国立国語研究所学術情報リポジトリ

Sentiment Analysis That Does Not Require Training Data in the Target Language Using Billingual BERT

メタデータ	言語: jpn 出版者: 公開日: 2021-03-05 キーワード (Ja): キーワード (En): 作成者: 荘司, 響之介, 曹, 鋭, 白, 静, 馬, ブン, 新納, 浩幸, Syouji, Kyonosuke, Cao, Rui, Bai, Jing, Ma, Wen メールアドレス: 所属:
URL	https://doi.org/10.15084/00003159

二言語 BERT を利用したターゲット言語の教師データを必要としない感情分析

荘司響之介 (茨城大学大学院理工学研究科情報工学専攻) *
曹鋭 (茨城大学大学院理工学研究科情報工学専攻) †
白静 (茨城大学大学院理工学研究科情報工学専攻) ‡
馬ブン (茨城大学大学院理工学研究科情報工学専攻) §
新納浩幸 (茨城大学大学院理工学研究科情報工学専攻) ¶

Sentiment analysis that does not require training data in the target language using billingual BERT

Kyonosuke syouji (Graduate School of Science and Engineering, Ibaraki University)
Cao Rui (Graduate School of Science and Engineering, Ibaraki University)
Bai Jing (Graduate School of Science and Engineering, Ibaraki University)
Ma Wen (Graduate School of Science and Engineering, Ibaraki University)
Hiroyuki Shinnou (Graduate School of Science and Engineering, Ibaraki University)

要旨

文書分類のタスクを教師あり学習で解く場合、大量のラベル付きデータ（教師データ）が必要であり、このデータの構築コストが高いという問題がある。ただし、英語などのメジャーな言語に対しては、ラベル付けされたデータが既に存在していることも多い。この場合、英語側では分類器を学習できるため、その学習できた知識を、タスクの対象となっている言語側へ転移できれば、ターゲット言語での教師データを利用せずに、分類器を構築することができる。本論文ではそのような転移を行うために BERT を用いる。具体的には、 英語 BERT を用いて英語の訓練文書をベクトル化し、それをもとに分類器を学習する。次に、ターゲット領域の文書となる日本語の文書を、日本語 BERT を用いてベクトル化する。あらかじめ学習しておいた 2 言語間の BERT の変換器を用いて日本語の文書ベクトルを英語のベクトル空間に埋め込み、先の分類器によって識別する。これによって、ターゲット言語である日本語の訓練文書を利用せずに、日本語の文書の感情分析が可能となる。

1. はじめに

本論文では BERT(Devlin et al. (2018)) と Bilingual Word Embeddings (以下、BWE)(Mikolov et al. (2013)) を用いることで、教師データを利用しない感情分析を試みる。

* 19NM716G@vc.ibaraki.ac.jp
† 18ND305G@vc.ibaraki.ac.jp
‡ 19ND301R@vc.ibaraki.ac.jp
§ 19ND302H@vc.ibaraki.ac.jp
¶ hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

感情分析とはレビュー文書が肯定的なものか、否定的なものかを判定するタスクである。これは文書分類の一種であり、教師あり学習を用いて解決できる。しかし教師あり学習には大量のラベル付きデータ（教師データ）が必要であり、このデータの構築コストが高いという問題がある。ただし英語などのメジャーな言語に対しては、ラベル付けされたデータが既に存在していることも多い。この場合、英語側では分類器を学習できるため、その学習できた知識を、タスクの対象となっている言語側へ転移できれば、ターゲット言語での教師データを利用せずに、分類器を構築できる。本論文はそのような転移を行うために BWE の原理を利用する。

BWE とは英語や日本語などの異なる言語の分散表現を共通に扱う枠組みである。例えば英語 “dog” の分散表現と日本語「犬」の分散表現は、学習基のコーパスが異なるために、異なるベクトルであるが、概念としては同じなので、同一のベクトルとして表現できるはずである。このようなアイデアのもと、異なる言語の分散表現を同一したもの、あるいはそれらの変換を実現したものが BWE である。本論文ではそのような同一化を単語ベクトル単位ではなく文書ベクトル単位で行う。

提案手法としては、英語のラベル付き文書を BERT を用いてベクトル化し、そのベクトルを基に分類器を学習する。次にターゲット領域の文書となる日本語文書を BERT を用いてベクトル化し、それを上記の　BWE のように変換したものを先の分類器によって識別する。これによってターゲット領域側のラベル付き文書を全く利用せずに、感情分析が可能となる。

実験では、提案手法（英語の教師データを用いて日本語の感情分析を行う手法）と、単純な英語のみの感情分析を比較した。

2. 関連研究

本研究は事前学習モデルである BERT を BWE に応用したと捉えられる。ここでは事前学習モデルと BWE の 2 つの観点から関連研究を述べる。

事前学習モデルは大規模なコーパスからあらかじめ学習させた状態の言語のモデルである。様々なモデルの形態があるが、標準的にはそのモデルを利用して、入力単語列をその埋め込み表現列へ変換する。利用法としてはその変換された埋め込み表現列を直接タスクに利用する feature based の利用法と、タスクに対するネットワークを学習すると同時にそのモデル自体も学習する fine-tuning の利用法がある。BERT は feature based の利用法を用いることで文脈を考慮した単語の分散表現を得ることができる。そのようなモデルとして ELMo(Peters et al. (2018)) があげられるが、BERT はその改良版と位置づけられるモデルであり、多くの自然言語処理のデータセットに対して SOTA を実現している。

1 で述べたように、BWE とは異なる言語の分散表現を共通に扱う枠組みである。手法は様々だが、標準的には対訳のソース言語とターゲット言語の分散表現から、ソース言語をターゲット言語へ変換するモデルを学習する。そのため、BWE は本質的には変換器を学習するタスクとみなすことができる。例えば、Mikolov ら (Mikolov et al. (2013)) は、単語間の幾何学的関係は言語を通じて類似しているとういう事実に注目し、異なる言語間のベクトル空間間の関係を線形マッピングにより捉えた。そのための変換行列を次の最適化問題により学習した。

$$\min_{W} \sum_{i=1}^{n} |Wx_i - z_i|^2$$

ここで、x_i はソース言語の単語ベクトル、z_i はその翻訳のベクトルである。

従来の研究では、このような変換を単語間で行っていた。また、Word2Vec のような文脈を考慮しないモデルによって単語の分散行列を得ていた。そこで本論文では、事前学習モデルである BERT を用いて文章の分散表現を取得し、文章間でソース言語とターゲット言語間の変換を行った。

3. 提案手法

ここではまず、BERT を用いて英語のラベル付きの訓練文書をベクトル化し、それを利用して分類器を学習する。次に、日本語の文書を BERT でベクトル化したものをを日本語から英語への変換器で変換し、先の分類器で分類する。

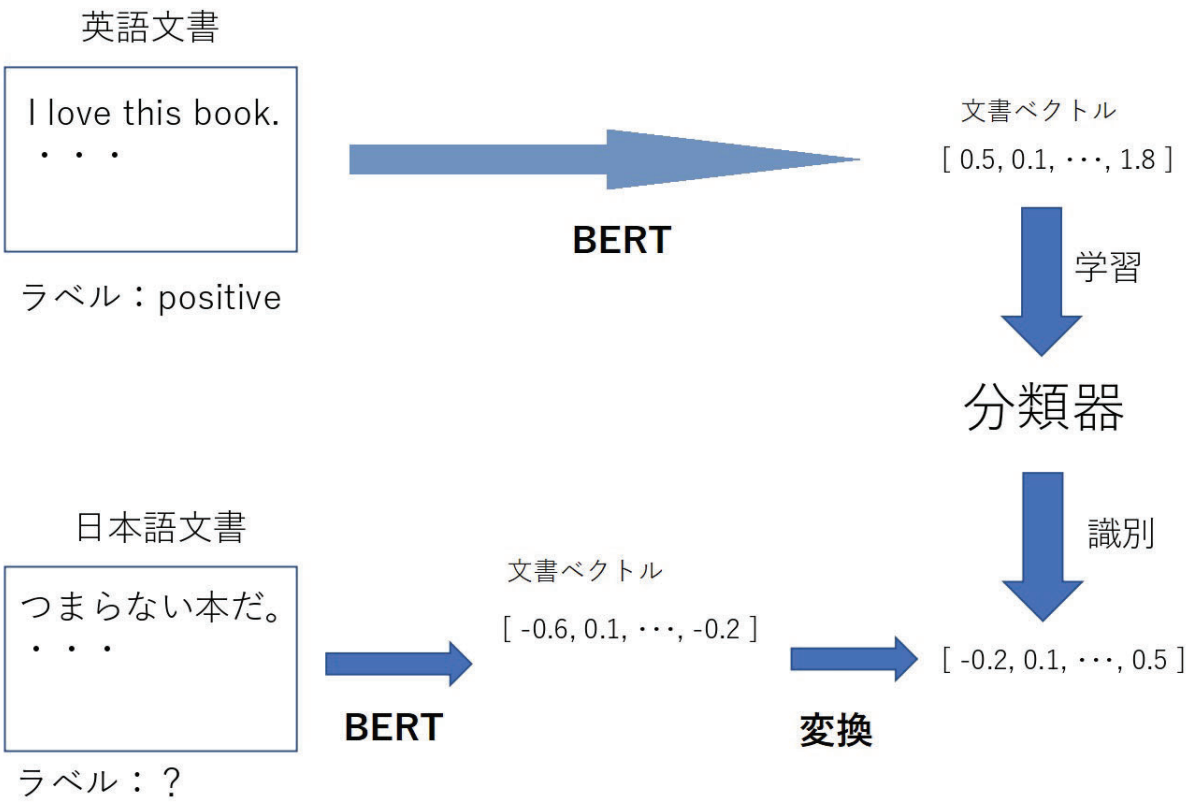


図 1　提案手法の概要図

3.1 文書のベクトル化

BERT を用いて文書中の各単語のベクトルを取得し以下のように文書全体のベクトルを求める。

$$S = \frac{w_1 + w_2 + ... + w_i}{i}$$

ここで、S は文のベクトル、w_i は i 番目の単語のベクトルを表している。

3.2 変換器の学習

変換器を学習するにあたり、まず、日英の対訳コーパスコーパスの文章を BERT にかけて各単語をベクトル化し、それらを 3.1 の方法で計算することで文章をベクトル化した。

ここで、対訳コーパスは以下で公開しているものから 5000 文ずつ使用した。

`http://www.edrdg.org/wiki/index.php/Tanaka_Corpus`

また、 BERT のモデルは、Japanese_L-12_H-768_A-12_E-30_BPE（日本語）と cased_L-12_H-768_A-12（英語）を使用した。それぞれ以下で公開されている。

- 日本語

  urlhttp://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT 日本語 Pretrained モデル

- 英語

  urlhttps://github.com/google-research/bert

このようにして得られたベクトルを以下のようなネットワークにかけ変換器を学習した。

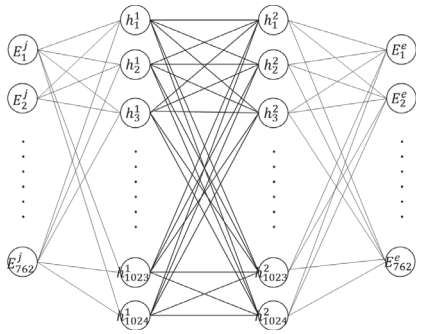


図 2　変換器の学習

ここで、$(E_1^j, ..., E_{762}^j), (E_1^e, ..., E_{762}^e)$ はそれぞれ同じ意味の日本語と英語の文のベクトルである。

4. 実験

4.1 英日の感情分析データ

感情分析の実験のために、以下のサイトから英語と日本語の感情分析のデータをダウンロードした。

https://webis.de/data/webis-cls-10.html

このデータは、日本語と英語でそれぞれ books、DVD、music の 3 つの領域を持ち、各データ（文書）数は 2000 である。また、日本語のテスト文書を英語に翻訳した文書を持つ。

4.2 実験結果

提案手法によって感情分析を行った。分別器の学習アルゴリズムには 3 層のニューラルネットワークをを用いた。英語の文書を教師データとして次の 3 つのテスト文書に対して実験を行った。

- 日本語の文書（提案手法により変換）
- 英語の文書
- 日本語から英語に翻訳した文書

結果を表 1 に示す。

表 1　BERT を用いた感情分析

テスト文書	books	DVD	music
日本語の文書	0.64	0.60	0.57
英語の文書	0.65	0.59	0.60
日英翻訳した文書	0.63	0.55	0.53

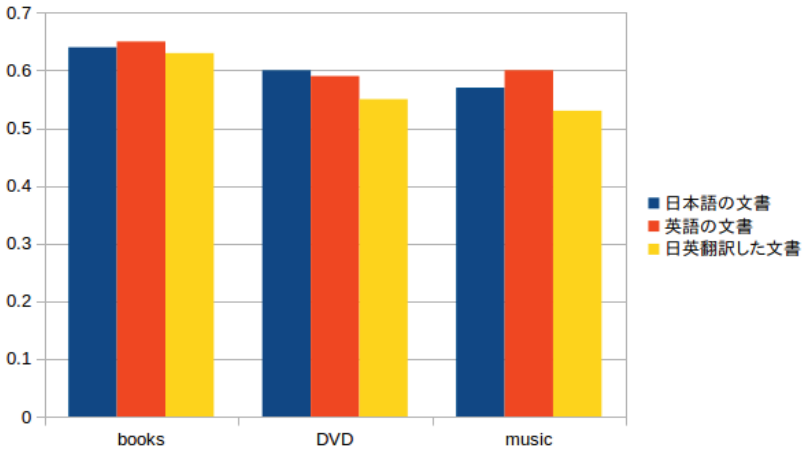


図 3　BERT を用いた感情分析

次に、上記の実験のうち、単語をベクトル化する部分を Word2Vec を用いて行った場合の結果を表 2 に示す (荘司ほか (2019))。

5. 考察

表 1 より、BERT を用いた感情分析では、「テスト文書が英語の場合」に対し「テスト文書が日本語の場合」の精度があまり変わらない。しかし、表 2 より、Word2Vec を用いた感情分

表 2　Word2Vec を用いた感情分析

テスト文書	books	DVD	music
日本語の文書	0.64	0.69	0.70
英語の文書	0.75	0.74	0.76
日英翻訳した文書	0.69	0.70	0.72

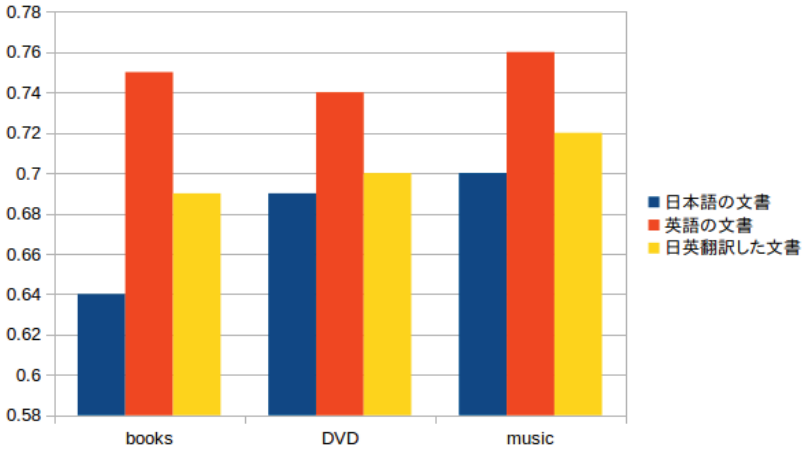


図 4　Word2Vec を用いた感情分析

析では、「テスト文書が英語の場合」に対し「テスト文書が日本語の場合」の精度が 6 10% 低下することがわかる。つまり、BERT を用いた場合、文書ベクトルを他言語へ変換することによる精度の低下が少ないといえる。その理由は、Word2Vec が単語そのものの意味をベクトル化するのに対し、BERT では文脈に沿った単語の意味をベクトル化するからであると推測できる。例えば、”mean”という単語は「意味する」、「卑劣な」、「平均」などの意味を持つが、Word2Vec で表すと全て同じ埋め込み表現になってしまう。さらにそれを変換するため、意味合いが異なる埋め込み表現なり、Word2Vec を用いた場合の精度が低下すると考えられる。

また、BERT を用いた場合、日英翻訳文書に対しての精度と提案手法の精度はほぼ同等である。翻訳は他言語間で感情分析をするもっとも単純な方法であるが、対訳コーパスなどの多くのリソースを必要とする。そのため、精度が同等であるならば、提案手法に優位性があるといえる。

6.　おわりに

本論文では教師データを利用しない感情分析を試みた。英語というメジャーな言語側での分類器の学習、つまりターゲット領域（日本語）の教師データ無しでの分類機の学習である。

実験では、BERT を用いた方法と Word2Vec を用いた方法を比較し、BERT を用いた場合のほうが、「テスト文書が英語の場合」に対する「テスト文書が日本語の場合」の精度の低下が少ないことが分かった。しかし、本論文の実験では、「テスト文書が英語の場合」の精度が

低く、今後上げていくことが出来ればそれに伴って「テスト文書が日本語の場合」の精度も上がっていくと考えられる。今後はその方向で研究を進めていきたい。

文　献

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever (2013). "Exploiting similarities among languages for machine translation." *arXiv preprint arXiv:1309.4168*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep contextualized word representations." *CoRR*, abs/1802.05365.

荘司響之介・新納浩幸・小宮嘉那子 (2019).「Bilingual Word Embeddings によるターゲット言語の教師データを必要としない感情分析」　.　言語処理学会第 25 回年次大会.