BERT を用いた文埋め込みモデルの単語の暗黙的な重み付け

栗田宙人 [1]　小林悟郎 [1,2]　横井祥 [1,2]　乾健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所
{hiroto.kurita, goro.koba}@dc.tohoku.ac.jp {yokoi, kentaro.inui}@tohoku.ac.jp

概要

事前学習済みマスク言語モデルに追加学習を加えた文埋め込みモデルが続々と提案されており，幅広い後段タスクで高い性能を達成している．不思議な点として，以前主流であった静的単語埋め込みを用いた手法では単語を陽に重み付けるという工夫が肝要であったにも関わらず，マスク言語モデルを用いた文埋め込みはこの工夫抜きで既存手法を凌駕している．本研究では，BERT を用いた文埋め込みモデルたちが各単語を単語頻度に基づく情報量 $-\log P(w)$ で暗黙的に重み付けていることを，特徴寄与測定手法 Integrated Grad との経験的な相関を通して明らかにする．さらに，この重み付け傾向は事前学習済みの BERT を文埋め込みに適するよう追加学習する過程で強まっていることを報告する．

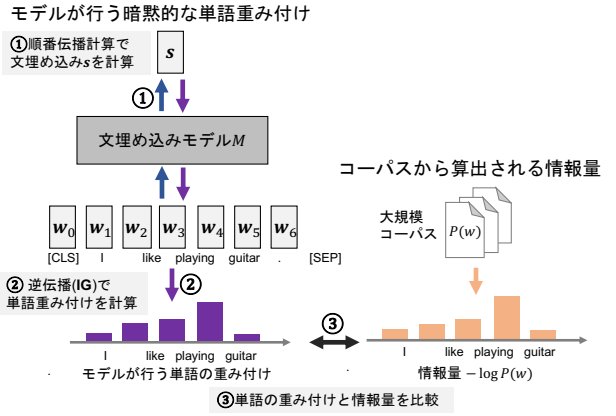

図 1　本研究の概念図．モデルが文埋め込みを計算する際の各単語の寄与を逆伝播計算により求め，求めた単語の寄与と大規模コーパスから計算された単語の情報量の比較を行う．

1　はじめに

自然言語文をベクトル表現で表した文埋め込みは文の意味的類似度の計算，情報検索，文書分類など，自然言語処理分野で幅広く使用される有用な道具である．代表的な文埋め込み生成技術として，文を構成する各単語の静的単語埋め込みの平均を取る手法がある程度効果的であることが経験的・理論的に知られている [1, 2, 3]．このとき，単純な平均ではなく，ストップワードの除去や TF-IDF[4] に代表される単語の逆頻度に基づく重み付けが大きな効果をもたらすことが知られている．

一方，近年では動的単語埋め込みを生成するマスク言語モデルである BERT[5] や RoBERTa[6] を追加学習した文埋め込みモデルが続々と提案され，意味的類似度計算などの後段タスクで静的単語埋め込みを用いる手法を凌駕している [7, 8]．これらのモデルでは，静的単語埋め込みを用いた手法で効果的だった明示的な重み付けは基本的に行われない．しかし，マスク言語モデルを基にしているため，内部の複雑な非線形ネットワークにより単語が混ぜ合わせられ，暗黙的に単語を重み付けながら文埋め込みを生成していると考えられる．成功を収めている最近の文埋め込みモデルはどのような重み付けを学んでいるのだろうか．

本研究では BERT を用いた文埋め込みモデルは内部で単語頻度に基づく情報量 $-\log P(w)$ に比例した暗黙的な単語重み付けを行なっていることを明らかにする．また，この重み付け傾向は BERT の追加学習を通じて強まっていることわかった．すなわち，BERT を用いた文埋め込みモデルは既存手法でポストホックに外から行われていた重み付けを追加学習の過程で自ら獲得していることがわかる．

2　分析の方針

本稿では BERT を用いた文埋め込みモデルが内部で暗黙的に行う単語の重み付けと頻度に基づく単語の情報量とを比較する（図 1）．モデルが行う重み付けは勾配を用いた特徴量帰属手法 Integrated Gradients を用いて定量化し，単語の情報量はコーパス上で算出された単語頻度を元に計算する．

2.1 準備：BERT を用いた文埋め込み

モデルは，入力文 $s = (w_1, w_2, ..., w_{|s|})$ の各単語を単語埋め込み層にて単語埋め込みに変換し（$w_i \mapsto \boldsymbol{w}_i \in \mathbb{R}^d$），この単語埋め込みをまとめたベクトル列 $\boldsymbol{W} = [\boldsymbol{w}_1, .., \boldsymbol{w}_{|s|}] \in \mathbb{R}^{|s| \times d}$ を残りのモデル M に入力して文埋め込み $\boldsymbol{s}$ を計算する（$M: \boldsymbol{W} \mapsto \boldsymbol{s} \in \mathbb{R}^d$）．ここで，$d$ は埋め込みの次元数を表す．なおモデル M の最後尾で文ベクトル $\boldsymbol{s}$ を作るためにおこなわれる最後の処理は，各トークンの隠れ状態のプーリングによるまとめ上げであるが，このプーリングには文に含まれる単語の表現の平均プーリング（MEAN）と文頭に挿入される特殊トークン [CLS] の隠れ状態を用いる CLS プーリング（CLS）の 2 種類が主に使われている．

2.2 モデルが行う単語の暗黙的な重み付け

モデルが文埋め込みを生成する際に内部でどのように各単語を重み付けているかを定量化するために，勾配計算を用いた特徴量（ここでは単語）の帰属手法である Integrated Gradients（IG）[9] を用いた．特徴量帰属手法は他にも様々あるが，IG は複雑なニューラルネットに適用でき，さらに単純な微分を用いる手法の問題点が積分によって克服されている，近年もっとも標準的に採用されているアプローチである [10, 11]．

今回の問題では，文中の単語 w_i がニューラルネットで文埋め込み $\boldsymbol{s}$ を構成する際にどの程度寄与しているかの度合い $c(w_i, \boldsymbol{s})$ を，単語ベクトルの各要素の文ベクトルの各要素への寄与度 $\mathrm{IG}(\boldsymbol{w}_i[j], \boldsymbol{s}[k]; M)$ に帰着させて以下のように計算することができる [12]．

$$c(w_i, \boldsymbol{s}) \coloneqq \sqrt{\sum_j \sum_k \mathrm{IG}(\boldsymbol{w}_i[j], \boldsymbol{s}[k]; M)} \tag{1}$$

$$\mathrm{IG}(\boldsymbol{w}_i[j], \boldsymbol{s}[k]; M) \coloneqq (\boldsymbol{W}_i^{(k)} - \boldsymbol{B}_i^{(k)}) \times \int_{\alpha=0}^{1} \frac{\partial M(\boldsymbol{B} + \alpha \times (\boldsymbol{W} - \boldsymbol{B}))^{(j)}}{\partial \boldsymbol{W}_i^{(k)}} d\alpha \tag{2}$$

ただし，$\boldsymbol{v}[j]$ はベクトル v の j 次元目の値．$\boldsymbol{B} = [\boldsymbol{b}_1, ..., \boldsymbol{b}_{|s|}]$ は積分の「ベースラインの位置」を表すベクトル列で，文頭と文末に挿入される特殊トークン以外の単語を [PAD] に置き換えた [CLS],[PAD],..,[PAD],[SEP] に対応するベクトル列を用いた．IG はモデルへの入力をベースライン $\boldsymbol{B}$ から実際の入力 $\boldsymbol{W}$ まで徐々に変化させながら対象特徴量の出力への偏微分を計算していき，それらを積分することで（経路全体での寄与を足し合わせることで）合計の寄与を計算する．式 1 では文ベクトル側と単語ベクトル側の要素のすべての組み合わせについての寄与の二乗和を計算することで単語の文への寄与を算出している．

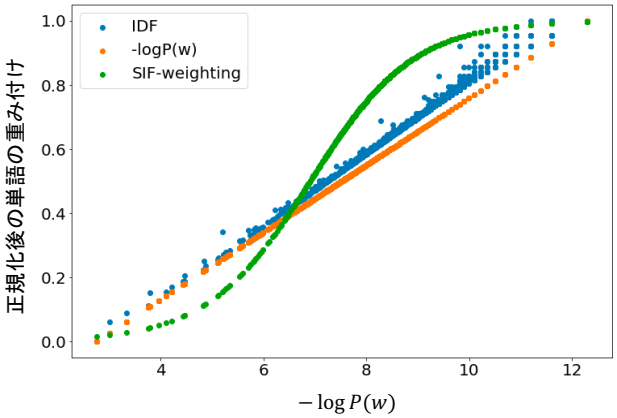

図 2 単語頻度に基づく情報量 $-\log P(w)$ と各種重み付け手法の比較．各種重み付けは STS-Benchmark[13] の頻度情報を基に作成した．

2.3 単語の持つ情報量

本研究では，頻度に基づく単語の情報量 $-\log P(w)$ を文埋め込みモデル内部で行われる重み付けと比較する．情報量とは確率 $P(x)$ で発生する事象 x を観測した時に得られる情報の量であり，$-\log P(x)$ で定義される．文（単語列）を受け取ってその意味を扱う文埋め込みモデルでは，$P(\cdot)$ を単語頻度分布，$P(w)$ を単語 w が出現する確率とすれば，$-\log P(w)$ は単語 w を観測した時に得られる情報の量に相当し，低頻度語ほど情報量は大きい．

情報量 $-\log P(w)$ は既存の単語重み付け手法と深い関係を持つ．文埋め込みにおいては，ストップワードの除去や TF-IDF，SIF weighting など，単語の逆頻度に基づく重み付けが効果的であると知られている [4, 14]．これらの手法は文の意味を理解する上で不必要な高頻度語を取り除いたり，文の意味を決定づけると期待される低頻度語を重視しているとみなすことができ，情報量 $-\log P(w)$ の概念と一貫する．情報量 $-\log P(w)$ は逆頻度に基づく最も基本的な重み付けと言え，実際これらの既存の重み付け手法とも類似している（図 2）．これらの理由から，文埋め込みモデル内部で行われる各単語への重み付けの比較対象として情報量 $-\log P(w)$ を用いる．

3 実験

BERT を用いた文埋め込みモデル内部での各単語への重み付けを単語頻度に基づく情報量 $-\log P(w)$ に照らし合わせながら調査する．

モデル BERT を追加学習した文埋め込みモデルには SentenceBERT（SBERT）[7] と SimCSE[8] を用いた．SimCSE は追加学習時にラベル付きデータを使っていない Unsupervised SimCSE と，使っている Supervised SimCSE の 2 種類を扱う．また，ベースラインとして文埋め込み用の追加学習を行う前の BERT-base (uncased) を用いた．特に，平均プーリングを採用した BERT (MEAN) と CLS プーリングを採用した BERT (CLS) の 2 種類を扱う．なお，SimCSE および SBERT はそれぞれ BERT (CLS) および BERT (MEAN) を追加学習したモデルである．

データセット モデルへの入力文には，文類似度計算タスクの代表的なデータセットである STS Benchmark[13] の検証データおよびテストデータを用いた．今回は簡単のため，サブワード分割が発生せず，文長 7 の合計 617 文のみを対象とした．また，単語の情報量 $-\log P(w)$ の計算に用いる単語頻度は BERT の事前学習時コーパスを Wikipedia と BooksCorpurs[15] から再現して算出した．

3.1 定量分析

モデルに各文を入力し，2.2 節で述べた方法でモデル内部における各単語への重み付けを計算する．続いて，計算された各単語への重み付けと頻度に基づく単語の情報量 $-\log P(w)$ の関係を比較するため両者のピアソン相関係数を測定する．また，ベースラインとして単語頻度 $P(w)$ とのピアソン相関係数も測定する．

BERT を用いた文埋め込みモデルは $-\log P(w)$ に従って各単語を重み付けている：表 1 より，文埋め込み用の追加学習を行っていない BERT (CLS) および BERT (MEAN) における単語への重み付けは $-\log P(w)$ と弱い正の相関があった．つまり，追加学習前の BERT は緩やかに情報量に基づいた単語重み付けを行なっている事が分かった．追加学習後のモデルである SBERT および SimCSE では追加学習前の BERT (MEAN/CLS) よりも強い正の相関が見られ，追加学習後のモデルはより情報量に従って単語を重み付けていることが分かった．これらの結果は，文埋め込み用の追加学習によってモデルが暗黙的に単語の情報量に基づく重み付けを獲得していることを示唆する．ただし，同程度の情報量を持つ単語でもモデルの重み付けには分散が見られ，特に高情報量（低頻度語）側で分散が大きくなっていた（図 3）．このことから，$-\log P(w)$ だけでは文埋め込みモデルが行う重み付けを完全には説明できていないと考えられる．例えば，同じ単語でも出現によって異なる重み付けがされており，周囲の単語との関係性などにも依存して重み付けを変えている可能性がある．高情報量での分散傾向の調査は今後の展望の一つである．

表 1 各モデルにおける単語の貢献度とのピアソンの相関係数

モデル（プーリング方法）	$-\log P(w)$	$P(w)$
BERT (CLS)	0.38	-0.26
→ Unsupervised SimCSE (CLS)	0.80	-0.63
→ Supervised SimCSE (CLS)	0.59	-0.37
BERT (MEAN)	0.51	-0.49
→ SBERT (MEAN)	0.72	-0.50

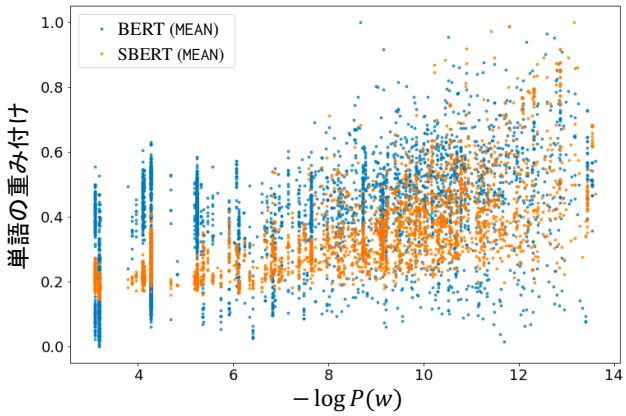

図 3 BERT (MEAN) と SBERT における $-\log P(w)$ と単語の重み付けの散布図．他のモデルの散布図は付録 A 参照．

重み付けは $P(w)$ ではなく $-\log P(w)$ に従う：表 1 において $-\log P(w)$ との相関係数と $P(w)$ との相関係数の絶対値を比較すると，どのモデルにおいても一貫して $-\log P(w)$ との相関の方が強い．つまり，モデル内部での単語の重み付けは単純な単語頻度 $P(w)$ というよりも，log をかけた情報量 $-\log P(w)$ に従っている．

3.2 定性分析

各モデルにおける単語への重み付けの具体例を図 4 に示す．“*a woman is cutting an onion* .” を入力した例では，追加学習前の BERT (CLS/MEAN) はどの単

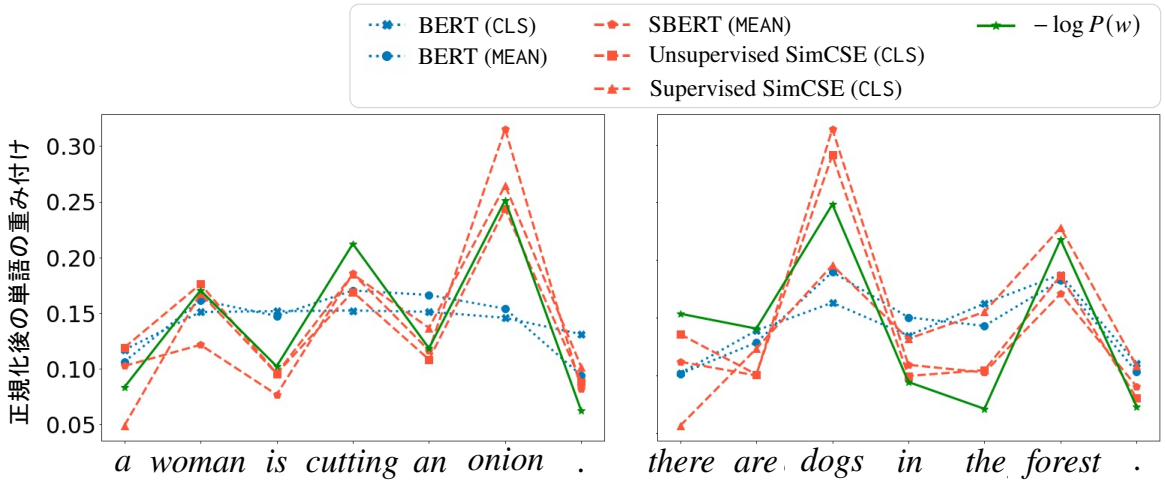

図 4　文ごとの単語の貢献度と $-\log P(w)$.

語にもほぼ均等に重み付けを行なっているが，追加学習後のモデルはよりピーキーな重み付けを行っており，$-\log P(w)$ と非常に似ていることがわかる．SBERT の重み付けが特にピーキーで，文中で情報量が最も高い *onion* を最も高く重み付けている．

“*there are dogs in the forest* .” を入力した例でも同様の傾向が見られ，情報量が最も高い *dogs* に SBERT が最も高い重み付けを行なっていた．これらのことから，BERT を追加学習したモデルは $-\log P(w)$ に近い形で重み付けを行うが，モデルによって重み付け傾向がやや異なることが示唆される．

4　関連研究

静的単語埋め込みを用いた文埋め込み　静的単語埋め込みが持つ加法構成性により，単純な単語埋め込みの平均を用いても質の高い文埋め込みを構成できることが知られている [1, 2, 3]．さらに，各単語埋め込みを単純に平均する代わりに，単語の逆頻度に基づいて（TF-IDF や SIF-weighting など）重み付け和することで，文埋め込みの質が向上する [14, 4]．本研究では，文埋め込み構築において効果的であると知られるこれらの重み付け手法と関係深い単語の情報量 $-\log P(w)$ を比較対象に用いた．

マスク言語モデルを用いた文埋め込み　近年，自然言語処理分野の様々なタスクで成功を納めているマスク言語モデル [5, 6] を追加学習した文埋め込みモデルが盛んに開発されている．Reimers ら [7] は BERT を自然言語推論タスクを用いて追加学習させることで文埋め込みに適した SBERT を提案した．また，最近では画像分野の表現学習において対照学習が成功を収めている（SimCLR[16] など）ことに影響を受け，DeCLUTR[17], SimCSE[8], など，対照学習を取り入れた文埋め込み手法が続々と提案されている．これらのマスク言語モデルを用いた文埋め込みモデルでは，既存手法のように単語を明示的に重み付けることは基本的にない．例外としては，Wang ら [18] は SBERT の各層から取り出した単語埋め込みを明示的に重み付けて文埋め込みを構築することでより性能を向上させる手法を提案している．本研究では，マスク言語モデルを用いた文埋め込み手法の中でも代表的である SBERT と SimCSE を対象とし，モデル内部で暗黙的に行われている単語の重み付けについて調査した．

5　おわりに

本稿では，BERT を用いた文埋め込みモデルが内部では単語の情報量 $-\log P(w)$ に基づいて単語を重み付けていることを明らかにした．また，この重み付け傾向は BERT の追加学習を通じて獲得されていることも示し，BERT を用いた文埋め込みモデルは既存手法でポストホックに外から行われていた重み付けを追加学習の過程で自ら獲得していることが示唆される．

今後は，モデルが行う単語重み付けと文脈付きの情報量や単語の品詞などを比較し，$-\log P(w)$ 単体では捉えきれなかったモデルが「暗黙的に行う重み付け」についてさらに調査を進める．また，他の文埋め込みモデルや情報検索などで使用されているテキスト埋め込みモデルにまで分析の範囲を拡大し，様々な学習がモデルの単語重み付けに与える影響を分析する方向性も興味深い．

謝辞

本研究は JSPS 科研費 JP22H05106，JP22J21492，JST，CREST，JPMJCR20D2 の助成を受けたものです．また，本研究に関して多くのアドバイスを下さった栗林樹生氏をはじめ Tohoku NLP グループの皆様に感謝致します．

参考文献

[1] Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. **Cognitive science**, Vol. 34, No. 8, pp. 1388–1429, 2010.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, **Advances in Neural Information Processing Systems**, Vol. 26. Curran Associates, Inc., 2013.

[3] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. In **International Conference on Learning Representations**, 2016.

[4] Ignacio Arroyo-Fernández, Carlos Francisco Méndez-Cruz, Gerardo Sierra, Juan Manuel Torres-Moreno, and Grigori Sidorov. Unsupervised sentence representations as word information series: Revisiting tf–idf. **Computer Speech and Language**, pp. 107–129, July 2019.

[5] 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.

[6] 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, 2019.

[7] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[8] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[9] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In **Proceedings of the 34th International Conference on Machine Learning - Volume 70**, ICML'17, p. 3319–3328. JMLR.org, 2017.

[10] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[11] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. Vol. 35, pp. 12963–12971, May 2021.

[12] Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In **International Conference on Learning Representations**, 2020.

[13] 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 the 11th International Workshop on Semantic Evaluation (SemEval-2017)**, pp. 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[14] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In **International Conference on Learning Representations**, 2017.

[15] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In **2015 IEEE International Conference on Computer Vision (ICCV)**, pp. 19–27, 2015.

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In **Proceedings of the 37th International Conference on Machine Learning**, ICML'20. JMLR.org, 2020.

[17] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 879–895, Online, August 2021. Association for Computational Linguistics.

[18] Bin Wang and C.-C. Jay Kuo. Sbert-wk: A sentence embedding method by dissecting bert-based word models. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Vol. 28, pp. 2146–2157, 2020.

A $-\log P(w)$ と単語の重み付けの散布図

$-\log P(w)$ と各モデルが行う単語の重み付けの散布図を示す.

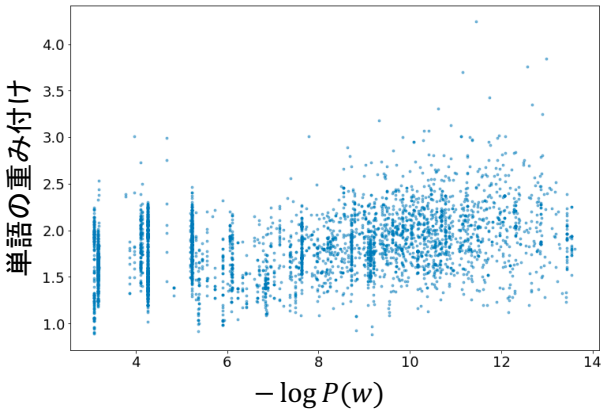

図 5 BERT(CLS) における $-\log P(w)$ と単語の重み付けの散布図

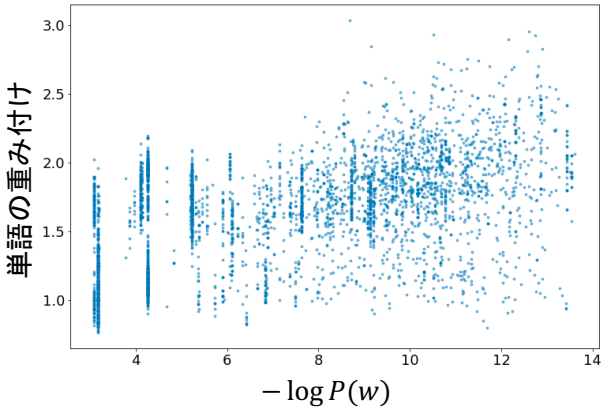

図 6 BERT(MEAN) における $-\log P(w)$ と単語の重み付けの散布図

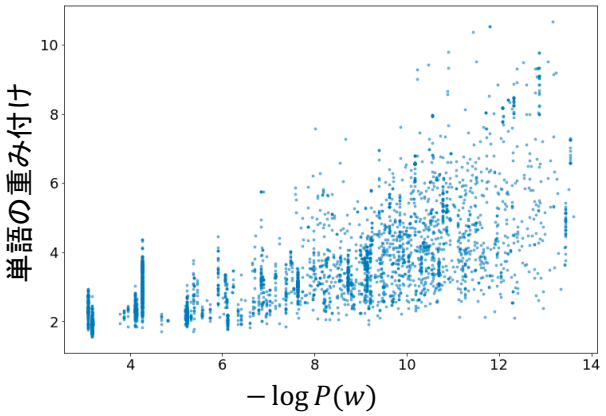

図 7 SBERT(MEAN) における $-\log P(w)$ と単語の重み付けの散布図

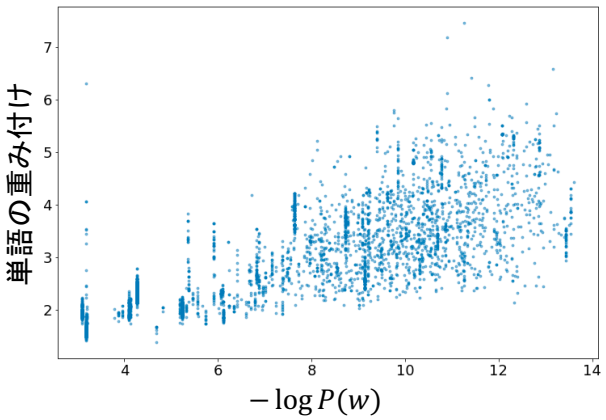

図 8 Unsupervised SimCSE(CLS) における $-\log P(w)$ と単語の重み付けの散布図

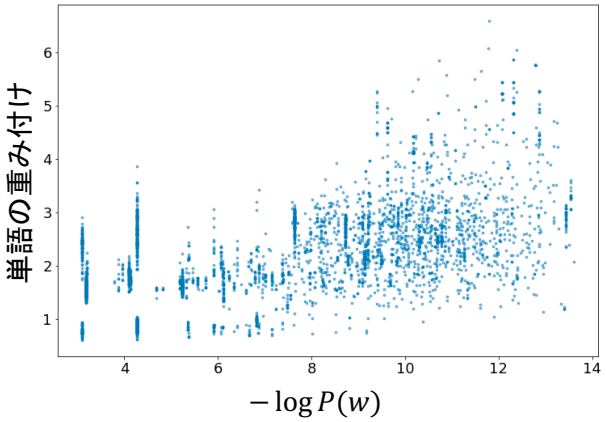

図 9 Supervised SimCSE(CLS) における $-\log P(w)$ と単語の重み付けの散布図