BERTed-BCCWJ: 多層文脈化単語埋め込み情報を付与した『現代日本語書き言葉均衡コーパス』データ

浅原 正幸 *　　加藤 祥
国立国語研究所　国立国語研究所

1. はじめに

文中の単語に対し、その生起する文脈に基づいて異なるベクトルを付与する技術、文脈化単語埋め込みが提案されている。文脈化単語埋め込みモデルの構築・利用には、大規模なテキストコーパスと高性能な GPU/TPU サーバが求められ、個人が文脈化単語埋め込みをコーパスに付与することは難しい。そこで、『国語研日本語ウェブコーパス』(NWJC) (Asahara et al. 2014) から訓練した、事前学習済み BERT (Devlin et al. 2019) モデル NWJC-BERT (浅原ほか 2020) による多層文脈化単語埋め込み情報を、『現代日本語書き言葉均衡コーパス』(BCCWJ) (Maekawa et al. 2014) に対して付与した BERTed-BCCWJ を構築した。工学側のメリットとして、BERT の中間層のベクトルの特徴と、BCCWJ に付与された各種アノテーションとを対照させることで、転移学習に有効であった中間層のベクトルの成分がどのような言語現象に関連付けられるかを知ることができる。言語学側のメリットとして、多義語に対しても語義の近さや連続性が評価できるほか、換喩・提喩などを含めた比喩表現に見られる語義の転換を定量的に評価できる。また、脳科学の分野では、言語を刺激として被験者実験を行う場合、単語埋め込み技術が刺激文のベクトル化に用いられ、脳活動データとの対照比較が可能になる。多種多様な言語情報がアノテーションされた、代表性を有するコーパスに多層文脈化単語埋め込み情報を付与し、共有化することにより、工学・言語学・脳科学をつなぐ共通の研究基盤を提供することになる。

著者が参画している新学術領域「時間生成学」では、現在・過去・未来の表現を呈示した場合に、どの脳部位が活動するかを明らかにすることを目的として研究を進めている。本稿では BERTed-BCCWJ の構築方法を示すとともに、BCCWJ-NWJC (加藤ほか 2019) に付与された分類語彙表番号に基づいて、現在・過去・未来の言語情報がベクトル空間上にどのように表現されるか検討する。

2. BERTed-BCCWJ: 構築方法

本節では、文脈化単語埋め込み情報を BCCWJ に付与した手法について示す。まず、文脈化単語埋め込み情報付与に用いた NWJC-BERT について紹介する。事前学習モデルとして BERT を NWJC 12.8 億文から訓練した。本研究では UniDic の語彙素表記に基づいて訓練する。BERT の訓練時の語彙 (vocab.txt) は、通常、分かち書きされたコーパスの表層形の出現順位に基づき選択するが、UniDic の機能語全て 154 語彙素表記と UniDic-分類語彙表番号対応表 (WLSP2UniDic) に出現する 48,790 語彙素表記（機能語と重複あり）と制御語 5 種の合計 48,914 語とした。これにより、UniDic の 54.6% (468,460/872,831 エントリ) を被覆するほか、分類語彙表番号による語彙の評価が行うことができる。NWJC-BERT の詳細は浅原ほか (2020) を参照されたい。

NWJC-BERT のモデルに基づいて、BERT の extract_features.py を基に、BCCWJ コアデータの語彙素に対して、単語単位に BERT の 12 層のベクトルを付与した。また、文単位にも [CLS] のベクトルを割り当てた。なお、vocab.txt に登録されていない語彙素 ([UNK] 相当 19.18% 242,575/1,264,926)

* masayu-a@ninjal.ac.jp

表 1　BCCWJ-WLSP アノテーション例

短単位書字形	語彙素番号	分類番号	類	部門	中項目	分類項目
「			-	-	–	—-
いま	2460	1.1641	1:体	1:関係	16:関係-時間	1641:関係-時間-現在
は	29321		-	-	–	—-
8	29802	1.1960	1:体	1:関係	19:関係-量	1960:関係-量-数記号（一二三）
割	41343	1.1962	1:体	1:関係	19:関係-量	1962:関係-量-助数接辞
の	28989		-	-	–	—-
人	31500	1.2000	1:体	2:主体	20:主体-人間	2000:主体-人間-人間
が	7889		-	-	–	—-
残りご飯	255971	1.1931	1:体	1:関係	19:関係-量	1931:関係-量-過不足
を	41407		-	-	–	—-
冷蔵	40606	1.3842	1:体	3:活動	38:活動-事業	3842:活動-事業-炊事・調理
・			-	-	–	—-
冷凍	40617	1.3842	1:体	3:活動	38:活動-事業	3842:活動-事業-炊事・調理
し	19537	2.3430	2:用	3:活動	34:活動-行為	3430:活動-行為-行為・活動
て	24874		-	-	–	—-
しまい	15738	2.1503	2:用	1:関係	15:関係-作用	1503:関係-作用-終了・中止・停止

にもベクトルが割り当てられるが、全体の 0.35% (4,553/1,264,926) にあたる全角空白などの一部の記号やアスキーアートに対してはベクトルを割り当てていない。また、BCCWJ 非コアデータ全体については、最終層 (-1 層) のベクトルのみを付与した。

3.　BCCWJ-WLSP との対照分析

本節では、BCCWJ に対して分類語彙表番号を付与したデータ BCCWJ-WLSP を用いて、BERTed-BCCWJ の単語埋め込みを評価する。BCCWJ-WLSP は、表 1 のように BCCWJ の一部 347,094 語に対して、語義の曖昧性を解消しながら人手で分類語彙表の分類番号を付与したものである。この BCCWJ-WLSP と BERTed-BCCWJ を重ね合わせ、BCCWJ-WLSP の分類項目により認定される、現在 (.1641 関係-時間-現在) 650 語・過去 (.1642 関係-時間-過去)563 語・未来 (.1643 関係-時間-未来) 253 語、合わせて 1466 語のベクトルを抽出した。BERT の 12 層 (-1 層〜-12 層) のベクトルを t-SNE 法に基づき 2 次元に写像して可視化したものを図 1 に示す。図における現在（青）・過去（緑）・未来（赤）の分布を見ると、最下層の -12 層のレベルで、まとまったいくつかのクラスタを構成していることがうかがえる。

次に、-1 層に対して表層形を割り当てたものを図 2 に示す。同図において分類語彙表における現在・過去・未来が混在する個所について注目する。図中 A の四角の中には多義語「今度」が含まれ、B の四角の中には、多義語「先」が含まれる。(1)-(4) に例を示す。(x,y) は tSNE により 2 次元上に写像した図中の座標である。

(1) しかしどうもまだ気がのらないようで，しばらく漫然と机を眺めていたが，【今度】は机のいちばん下の引き出しを開けた。中から何枚か，ＢｅＯＳのＣＤ-ＲＯＭが出てくる。
現在 ：PM25_00084 (x,y)=(5.33, 22.79)

(2) 【 今度 】いつポケモンカードくれるの？
未来：PB11_00006 (4.06, 20.61)

(3) 厚生労働省が【先】に公表した医療制度改革試案に対し、本間正明・阪大教授ら四人の同会議民間議員が意見書を提出し
過去：PN1e_00002 (3.81, 0.49)

(4) メーカーにマシンを返すのは【先】なんで後処理のことは悩まなくていいんですが。
未来：PM25_00084 (-3.78, -0.66)

一方、図 2 中 C の四角の中には、非常に近い距離で現在と未来が配置されている。(5)-(7) に例を示す。BCCWJ-WLSP は国語研短単位に対して語義情報を付与しているが、固有名詞の一部など、特定の時間を指示しない表現が多くみられた。

図 1　BERTed-BCCWJ における現在（青）・過去（緑）・未来（赤）の表現：左上 -1 層〜右下 -12 層

(5) Ｔｈｅ　Ｃｌｉｎｉｃａｌ　【Ｃｕｒｒｅｎｔ】加齢黄斑変性症の新しい治療法
現在：PM26_00004 (-29.44, -18.45)

(6) だが「彼女の穴を埋める存在は見当たらない」（ＵＳＡ【トゥデー】紙）のが実情で、
現在：PN2e_00004 (-29.24, -19.21)

(7) この【トゥモロー】ランドのサテンワンピースは、胸元の切り替えとハイウエストマークで脚長に見えます。
未来：PM51_00077 (-29.32, -18.97)

なお、現在を表す表現「今（いま）」は 232 例確認され、複数のクラスタに分類された。(8) の表現は連体修飾の用例と近接し、(9) の表現は「今では」「今でも」といった用例と近接し、(10) の表現は「今国会」「今年度」などの接頭辞の用例と近接する傾向がみられた。このように高頻度の語においては、その出現する文法形式によりクラスタが分けられる傾向が確認された。

(8)【今】の高校生をどう見ていますか。
現在：PN3e_00006 (-34.30,7.33)

(9)【今】では古風なものとしてやはりやや低い目途にしか使われておらぬのは、
現在：PB48_00016 (-24.30,11.48)

(10) 米国は【今】協議を通じ、北朝鮮に対し各国とともにあらゆる手段でその態度変更を促す構えだ。
現在：PN4d_00004 (-14.91, 20.89)

4.　おわりに

本稿では、NWJC で訓練した BERT モデルによる多層文脈化単語埋め込みを付与した、 BCCWJ データについて紹介した。分類語彙表の関係-時間-{ 現在, 過去, 未来 } の語のベクトルを評価し、多義語について異なるベクトルを付与することを確認した。今後、言語刺激と脳活動データを NWJC-BERT を介して対照比較することで、時間知覚に関連する脳部位を特定する研究を進める。

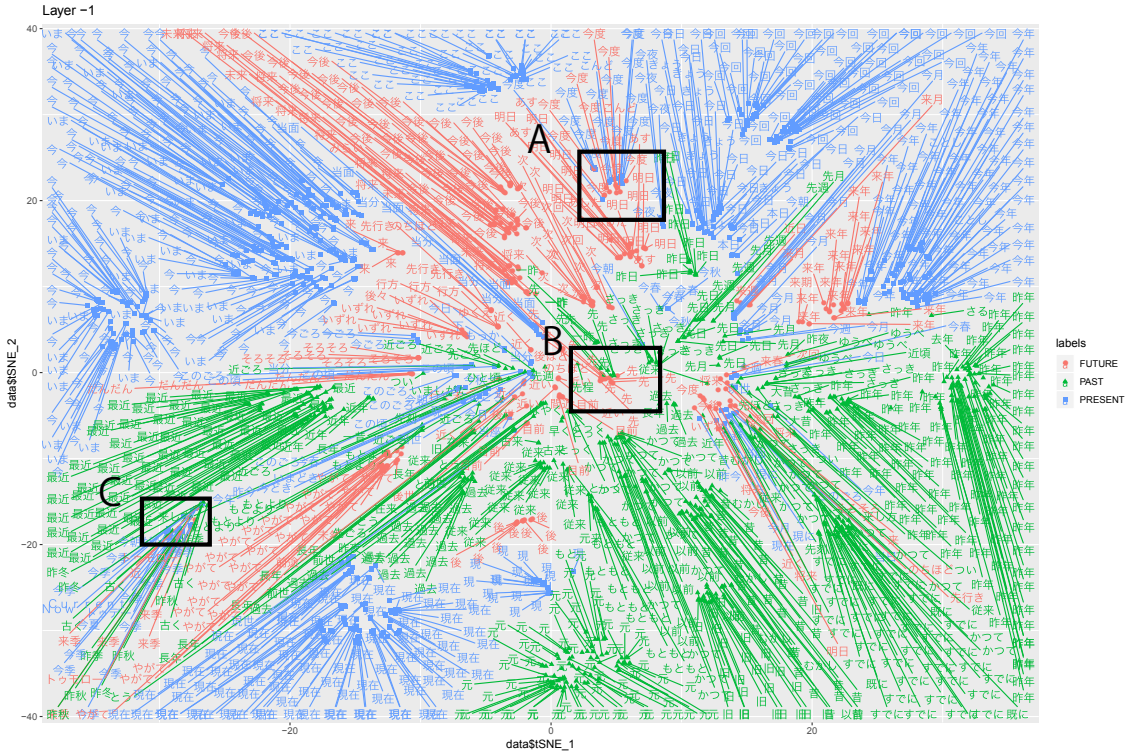

図 2　BERTed-BCCWJ における時間表現の分布：-1 層（表層形つき）

同データは、GPU サーバを有しない言語学の研究者にとっても価値がある。従来の単語埋め込みのモデルでは、異なる語義間の距離を測ることが困難であったが、BCCWJ の用例とアノテーションを介して、定量的に評価できるようになった。適切なクラスタリング技術を使うことによって、語義の数や新語義の出現を推定できるほか、BCCWJ に基づく指標比喩データベースを分析して、提喩・換喩を含む比喩と基本義との距離を定量的に評価できる可能性がある。

謝　辞

本研究は国立国語研究所コーパス開発センター共同研究プロジェクトおよび科研費 JP17H00917, JP18H05521, JP18K18519 によるものです。

文　献

Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi (2014). “Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan.” *Alexandria: The Journal of National and International Library and Information Issues*, 25:1–2, pp. 129–148.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *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.

浅原正幸・西内沙恵・加藤祥 (2020).「NWJC-BERT: 多義語に対するヒトと文脈化単語埋め込みの類似性判断の対照分析」言語処理学会第 26 回年次大会発表論文集.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*, 48, pp. 345–371.

加藤祥・浅原正幸・山崎誠 (2019).「分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ」　日本語の研究, 15:2, pp. 134–141.