# 白色化が単語埋め込みに及ぼす効果の検証

佐々木翔大 [1,2]　Benjamin Heinzerling[1,2]　鈴木潤 [2,1]　乾健太郎 [2,1]
[1] 理化学研究所　[2] 東北大学
{shota.sasaki.yv, benjamin.heinzerling}@riken.jp
{jun.suzuki, kentaro.inui}@tohoku.ac.jp

## 概要

深層ニューラルネットワークに基づく自然言語処理では必須の要素となる静的あるいは文脈化単語埋め込みには，空間的な偏り（異方性）が存在し，その能力を十分にいかせていないことが実証されている．この異方性を緩和する方法として，白色化がある．白色化は標準的な線形変換であるが，実験的に性能が向上することが報告されている．しかし，性能向上の理由は自明ではなく，理論的あるいは実験的な分析が望まれる．本研究では，白色化が単語埋め込みに及ぼす意味的な影響を明らかにするための実験を行う．実験結果から，白色化は，静的単語埋め込みに対しては単語頻度バイアス除去効果を示し，文脈化単語埋め込みに対しては単語頻度バイアス除去効果以外の効果を持つことが示唆された．[1)]

## 1　はじめに

静的単語埋め込み（Static Word Embeddings, SWE）[1, 2]，および文脈化単語埋め込み（Contextualized Word Embeddings, CWE）[3, 4, 5] は，現代の自然言語処理システムにおいて必須の基盤技術である．このような埋め込みを作成する目的は，単語，フレーズ，文の “意味” を捉えた表現を計算することである．しかしながら，ジェンダーバイアス [6]，ソーシャルバイアス [7] などの学習データに固有のバイアスも反映してしまうことが報告されている．SWE については，高頻度単語が特定の方向に沿って集中する**単語頻度バイアス**が存在することが先行研究により示されている [8]．このような単語ベクトルの非一様な角度分布（**異方性**）は，埋め込み空間の非効率的な利用につながる．また，単語頻度バイアスの介在する埋め込み空間においては，頻度の高い単語は，意味が似ていないにもかかわらず，類似のベクトルで表現されることになる．このことから一般に，単語埋め込み空間の異方性は改善すべきであると考えられている．

単語埋め込み空間の異方性の影響を軽減するために，いくつかの等方化手法が提案されている．本研究では，最もシンプルな等方化手法として注目されている**白色化**に焦点を当てる．白色化は空間的に相関のある（異方的な）ベクトル集合を，相関のない（等方的な）ベクトル集合に変換する線形変換である．白色化は標準的なデータ変換技術であるが，特に CWE などの単語埋め込みへ適用する研究は最近になって登場した [9, 10]．これらの先行研究では，CWE を対象とした等法化手法の中で白色化が他の手法よりも優れていることを報告している．しかしながら，白色化の欠点は，様々な種類のバイアスや埋め込みの意味的特性への影響が十分に調査されていない点である．本論文では，白色化が単語埋め込みに及ぼす意味的な影響に関する初期分析を行う．

事前分析では，白色化の効果は単語頻度バイアス除去の効果を含んでいることが示された．そこで，白色化の効果が単語頻度バイアス除去のみであるかどうかを本論文の Research Question とする．白色化の効果をより明確化するために，単語頻度バイアス除去のみを行う手法を利用する．具体的には，埋め込みにおける単語頻度バイアスを除去することのみに着目した再構築に基づく単語頻度バイアス除去手法 (RFD) を提案し，白色化と RFD の挙動を比較する．その結果，SWE では単語頻度バイアスが，CWE では単語頻度バイアスに加え，それ以外のバイアスが除去されることが示唆された．

## 2　背景

### 2.1　SWE と CWE における異方性

単語埋め込みにおける異方性に関してはこれまで多くの議論がなされてきた．Mu ら [8] は，SWE の

1)　本研究で用いたコードは以下の URL で公開予定である：https://github.com/losyer/whitening_effect

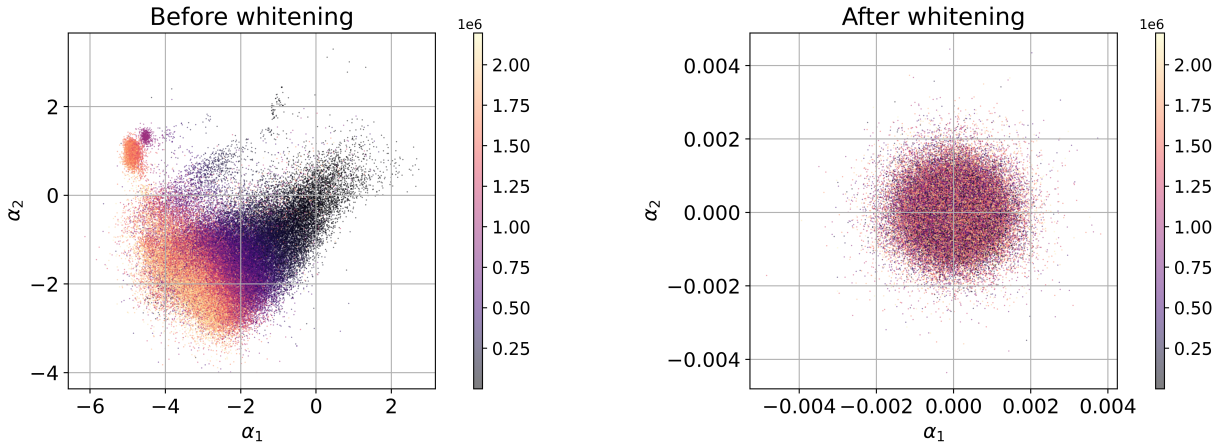



図 1: GloVe の単語埋め込みに白色化を適用する前（左図）と後（右図）の第 1，第 2 主成分 ($\alpha_1$．$\alpha_2$) の可視化．各点が単語に対応し，点の色が単語の頻度を表す．黒は高頻度，黄色は低頻度であることを表す．

第一主成分と第二主成分に単語の頻度情報が埋め込まれており，SWE の異方性の原因であると報告した．Li ら [11] は，BERT [4] の単語埋め込み層のベクトルに単語頻度のバイアスがあることを実証的に示した．また，Liang ら [12] は，単語頻度の対数と単語ベクトルのノルム，平均コサイン類似度の間に相関があることを報告した．頻度のバイアスに加え，CWE における外れ値の次元も最近注目されている．Luo ら [13]，Kovaleva ら [14] は，BERT と RoBERTa [15] の埋め込みにおいて，他の次元よりも有意に高い値を持つ次元を特定し，埋め込みにおける異方性の原因であることを示唆した．

## 2.2 白色化による等方化

白色化とは，ベクトル集合を，共分散行列が恒等行列であるベクトル集合に変換する線形変換である．共分散行列が恒等行列であるということは，変換によって各次元が無相関化され，分散が 1 になることを意味する．定義から，白色化された単語埋め込みは，より等方的になるといえる．

一般に機械学習において，白色化は学習データのバイアスを軽減する目的で，特徴量ベクトルの集合に適用されている [16, 17]．バイアスを軽減することで，深層学習モデルが高品質な表現を学習し，モデルの収束が早まることが報告されている．白色化は任意のベクトル集合に適用できる汎用的なアルゴリズムであるため，CWE で得られた文ベクトルにも適用することが可能である．Huang ら [9] は CWE の異方性の問題に対処するために，CWE に白色化を適用し，CWE の性能を向上したことを報告した．白色化は数学的には明確に定義された変換であるが，SWE や CWE に適用したときに，白色化によってどのような情報，どのように変換されるかを明らかではない．本研究では，白色化が SWE と CWE に及ぼす効果を明らかにすることを目的する．

# 3 事前分析

本節では，白色化の単語埋め込みに及ぼす効果を探るための予備的な分析を行う．Mu ら [8] は，GloVe [2] と Word2Vec [1] の単語埋め込み行列を対象に主成分分析を行い，その第 1，第 2 主成分が単語の頻度と相関があることを観察し，単語頻度バイアスの存在を示唆した．ここで Mu らと同様に単語埋め込み行列の主成分分析を，白色化の適用前/後の埋め込みを対象に行う．図 1 にその結果を示す．白色化適用前は，単語の頻度と主成分方向の成分の値に相関がある，つまり単語埋め込みに単語の頻度バイアスが存在することを示唆しており，これは Mu ら [8] の報告と一致する．しかしながら，白色化適用後の埋め込みの主成分には，単語の頻度バイアス観察できなかった．したがって，白色化には単語の頻度バイアスを軽減する効果がある事がわかった．

分析結果を受けて，本論文では白色化の効果をより明確化することを目的とする．具体的には，白色化が単語頻度バイアスの除去と等価であるのか，それ以外の効果を有するかを確認することが本研究の Research Question である．実験では，白色化と次の 4 節で導入する単語頻度バイアスの除去手法を同時にモデルに適用する実験を行う．白色化と単語頻度バイアス除去の両者を同時に適用した時，それぞれの効果が独立していれば，それぞれの効果による性能向上が期待できるが，両者の効果が重複していれ

ば，性能向上は限定的となることが予想される．

## 4 再構築に基づく単語頻度バイアス除去手法

本節では，元の単語埋め込みの品質に影響を与えず，単語頻度バイアスを除去することのみに焦点を当てた手法を導入する．敵対的学習を通してバイアス除去を達成する基本方針は Gong ら [18] の手法と同様である．単語埋め込みは，対象の単語が高頻度クラスに属するか低頻度クラスに属するかを識別しようとする識別器を欺くためにチューニングされる．Gong らが目的タスクに関する損失を定義した代わりに，本手法では GloVe や BERT などの事前学習された単語埋め込みの性質をできるだけ保持することを目的とした再構築損失を導入する．以降，本手法を単語埋め込みの再構築に基づく単語頻度バイアス除去手法 (Reconstruction-based frequency debiasing, RFD) と呼ぶ．

はじめに SWE に対する学習手順を説明する．語彙中の単語集合を $\mathcal{W}$，$\boldsymbol{e}(w)$ を事前学習された単語 $w$ の単語埋め込み，$\boldsymbol{v}(w;\theta^{\mathrm{emb}})$ を学習対象の単語埋め込みとする．ここで $\theta^{\mathrm{emb}} \in \mathbb{R}^{d\times V}$ は単語埋め込みの重み行列である．また，$d$ は単語埋め込みの次元数，$V$ (= $|\mathcal{W}|$) は語彙サイズである．SWE の再構築損失は以下のように定義される．

$$L_{R_{\mathrm{swe}}}(\mathcal{W};\theta^{\mathrm{emb}}) = \sum_{w\in\mathcal{W}} \left\|\boldsymbol{e}(w) - \boldsymbol{v}(w;\theta^{\mathrm{emb}})\right\|_2^2. \quad (1)$$

次に，[18] らに従って識別器損失を定義する．はじめに，語彙 $\mathcal{W}$ を高頻度単語集合 $\mathcal{W}_{\mathrm{pop}}$ と低頻度単語集合 $\mathcal{W}_{\mathrm{rare}}$ に 2 分割する．$\mathcal{W}_{\mathrm{pop}}$ は頻度の上位 $t$% の単語で構成され，$\mathcal{W}_{\mathrm{rare}} = \mathcal{W} \setminus \mathcal{W}_{\mathrm{pop}}$ とする．また $f_{\theta^D}$ を単語の頻度クラスを出力する 2 値分類を行う識別器とする．識別器損失は以下のように定義される．

$$\begin{aligned} L_{D_{\mathrm{swe}}}(\mathcal{W};\theta^D,\theta^{\mathrm{emb}}) = {} & \frac{1}{|\mathcal{W}_{\mathrm{pop}}|}\sum_{w\in\mathcal{W}_{\mathrm{pop}}} \log A \\ & + \frac{1}{|\mathcal{W}_{\mathrm{rare}}|}\sum_{w\in\mathcal{W}_{\mathrm{rare}}} \log(1-A), \end{aligned} \quad (2)$$

ここで $A = f_{\theta^D}(\boldsymbol{v}(w;\theta^{\mathrm{emb}}))$ とした．最後にパラメータ $\theta^{\mathrm{emb}}$ と $\theta^D$ を敵対的学習の手順で最適化する．

$$\arg\min_{\theta^{\mathrm{emb}}} \arg\max_{\theta^D} L_{R_{\mathrm{swe}}}(\mathcal{W};\theta^{\mathrm{emb}}) - \lambda L_{D_{\mathrm{swe}}}(\mathcal{W};\theta^D,\theta^{\mathrm{emb}}), \quad (3)$$

ここで $\lambda$ はハイパーパラメータである．CWE を対象にした際の定式化は付録 A.1 に記述する．

## 5 実験

**データセット**　単語埋め込みの品質を評価するために，Semantic Textual Similarity (STS) タスクを採用する．具体的には，公正な比較ための STS の標準的データセットである STS Benchmark (STS-B) データセット [19] を使用する[2)3)]．このデータセットは文のペアとその間の類似度スコアを人手で付与したもので構成されている．類似度スコアは 0 から 5 の範囲である．

**評価**　先行研究 [9, 20] に従い，正解類似度スコアとモデルによって算出される文類似度の間のスピアマン順位相関で評価する．モデルが出力する文間の類似度としては，文ベクトル同士のコサイン類似度を採用する．

**実験設定**　SWE として 2 種類の GloVe 埋め込み[4)]（`GloVe840B`, `GloVe6B`）と Google News Embeddings[5)](`GNews`) を使用する．`GloVe840B` は 8400 億トークンを含む Common Crawl データセットで，GloVe6B は 60 億トークンを含む Wikipedia と Gigaward データセットで学習されている．`GNews` は CBOW アルゴリズム [21] を用いて，1000 億トークンを含む Google ニュースデータセットで学習されている．CWE としては Huggingface Transformer Library [22] の `BERT-base` [4], `DistilBERT-base` [23], `RoBERTa-base` [15] モデルを使用する．

実験では，SWE と CWE それぞれについて，以下の 4 つの設定を比較する．

1. **Vanilla モデル**：後処理を適用しない生のモデル
2. **白色化モデル (`-wh`)**：白色化を適用したモデル
3. **単語頻度バイアス除去モデル (`-Fdeb`)**：単語頻度バイアス除去手法を適用したモデル
4. **単語頻度バイアス除去・白色化モデル (`-Fdeb-wh`)**：頻度バイアス除去手法を適用した後に白色化を適用したモデル

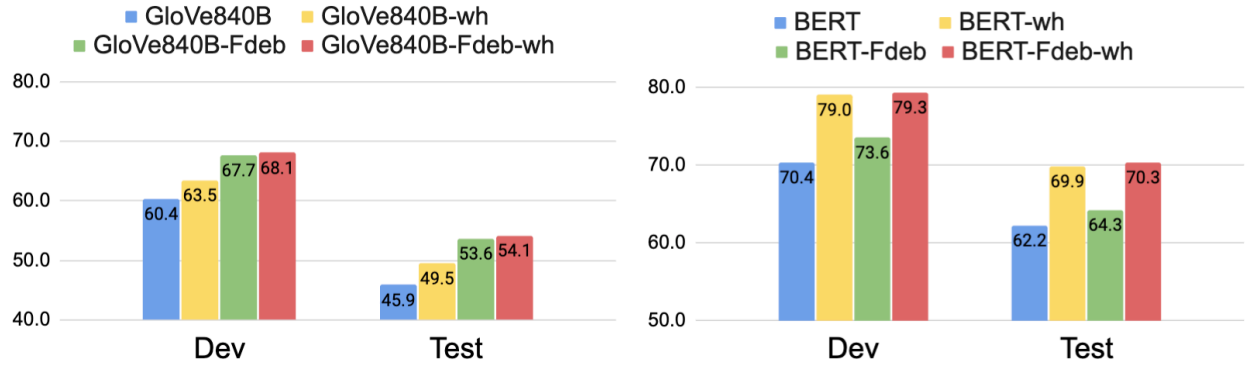
単語頻度バイアス除去手法には，4 節で導入した `RFD` を用いる．

2) STS タスクを用いて評価する際，シェアドタスクで年ごとに提供されたデータセット（例えば STS-14 など）が利用される場合がある．これが原因で訓練，開発，評価セットの区分に関して統一的な規範が存在せず，公平な性能比較が難しかった．Cer らは [19] はこうした現状を踏まえて，過去に提供されたデータセットから質の良いデータを選択し，訓練，開発，評価セットの公式区分を STS-B として提供した．

3) https://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark

4) https://nlp.stanford.edu/projects/glove/

5) https://code.google.com/archive/p/word2vec/




(a) GloVe840B

(b) BERT

図 2: STS-B の開発セット（Dev）と評価セット（Test）における実験結果．モデル性能は正解スコアとモデルの予測した類似度の間のスピアマン順位相関係数によって評価した．

SWE モデルについては，文中の全単語の単語埋め込みを平均することで文ベクトルを算出する．CWE モデルについては，Huang ら [9] に従い，第 1 隠れ層と最終隠れ層の単語埋め込みを平均する．その他の詳細な実験設定は付録 A.2 に示す．

**実験結果**　GloVe840B と BERT モデルの性能を図 2 に示す．その他のモデルの結果は付録 A の図 4，5 に示す．実験結果から以下のことが観察された．

（i） SWE，CWE のいずれの実験でも，白色化により性能が向上した．特に，CWE モデルの性能が大きく向上した．

（ii） GNews を除く全てのモデルにおいて，RFD，つまり単語頻度バイアス除去の効果による性能向上が確認された．特に，GloVe840B-Fdeb は評価セットにおいて GloVe840B よりも 7.7 ポイント性能が向上した．

（iii） SWE の-Fdeb-wh モデルは-Fdeb モデルに対して大きな改善は見られなかった．

（iv） SWE とは異なり，CWE の-Fdeb-wh モデルでは-Fdeb モデルに対して大きな改善 が見られた．

観測（i）は，先行研究 [9] の結果と一貫している．観測 (ii) について，GloVe などの SWE モデルに対する単語頻度バイアス除去による性能向上幅は CWE モデルに対するそれよりも高かった．これは図 3 から示唆されるように，GloVe の方が単語頻度バイアスが強いためであると推測される．

観測 (iii)，(iv) に関して，もし，白色化と単語頻度バイアス除去の効果が独立しているのであれば，両者をモデルに適用した際，それぞれの効果による性能向上が期待できるはずである．しかしながら SWE については，-Fdeb と-Fdeb-wh の間に有意な差は見られなかった（観測 (iii)）．このことから，SWE に対する白色化の効果は，単語頻度バイアス除去の効果と概ね等価であるか，両者の間に大きな重複があることが明らかになった．一方，観測 (iv) から，CWE では，SWE にはない CWE に固有のバイアスの補正など，単語頻度バイアス除去とは別の効果があることが示唆された．考えられる効果の一つとして，Luo ら [13]，Kovaleva ら [14] が報告した外れ値問題を補正する効果があると推測しているが，この点についてはさらなる調査が必要である．

## 6　おわりに

本研究では，等方化手法として昨今注目を集めている白色化に焦点を当て，白色化が単語埋め込みに及ぼす意味的な影響に関する初期分析を行った．特に，SWE と CWE に白色化を適用した際の単語頻度バイアスの変化を分析した．

事前分析では，白色化の効果は部分的に単語頻度バイアス除去の効果を含んでいることを示した．次に，単語頻度バイアスを除去することのみに特化した単語頻度バイアス除去手法 (RFD) を提案し，白色化と RFD の挙動を比較することで白色化の効果をより明確にすることを試みた．実験結果から，白色化は，静的単語埋め込みに対しては単語頻度バイアスの除去効果を示し，文脈化単語埋め込みに対しては単語頻度バイアス除去効果以外の効果を持つことが示唆された．

## 謝辞

本研究は JST CREST JPMJCR20D2，JSPS 科研費 JP21H04901，JP21K17814，JST ムーンショット型研究開発事業 JPMJMS2011 (fundamental research) の助成を受けて実施されたものである．

## 参考文献

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In **Proceedings of NIPS**, pp. 3111–3119, 2013.

[2] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In **Proceedings of EMNLP**, pp. 1532–1543, 2014.

[3] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In **Proceedings of NAACL**, pp. 2227–2237, 2018.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of NAACL**, pp. 4171–4186, 2019.

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[6] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In **Proceedings of NAACL**, pp. 629–634, 2019.

[7] Masahiro Kaneko and Danushka Bollegala. Unmasking the mask – evaluating social biases in masked language models. In **Proceedings of AAAI**, p. 13, 2022.

[8] Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In **Proceedings of ICLR**, p. 25, 2018.

[9] Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. WhiteningBERT: An easy unsupervised sentence embedding approach. In **Findings of EMNLP**, pp. 238–244, 2021.

[10] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. **arXiv preprint arXiv:2103.15316**, 2021.

[11] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In **Proceedings of EMNLP**, pp. 9119–9130, 2020.

[12] Yuxin Liang, Rui Cao, Jie Zheng, Jie Ren, and Ling Gao. Learning to remove: Towards isotropic pre-trained bert embedding. In **Artificial Neural Networks and Machine Learning – ICANN 2021**, pp. 448–459. Springer-Verlag, 2021.

[13] Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. Positional artefacts propagate through masked language model embeddings. In **Proceedings of ACL-IJCNLP**, pp. 5312–5327, 2021.

[14] Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. BERT busters: Outlier dimensions that disrupt transformers. In **Findings of ACL-IJCNLP**, pp. 3392–3405, 2021.

[15] 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. **arXiv preprint arXiv:1907.11692**, 2019.

[16] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, **Proceedings of International Conference on Artificial Intelligence and Statistics**, Vol. 15, pp. 215–223. PMLR, 2011.

[17] Marc' AurelioRanzato, Alex Krizhevsky, Geoffrey Hinton. Factored 3-way restricted boltzmann machines for modeling natural images. In Yee Whye Teh and Mike Titterington, editors, **Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics**, Proceedings of Machine Learning Research, pp. 621–628. PMLR, 2010.

[18] Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. FRAGE: Frequency-agnostic word representation. In **Proceedings of NIPS**, Vol. 31, p. 12 pages. Curran Associates, Inc., 2018.

[19] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In **Proceedings of SemEval**, pp. 1–14, 2017.

[20] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **Proceedings of EMNLP-IJCNLP**, pp. 3982–3992, 2019.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

[22] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of EMNLP**, pp. 38–45, 2020.

[23] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. **arXiv preprint arXiv:1910.01108**, 2019.

# A 付録

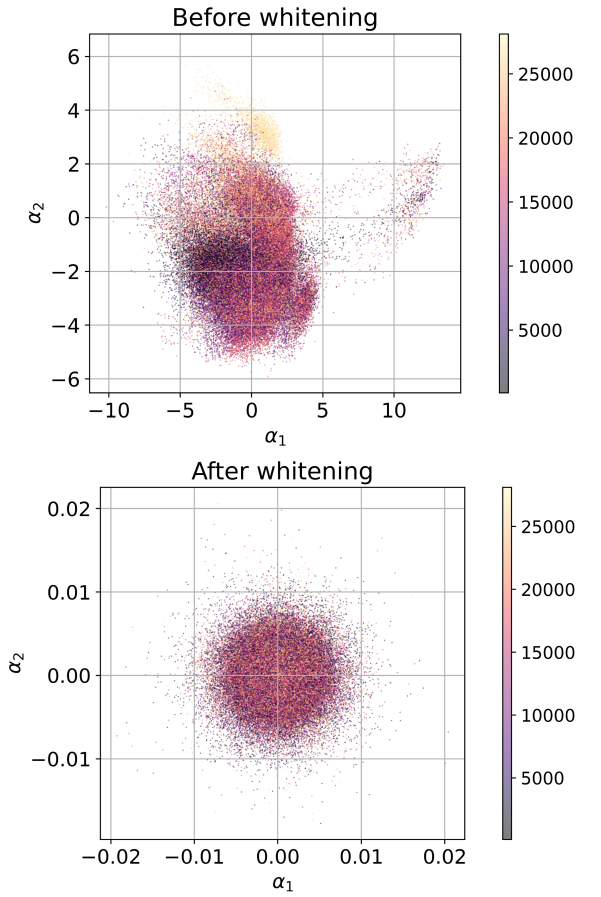



図 3: BERT の単語埋め込みに白色化を適用する前と後の第 1, 第 2 主成分 ($\alpha_1$, $\alpha_2$) の可視化.

## A.1 CWE を対象にした際の RFD

訓練コーパス $\mathscr{C}$ を用意し，$\mathscr{C}$ 内の文 $s$ のエンコードした際の隠れ層を対象に最適化を行う．ここで $\mathscr{W}_s$ を文 $s$ 内の単語（もしくはサブワード）の集合とし，$\mathscr{L}$ を CWE モデル中の層から選択されたのターゲット層の集合とする．また $s$ をエンコードした際の単語 $w$ の $l$ 番目の層のベクトルを $\boldsymbol{e}^l(w,s)$ とする．$\boldsymbol{v}^l(w,s;\theta^{\text{emb}})$ は $\boldsymbol{e}^l(w,s)$ と同様であるが，学習対象のパラメータ $\theta^{\text{emb}}$ を持つ学習対象の CWE モデルから得る埋め込みベクトルとする．CWE のための再構築損失と識別器損失は以下のように定義される．

$$L_{R_{\text{cwe}}}(\mathscr{C};\theta^{\text{emb}}) = \sum_{s\in\mathscr{C}}\sum_{w\in\mathscr{W}_s}\sum_{l\in\mathscr{L}} \left\|\boldsymbol{e}^l(w,s) - \boldsymbol{v}^l(w,s;\theta^{\text{emb}})\right\|_2^2, \quad (4)$$

$$L_{D_{\text{cwe}}}(\mathscr{C};\theta^D,\theta^{\text{emb}}) = \sum_{s\in\mathscr{C}}\sum_{w\in\mathscr{W}_s}\sum_{l\in\mathscr{L}} L'_{D_{\text{cwe}}}(w,l;\theta^D,\theta^{\text{emb}}), \quad (5)$$

$$L'_{D_{\text{cwe}}}(w,l;\theta^D,\theta^{\text{emb}}) = \frac{1}{|\mathscr{W}_{s,\text{pop}}|}\sum_{w\in\mathscr{W}_{s,\text{pop}}} \log B + \frac{1}{|\mathscr{W}_{s,\text{rare}}|}\sum_{w\in\mathscr{W}_{s,\text{rare}}} \log(1-B), \quad (6)$$

ここで $B = f_{\theta^D}(\boldsymbol{v}^l(w,s;\theta^{\text{emb}}))$ とした．ただし $\mathscr{W}_{s,\text{pop}} = \mathscr{W}_s \cap \mathscr{W}_{\text{pop}}$ and $\mathscr{W}_{s,\text{rare}} = \mathscr{W}_s \setminus \mathscr{W}_{s,\text{pop}}$ である．

目的関数は以下のように定義される．

$$\arg\min_{\theta^{\text{emb}}} \arg\max_{\theta^D} L_{R_{\text{cwe}}}(\mathscr{C};\theta^{\text{emb}}) - \lambda L_{D_{\text{cwe}}}(\mathscr{C};\theta^D,\theta^{\text{emb}}). \quad (7)$$

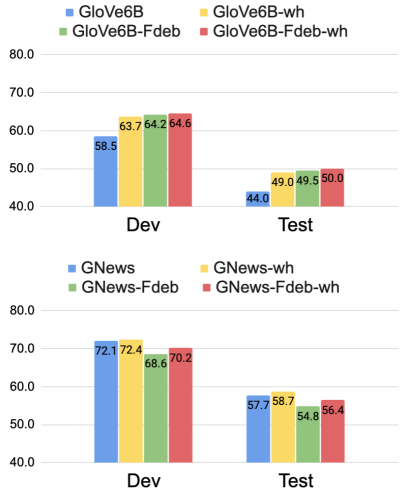



図 4: STS-B における SWE に関する実験結果（スピアマン順位相関係数 $\rho \times 100$）.

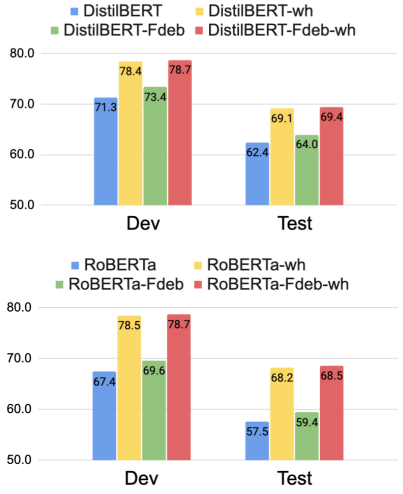



図 5: STS-B における CWE に関する実験結果（スピアマン順位相関係数 $\rho \times 100$）.

## A.2 実験設定の詳細

式 4, 5 のターゲット層集合としては，様々な選択肢がある．例えば，ターゲット層をモデル内部の全ての層とすることも可能である．本研究では，第 1 層と最終層を含むターゲット層集合の中で最も簡易なもの，つまり，BERT と RoBERTa では $\mathscr{L} = [1, 12]$, DistilBERT では $\mathscr{L} = [1, 6]$ を選択した．これは，第 1 層と最終層の単語埋め込みを平均すると最も良い性能を達すると報告した Huang ら [9] の研究に由来する．訓練コーパス $\mathscr{C}$ としては，STS データセットの文集合を用いた．

高頻度単語集合の閾値としては $t = 10$ を用いた．予備実験においてバッチサイズに変更を加えても性能に影響がなかったため，バッチサイズは 128 に固定した．式 3 と式 7 の $\lambda$ は $[0.02, 0.1]$ から探索した．これは，Gong ら [18] の実装[6] におけるデフォルト推奨設定が $\lambda = 0.02$ である一方で，彼らの論文内では $\lambda = 0.1$ を使用した旨が記載されていることに由来する．また，学習率は $[1e-3, 5e-3, 1e-2]$ から，学習エポック数は $[1, 3, 5, 10, 20, 30, 40, 50]$ から探索した．全ての探索は STS-B の開発セットで行った．

6) https://github.com/ChengyueGongR/Frequency-Agnostic