文書領域情報を有するBERTの階層位置に関する考察

著者	欧 陽恵子, 田中 裕隆, 曹 鋭, 白 静, 馬 ブン, 新納 浩幸
雑誌名	言語資源活用ワークショップ発表論文集
巻	4
ページ	169-173
発行年	2019
URL	http://doi.org/10.15084/00002566

文書領域情報を有する BERT の階層位置に関する考察

欧陽恵子 (茨城大学大学院理工学研究科情報工学専攻) *
田中裕隆 (茨城大学工学部情報工学科) †
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) ‖

A study of the layer in BERT with domain information of documents

Yanghuizi Ou (Graduate School of Science and Engineering, Ibaraki University)
Hirotaka Tanaka (Department of Computer and Information Sciences, Ibaraki University)
Rui Cao (Graduate School of Science and Engineering, Ibaraki University)
Jing Bai (Graduate School of Science and Engineering, Ibaraki University)
Wen Ma (Graduate School of Science and Engineering, Ibaraki University)
Hiroyuki Shinnou (Graduate School of Science and Engineering, Ibaraki University)

要旨

BERT は Transformer で利用される Multi-head attention を 12 層（あるいは 24 層）積み重ねたモデルである．各層の Multi-head attention は，基本的に，入力単語列に対応する単語埋め込み表現列を出力している．BERT の各層では低層から徐々に何からの情報を取り出しながら，その文脈に応じた単語の埋め込み表現を構築していると考えられる．本論文では領域適応で問題となる領域情報に注目し，BERT の出力の各層が持つ領域情報がどのように推移するのかを考察する．

1. はじめに

BERT (Devlin et al. (2018)) は言語の事前学習モデルであり，多くの自然言語処理システムにおいてその有効性が示されている．BERT の実体は Transformer の encoder で利用される Multi-head attention を 12 層（あるいは 24 層）積み重ねたネットワークである．各層では，入力された単語の埋め込み表現列に対して何らかの変換を行い，変換結果である単語の埋め込み表現列を出力する．この過程は単語の埋め込み表現が，層を経る毎に徐々に文脈に対応

* 19NM705X@vc.ibaraki.ac.jp
† 16T4032N@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

した意味を表す埋め込み表現に変換されると捉えることができる．

本論文では領域適応で問題となる領域に関する情報（領域情報）に注目する．BERT においても領域情報が層を経る毎に徐々に取り出されていると考えられる．そこで本論文では各層が持つ領域情報がどのように推移するのかを考察する．

手順としては単語 w を固定して，各領域 d ごとにその単語が出現する用例を集め，BERT にかける．この処理によって単語 w の BERT における階層 l の出力の埋め込み表現 $e(d,l)$ が得られる．この $e(d,l)$ の分布を領域毎に得て，その分布を比較することで領域情報がどのように推移するのかが確認できる．本論文では領域情報の推移を見るために $e(d,l)$ から領域 d を識別する SVM を学習し，$e(d,l)$ 自身を識別し，その正解率を測る(1)．

実験では単語として「自分」「感じ」「内容」の 3 つを利用した．データは Amazon レビュー文書を用いることで，3 つの領域 (books, DVD, music) を設定した．実験から対象単語ごとに層の位置と領域情報は異なることが確認できた．

2. 関連研究

一般にニューラルネットワークのモデルでは各階層毎に最終のタスクを解くために必要な特徴抽出がなされていると考えられる．例えば，画像識別ではネットワークの下層においては，ターゲットのラベルの種類に依らずに画像の特徴が抽出されており，これを利用して画像識別の転移学習が容易に行える．

言語の事前学習モデルでも実際のタスクでの利用時に，各層が出力する情報の違いが利用されることも多い．例えば ELMo (Peters et al. (2018)) は 2 層の双方向 LSTM を用いているが，語義曖昧性解消や品詞タグ付けでは各層別に利用して評価している．また BERT (Devlin et al. (2018)) においても feature based な利用を行うときには下位層の単語埋め込み表現列も同時に利用した方がよいことが指摘されている．

また我々のグループでは，感情分析の領域適応に対して，BERT の最上位層の出力が必ずしも下層の出力とよりも有用であるとは限らないことを示した (白静ほか (2019))．本研究はこの研究に関連したもので，この論文の主張を確認する目的も有している．

3. 文書領域情報を有する BERT の階層位置

単語の分散表現はその単語に対して一意に定まっているため文脈に応じた意味を表現できていない．その単語がどの領域で利用されていても，その分散表現は同じである．一方，ELMo や BERT などの事前学習モデルは，その文脈に応じた単語の埋め込み表現を出力するために，領域の情報も有していると考えられる．

問題は領域の情報をどのように調べるかである．ここではベクトル表現された単語の意味を領域毎に収集し，領域毎にクラスタを生成させる．そのクラスタがより明確に分類されるほど，領域の情報が埋め込まれていると考える．そして「クラスタがより明確に分類される」という基準には，線形カーネルを用いた SVM における正解率を用いることにする．この正解率

(1) l は最上層の -1 から最下層の -12 までの各層について正解率を測る．

が高いほど，より適切なクラスタが形成されていると考えることができる（図 1 参照）．

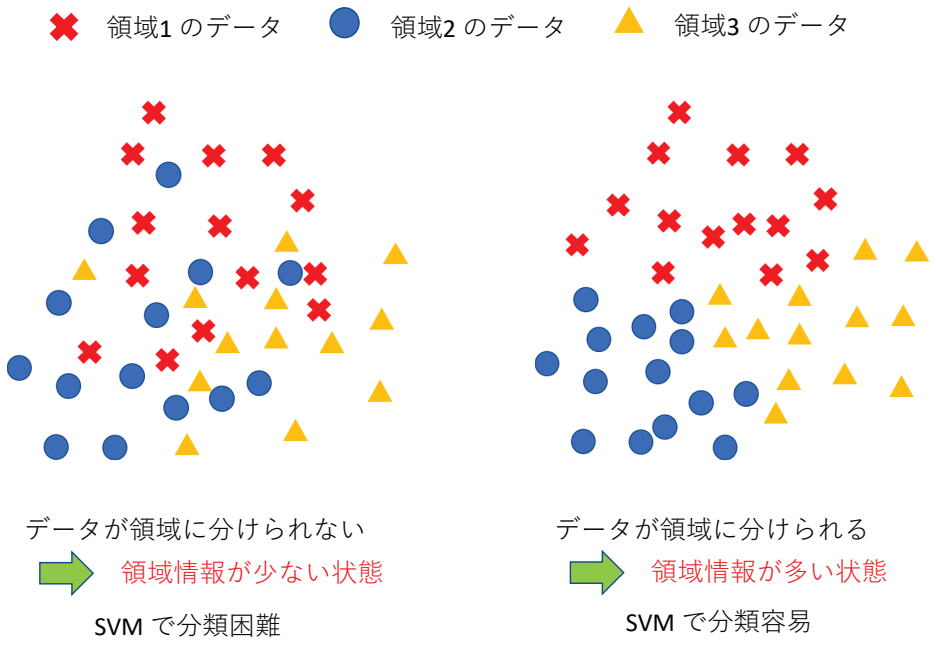

図 1　領域識別の容易性と領域情報の大小の関係

4.　実験

実験では以下のサイトで公開されている Amazon のレビュー文書を利用した．

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

上記データセットは (B) books，(D) DVD，(M) music の 3 つの領域を持つ．それぞれの領域から「自分」「感じ」「内容」の 3 つの単語の用例を取り出した．取り出した用例数を表 1 に示す．1 用例内に対象単語が複数出現することもあるため，実際にそこから取り出す埋め込み表現の数（データ数）は用例数以上になることに注意する．

表 1　対象単語とデータ数

	「自分」			「感じ」			「内容」		
領域	B	D	M	B	D	M	B	D	M
用例数	813	515	400	916	968	1,180	954	544	287
データ数	892	544	424	946	1,005	1,229	981	557	295

各用例を BERT にかけ，対象単語の各層における埋め込み表現を得た．領域の種類をラベルにして，各層ごとに埋め込み表現を収集し，それらから SVM により分類器を作成した．そ

の分類器により学習に利用したデータの識別を行い，正解率を測った．結果を表 2 に示す．

表 2　実験結果

階層	「自分」	「感じ」	「内容」
-1	0.994	0.862	0.997
-2	0.992	0.856	0.998
-3	0.990	0.848	0.997
-4	0.987	0.834	0.995
-5	0.991	0.847	0.997
-6	0.971	0.848	0.996
-7	0.959	0.858	0.993
-8	0.961	0.859	0.988
-9	0.956	0.871	0.981
-10	0.960	0.879	0.976
-11	0.960	0.873	0.974
-12	0.913	0.843	0.935

層の位置に対する正解率の推移を図 2 に示す．

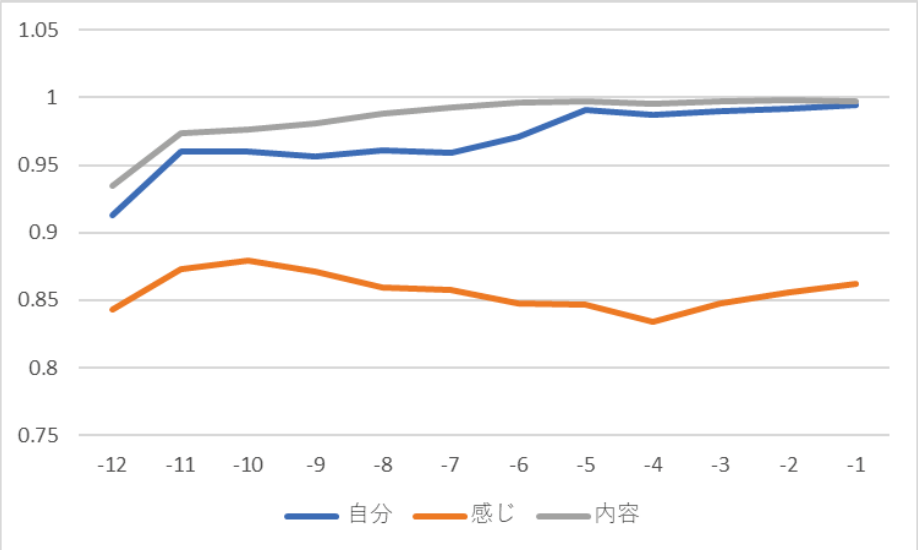

図 2　層位置に対する正解率の変化

表 2 と図 2 から，どの単語に対しても最下層から一つ上の層に移動した段階で，領域を識別するための情報がかなり獲得できていることがわかる．単語「自分」と「内容」に関しては，その後，層が上がるに従い，徐々に領域を識別するための情報が増えている．単語「感じ」に関しては領域を識別するための情報の増減は単純ではない．

5. 考察

実験結果の一般的な解釈は困難である．ただしこの結果から言えることの 1 つは，対象単語ごとに層の位置と得られる領域を識別するための情報（領域情報）の関係は異なるということである．つまり BERT の出力を領域適応に利用する場合には，単純に最上位の層の出力を使えば最適とは限らない．これは論文 (白静ほか (2019)) で指摘された点と一致する．

また本実験で示した領域を識別するための情報が最大の層，つまり実験での正解率が最も高い層が領域適応に利用する最適な層であるとは限らないことに注意したい．例えば感情分析では，データがポジティブがネガティブかを識別することが重要であり，その領域を識別することとは無関係である．領域シフトが生じている場合であっても，ポジネガの識別と領域の識別とは無関係であることに変わりはない，ただし両者は相関があるはずである．通常，領域の識別精度が高いほどデータの表現に情報が多く，ポジネガの識別精度も高くなるはずである．

もう 1 点注意したいのは，領域適応では各領域の共通空間にデータをマップする手法が有効であり，このときマップされたデータには，領域の情報が消えている点である．つまり BERT からの出力を単純に領域適応に利用する場合には，どの層においても領域の情報が十分残っており，BERT からの出力を共通空間にマップされたデータとして扱うことはできない．

BERT の出力を領域適応に有効に利用するには最上位層の出力だけではなく，下位の層の出力も利用すべきだと考えている．具体的な利用法を今後考えていきたい．

6. おわりに

本論文では領域適応で問題となる領域情報に注目し，BERT の出力の各層が持つ領域情報がどのように推移するのかを考察するために，単語を固定し，各領域から用例を集め，BERT にかけ，各層におけるその単語の埋め込み表現を得た．この埋め込み表現から領域をラベルとして SVM を構築し，その埋め込み表現自身の識別を行うことで，各層の持つ領域情報を調べた．実験では単語として「自分」「感じ」「内容」の 3 つを利用した．実験から対象単語ごとに層の位置と領域情報は異なることが確認できた．今後は BERT の最上位層の出力だけでなく，下位の層も利用した BERT の利用法を考えたい．

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*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). “Deep Contextualized Word Representations.” *NAACL-2018*, pp. 2227–2237.

白静・田中裕隆・曹鋭・馬ブン・新納浩幸 (2019).「BERT の下位階層の単語埋め込み表現列を用いた感情分析の教師なし領域適応」　情報処理学会研究報告自然言語処理（NL），2019-NL-240:17, pp. 1–6.