番目の文書に $Q_k$ が対応）
- Kullback-Leibler 情報量 $D_{KL}(P \parallel Q_k)$ 計算し，その値が大きいラベル付きデータを削除する

### 3.1 BERT を用いた文書分類タスクの転移学習

本稿では，下記の手順により BERT を用いた文書分類タスクの転移学習を行う．

1. BERT をソース領域のラベル付きデータで fine-tuning して文書分類器を作成する．
2. 作成した文書分類器により，ターゲット領域で文書分類を行う．

BERT の fine-tuning では，ターゲット領域のラベル付きデータは使用しない．

### 3.2 Kullback-Leibler 情報量

Kullback-Leibler 情報量は，情報理論や統計学において 2 つの確率分布の類似度を表す尺度である．ある 2 つの離散確率分布 $P, Q$ に対する Kullback-Leibler 情報量 $D_{KL}(P \parallel Q)$ は，式 (1) となる．

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{1}$$

ここで $P(i), Q(i)$ はそれぞれ離散確率分布 $P, Q$ に従って $i$ が選ばれる確率である．

### 3.3 文書の離散確率分布表現

文書を離散確率分布として表現するためには，まず文書をベクトル化しなければならない．文書のベクトル化には Bag-of-Words を用いる．このとき，単語の重み付けには文書内での単語の出現頻度 (Term Frequency) を用いる．文書 $Doc$ のベクトル表現は式 (2) のようになる．

$$(tf_1, tf_2, \ldots, tf_{N_w}) \tag{2}$$

ここで，$tf_i$ は単語分割時に用いた辞書で $i$ 番目に登録されている単語が文書 $Doc$ 内で登場した回数，$N_w$ は単語分割時に用いた辞書に収録されている単語数である．

式 (2) より文書を実数値のベクトルで表現できるようになったが，単語を確率変数とし，文書を離散確率分布として表現するためには，式 (2) のベクトルをさらに変換する必要がある．そこで，式 (2) のベクトル表現を式 (3) のように変換する．

$$(v_1, v_2, \ldots, v_{N_w}) \tag{3}$$
$$v_i = \frac{tf_i}{Z}, \quad Z = \sum_{i=1}^{N_w} tf_i$$

確率は次の 2 条件を満たす．

- $v_i \geq 0$（確率の値は非負の実数）
- $\sum_{i=1}^{N_w} v_i = 1$（確率の総和は 1）

式 (3) のベクトルは，$Z$ の定め方より $\sum_{i=1}^{N_w} v_i = 1$（確率の総和は 1）を満たす．また，$tf_i$ は明らかに非負の値をとるため，$v_i \geq 0$（確率の値は非負の実数）を満たす．したがって，上記の方法により文書を離散確率分布として表現できる．

### 3.4 Mis-leading データの削除

Mis-leading データは，ソース領域の訓練データのうちターゲット領域での識別精度に悪影響を及ぼすものである．本節では Mis-leading データを削除する方法を説明する．

訓練データとなるラベル付きの文書の集合を $D_S = \{Doc_1, Doc_2, \ldots, Doc_N\}$ とする．$D_S$ には $N$ 件の文書が属し，$D_S$ に属する文書はすべてソース領域に属する．また，ラベルなしの文書のうちターゲット領域に属するものの集合を $D_T$ とする．

Mis-leading データを削除する手順は以下の通りである．

1. $D_T$ に属する文書すべてを 1 つの文書とみなし，3.3 節で説明した手法により離散確率分布 $P$ を得る．
2. $D_S$ に属する文書に 3.3 節で説明した手法を適用し，離散確率分布 $Q_1, Q_2, \ldots, Q_N$ を得る．$k$

番目の文書 $Doc_k$ に対応する離散確率分布が $Q_k$ である．

3. 各離散確率分布 $Q_1, Q_2, \ldots, Q_N$ について Kullback-Leibler 情報量 $D_{KL}(P \parallel Q_k)$ を計算する．
4. ラベルに基づいて訓練データをグループ分けする．各グループ内で Kullback-Leibler 情報量 $D_{KL}(P \parallel Q_k)$ の値の大きい順に文書を並べ替え，$D_{KL}(P \parallel Q_k)$ の値の大きい方から一定数の文書を取り出す．取り出した文書，すなわち $D_{KL}(P \parallel Q_k)$ の値の大きい文書を Mis-leading データとみなし，訓練データの集合から削除する．

### 3.5 BERT の追加学習

BERT を fine-tuning して文書分類器を作成する前に，タスク適応型事前学習を行う．本稿ではデータセットに含まれる全領域のラベルなしデータを用いて追加学習を行い，その後ターゲット領域のラベルなしデータを用いて追加学習を行った．

## 4 実験

BERT を用いた文書分類タスクで転移学習を行うとき，Mis-leading データの削除と BERT の追加学習を行った場合の識別精度を確認した．Mis-leading データの削除と BERT の追加学習を行わずに転移学習を行った場合（ベースライン）の識別精度と提案手法を用いたときの識別精度を比較し，提案手法の有効性を確認した．

### 4.1 事前学習済みモデル

東北大学で公開されているモデル (BERT-base-japanese) を使用した．

### 4.2 実験用データセット

実験には Webis-CLS-10 データセットを用いた．このデータセットには日本語及び英語の Amazon レビュー文書が収録されている．本実験では日本語の文書を用いる．ラベルは星の数であり，1 から 5 までの 5 段階である．ただしラベルが 3（星 3 つ）のデータは存在しない．本実験ではラベルが 4, 5 のデータを positive, ラベルが 1, 2 のデータを negative として感情分析（2 値分類）を行った．

このデータセットには books, dvd, music の 3 つの領域がある．各領域には訓練データ 2000 件，テストデータ 2000 件が収録されている．この 2 つはラベル付きデータである．また，訓練データ・テストデータとは別にラベルなしデータが収録されている．データセットの内訳を表 1 に示す．

**表 1** Amazon レビュー文書のデータセットの内訳

| | books | dvd | music |
|---|---|---|---|
| 訓練データ | 2000 | 2000 | 2000 |
| テストデータ | 2000 | 2000 | 2000 |
| ラベルなしデータ | 169780 | 68326 | 55892 |

実験では，ソース領域の訓練データを fine-tuning の訓練データとした．また，ターゲット領域の訓練データを検証用データとし，識別精度の最終的な評価にはターゲット領域のテストデータを用いた．

### 4.3 Mis-leading データの削除

3.4 節で説明した手法により，ソース領域の訓練データから Mis-leading データを削除した．文書の単語分割には BERT-base-japanese の tokenizer を用いた．削除する件数は 200, 400, 600, 800, 1000 のいずれかとし，検証用データの識別精度をもとに削除する件数を選択した．

### 4.4 BERT の追加学習

BERT-base-japanese に対してタスク適応型事前学習を行った．まず全領域のラベルなしデータ及び訓練データを用いて 10epoch の追加学習を行った．ここで，ラベルなしデータの件数は領域により異なる．本稿では books と dvd についてはラベルなしデータをランダム抽出し，music のラベルなしデータと同じ件数のラベルなしデータを用意した．music については全てのラベルなしデータを用いた．全領域のラベルなしデータ及び訓練データを用いた追加学習により得られたモデルをモデル A とする．

更に，モデル A に対してターゲット領域のラベルなしデータ及び訓練データを用いて 10epoch の追加学習を行った．どの領域についても，モデル A の追加学習で用いたものと同一のラベルなしデータで追加学習を行った．

なお，追加学習では Masked Language Model のみを行い，Next Sentence Prediction は省略した．

### 4.5 文書分類器の作成

モデル A に対してターゲット領域のラベルなしデータ及び訓練データを用いて追加学習を行っ

**表 2** 実験結果（正解率）

| ソース | ターゲット | ベースライン | 提案手法 |
|---|---|---|---|
| books | dvd | 0.8610 | 0.8895 |
| books | music | 0.8550 | 0.8870 |
| dvd | books | 0.8420 | 0.8785 |
| dvd | music | 0.8665 | 0.9005 |
| music | books | 0.8425 | 0.8720 |
| music | dvd | 0.8535 | 0.8735 |

たモデルを fine-tuning し，文書分類器を作成した．fine-tuning にはソース領域の訓練データのうち，3.4 節で説明した手法により Mis-leading データを削除したものを用いた．

### 4.6 実験結果

提案手法を用いて転移学習を行った場合の識別精度及び提案手法を用いずに転移学習を行った場合（ベースライン）の識別精度を表 2 に示す．ベースラインの数値は，追加学習を行っていない BERT-base-japanese をソース領域の訓練データ全てで fine-tuning して作成した文書分類器をターゲット領域で用いた場合の正解率である．ソース領域とターゲット領域の組み合わせは計 6 パターンあるが，全パターンについて文書分類タスクを行い，識別精度を評価した．

表 2 の通り，ソース領域とターゲット領域の組み合わせの全パターンで提案手法での正解率がベースラインの正解率を上回った．

## 5 考察

以下，ソース領域とターゲット領域の組を（ソース領域，ターゲット領域）と表記することがある．

ソース領域とターゲット領域の各組み合わせについて，Mis-leading データを削除するために計算した Kullback-Leibler 情報量 $D_{KL}(P \parallel Q_k)$ のばらつきを図 1 の箱ひげ図に表した．

ベースラインの正解率と提案手法の正解率を比較したとき，正解率の上がり幅が最も大きかったソース領域とターゲット領域の組は (dvd, books) である．また，正解率の上がり幅が最も小さかった組は (music, dvd) である．

図 1 の箱ひげ図より，(dvd, books) では他の組よりも Kullback-Leibler 情報量の散らばりが大きく，Kullback-Leibler 情報量の中央値は小さいとわかる．一方で，(music, dvd) では他の組よりも Kullback-

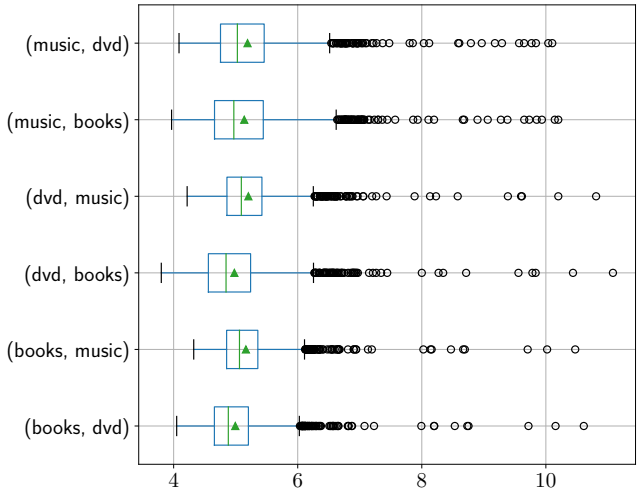


**図 1** Kullback-Leibler 情報量 $D_{KL}(P \parallel Q_k)$ のばらつき．箱の左のラベルはソース領域とターゲット領域の組（ソース，ターゲット）．

Leibler 情報量の散らばりが小さく，Kullback-Leibler 情報量の中央値は比較的大きいとわかる．このような Kullback-Leibler 情報量のばらつきが，Mis-leading データを削除する効果の大きさに影響していると考えられる．

## 6 おわりに

本稿では，BERT を用いた文書分類タスクで転移学習を行うときに Mis-leading データの削除と BERT の追加学習を行い，識別精度の改善を試みた．実験により，Mis-leading データの削除と BERT の追加学習を行わない転移学習と比較したときに識別精度が改善されることを確認できた．

## 謝辞

本研究は JSPS 科研費 JP19K12093 および 2021 年度国立情報学研究所公募型共同研究（2021-FC05）の助成を受けています．

## 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] Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. To transfer or not to transfer. In **NIPS 2005 workshop on transfer learning**, Vol. 898, 2005.

[3] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics.

[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. **CoRR**, Vol. abs/1907.11692, , 2019.