高品質な単語ベクトル取得のための言語モデルの層選択

上野和樹 [1] 永田亮 [1,2]
[1] 甲南大学 [2] 理化学研究所
s2371019@s.konan-u.ac.jp nagata-nlp2026 @ ml.hyogo-u.ac.jp.

概要

文脈付き単語ベクトルは言語研究に盛んに用いられているものの，言語モデルのどの層の出力を単語ベクトルとすべきかの明確な基準が存在しない．本稿では，最適な層を選択するための汎用的な手法を提案する．その実現のため，meaning-frequency law [1] という経験則を利用する．実験により，(1) 最適な層は最終層などとする従来の知見は必ずしも正しくはなく，(2) 提案手法は，常に最適な層を選択できるわけではないが，汎用的に層選択が行える，(3) そのようにして選択した層は，最大性能からの性能低下を抑えることに有効である，ことを示す．

1 はじめに

言語モデルから得られる文脈付き単語ベクトル（以降，単に単語ベクトルと省略）は単語の用法や語義に関する言語分析に盛んに用いられている（例：文献 [2, 3, 4, 5]）．単語ベクトルは言語データのみから学習され，語義，用法など単語の属性を反映しており，言語研究に有益である．特に，対象単語の前後両方の文脈情報を考慮できるマスク言語モデルの単語ベクトルの利用が盛んである．

深層学習に基づいた言語モデルは複数の層から成るが，どの層の出力を単語ベクトルとすべきかについて明確な基準が存在しない．上述を含む多くの既存研究では，言語モデルの最終層の出力，全層の出力の平均などを慣習的に用いることが多い．また，Ethayarajh [6] は，最終層の直前の層（以降，前最終層と省略）の出力が，文脈を最もよく識別すると報告している．しかしながら，この結果は語義識別能力を直接評価したものではない．単語ベクトルを得るための最適な層を簡便かつ効果的に知ることができれば，より良い単語ベクトルの利用につながる．

そこで，本稿では，最適な層を選択するための汎用的な手法を提案する．ここで汎用的とは，層選択に必要となるのは生コーパスのみであり，性能を見積もるための検証データなど追加の情報は必要ないことを意味する．この特性を実現するため，提案手法は，言語学で知られる Zipf の meaning-frequency law [1] に基づく．同法則は，高頻度な単語ほど語義数が多いという経験則である．形式的には，単語頻度を f，単語の語義の豊富さを v としたとき，

$$\log(v) = \delta \log(f) + c \tag{1}$$

が成り立つという法則である．文献 [7] で，傾き δ と言語モデルの最終層から得られる単語ベクトルの語義識別能力には高い相関がみられることが示されている．この知見に従い，単語ベクトルの語義識別能力の指標として傾き δ の値を使うというのが提案手法の基本方針である．

本研究の貢献は次の 3 点である：(1) 単語ベクトルを得るための最適な層は，言語モデルと対象言語に応じて大きく異なることを示す（すなわち，従来の知見が正しくないことがある）；(2) Meaning-frequency law を利用することで，汎用的な層選択を実現する；(3) 実コーパスを用いてその有効性を示す．

2 関連研究

マスク言語モデルから得られる単語ベクトルを言語研究に利用することが盛んに行われている．代表例に，語義の通時的な変化を取り扱う意味変化検出（例：文献 [2, 3, 4]）がある．また，語義／用法変化に関する仮説の検証 [5] や文法化度の定量化 [8] など応用範囲も広がりつつある．

一方で，どの層の出力を単語ベクトルとして利用すべきかの明確な基準はない．慣習的に，最終層の出力を単語ベクトルとして利用することが多い（上述のうち，文献 [2, 5, 8] が該当）．また，全層の平均，最終 4 層の平均なども用いられる [2, 3, 4]）．

関連して，単語ベクトルの文脈識別度を報告した Ethayarajh [6] の研究がある．同報告では，ある単語

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

タイプにおいて，ランダムに選択された二つの単語の使用事例に対して単語ベクトル間の余弦類似度を測定し，文脈識別度としている．最終層に近いほど文脈識別度は高くなると報告されている．また，言語研究での利用が多いマスク言語モデル（BERT）では，前最終層[1)]の文脈識別度が最も高いとされている．ただし，文脈識別度は関連が深いものの直接語義識別能力を評価したものではない．

別の関連研究として，単語ベクトルに基づいた meaning-frequency law の検証 [7] がある．この研究では，単語ベクトルの方向のばらつきに基づいて語義の豊富さ v を定量化する．単語ベクトルの方向は，周辺単語すなわち文脈により決定される．そのようにして求めた語義の豊富さ v を式（1）に回帰して得られる傾き δ は，語義識別能力に強く相関することを報告している．また，モデルサイズが小さい場合や未学習の言語では式（1）からの乖離が大きくなることも示している．

3 最適層の選択手法

提案手法は，meaning-frequency law の検証を通じて得られる式（1）の傾き δ に基づく．具体的な手順は，次のとおりである．

与えられたコーパスを文単位で言語モデルに入力し，各層から単語ベクトルを得る．このとき，単語の頻度 f も併せて求めておく．以降の処理は層別に行う．

得られた単語ベクトルを処理して，単語タイプごとに語義の豊富さ v を求める．文献 [7] に従い，語義の豊富さは，$v \equiv \frac{1-l^2}{l(d-l^2)}$ とする．ただし，l と d は，それぞれ，単語タイプごとに単語ベクトルを平均した結果のノルムと単語ベクトルの次元である．

得られた単語頻度 f と語義の豊富さ v を式（1）に回帰して傾き δ を得る．その際，単語頻度と語義の豊富さを bin にまとめそれぞれの平均値を回帰に用いる．本稿では bin サイズは文献 [7] に従い 100 とする．以上の処理は，層別に行うため，層の数だけ δ の値が得られる．

2 節で述べたように，語義識別能力が高いほど，傾き δ の値が大きくなることが知られている．この知見に基づき，本稿では傾き δ が最大となる層を最適層として選択する．ただし，補助情報として，回帰における決定係数 R^2 も用いる（本稿では $R^2 > 0.85$ とする）．なぜなら，決定係数が小さい場合は，meaning-freuency law がそもそも観測できず，語義識別能力は低いと判断されるためである．

4 傾きに基づいた最適な層の調査

提案手法を用いて最適な層を調査した．対象としたのは次の 4 モデルである：英語：bert-base-uncased[2)]，bert-large-uncased[3)]；日本語：bert-base-japanese-v3[4)]，bert-large-japanese-v2[5)]．以降では英語 bert-base のように略記する（文脈から明らかな場合は対象言語も省略する）．英日いずれも bert-base は 12 層，bert-large は 24 層から成る．モデルに合わせて，英語（CCOHA [9] の 2000 年代の文書）と日本語（BCCWJ [10]）のコーパスを利用した．頻度上位 20,000 件を対象にして回帰を行った．

まず，meaning-frequency law が観測できるかを確認する．図 1 に，英語 bert-base の奇数層における単語頻度 f と語義の豊富さ v のプロットを示す．第 1 層目（L01）以外は，決定係数が高く，同法則が観測できることがわかる．紙面の関係から割愛するが，偶数層および他のモデルいずれにおいても同様な傾向がみられた（付録 A.1，A.2 を参照のこと）．

詳細な分析のため，図 2 に層別の傾き δ の大きさを示す．横軸と縦軸はそれぞれ層番号と傾き δ の値に対応する．また，赤色の点が δ の最大値を表す．

図 2 より，傾きが最大となる層はモデルと言語により異なることがわかる．傾き δ の大きさが語義識別能力に対応するという仮定が正しいとすると，最終層や前最終層がよいとする知見が正しくない場合があることになる．次節では，本節の結果と実際の語義識別能力を比較し，この予想を吟味する．

5 傾きと語義識別能力の関係

本節では，層別の傾き δ を語義識別能力と直接比較することで提案手法の有効性を評価する．評価データは，意味変化検出用の評価データ DWUG Version 3.0.0 [11] の英語データとした．同データには，46 種類の単語について二種類の情報（変化の有無，変化の度合い）が収録されている．これに従い，意味変化の有無の検出と変化の度合いを推定す

1） 1 節で定義したように，最終層の直前の層である．

2） https://huggingface.co/google-bert/bert-base-uncased

3） https://huggingface.co/google-bert/bert-large-uncased

4） https://huggingface.co/cl-tohoku/bert-base-japanese-v3

5） https://huggingface.co/cl-tohoku/bert-large-japanese-v2

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

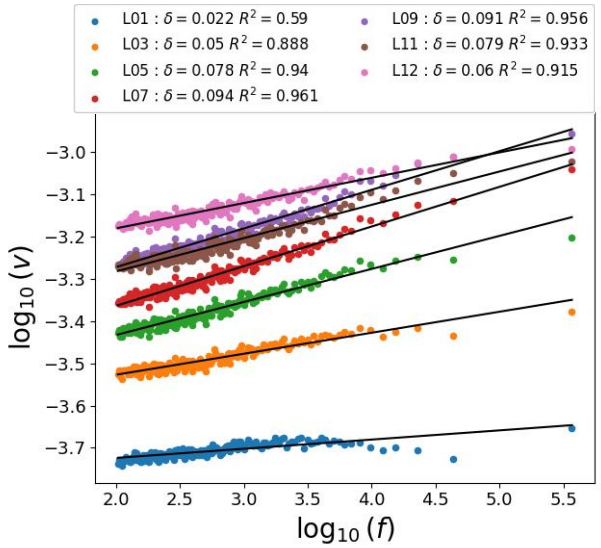


図 1　頻度 f と語義の豊富さ v の関係．コーパス：CCOHA 2000 年代文書．L01 などは層番号に対応．L12 が最終層．

る二つのタスクを実施した．従来にならい，前者は検出精度，後者はスピアマンの順位相関を評価尺度とした．検出／推定手法として，文献 [12] で最高性能と報告されている事例間余弦距離の平均に基づく手法を選択した．また，提案手法でも用いている語義の豊富さ v に基づいた意味変化検出手法 [13] も対象とした．ただし，この手法は変化の度合い推定について有効でないため対象外とした[6]．言語モデルは，英語の bert-base と bert-large を利用した．以上の条件で求めた性能値と前節で得た傾き δ の値を比較した．

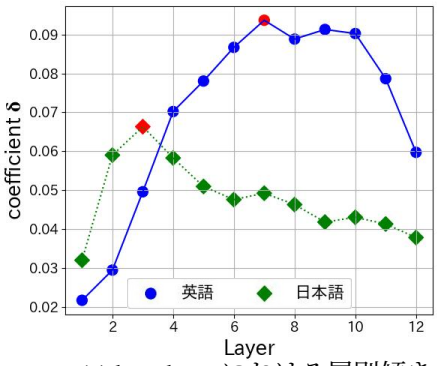
表 1 に結果を示す．太字は，行内で最大の性能値を表す．また，カッコ内の数値は，最適層の性能値からの差となる．したがって，カッコ内の数値が 0 である場合，最適層を選択できたことになる．また，比較のため，最終層，前最終層の性能も示して

表 1　選択された層における各タスクの性能．

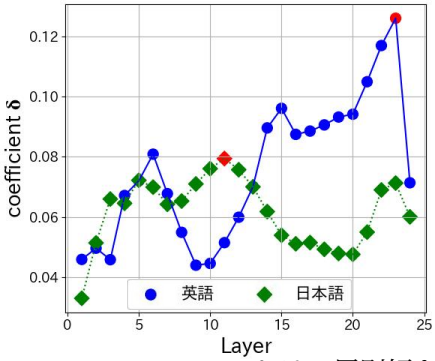
モデル	手法	最終層	前最終層	提案手法
意味変化検出（精度）				
bert-base	手法 [12]	0.58 (−0.10)	0.57 (−0.11)	**0.61** (−0.07)
	手法 [13]	0.63 (−0.02)	**0.65** (0)	0.57 (−0.08)
bert-large	手法 [12]	**0.58** (−0.08)	**0.58** (−0.08)	**0.58** (−0.08)
	手法 [13]	**0.59** (−0.04)	**0.59** (−0.04)	**0.59** (−0.04)
意味変化の度合い推定（順位相関係数）				
bert-base	手法 [12]	0.51 (−0.10)	0.44 (−0.17)	**0.53** (−0.08)
bert-large	手法 [12]	**0.52** (−0.16)	0.42 (−0.26)	0.42 (−0.26)

カッコ内の数値は最適層の精度からの差分．太字は対応する手法における最大性能を示す．

6）　理論上は適用可能であるが，どの層を用いても順位相関係数が零に近い値になるため除外した．



(a) bert-base における層別傾き



(b) bert-large における層別傾き

図 2　層別の傾き δ の大きさ．

いる．なお，図 2（b) からわかるように，bert-large については，提案手法は前最終層を選択していることに注意されたい（したがって，表中の前最終層の性能と一致する）．

表 1 より，全ての層選択手法を通じて，最適層を選択できたのは一ケースのみであり，そもそも最適層を選択するのは難しい問題であることがわかる．提案手法は必ずしも最適層を選択できるわけではないが，bert-base を用いた手法 [12] については，比較した層の中での最大性能を達成できている．より詳細な分析のため，層別に，傾き δ と各手法の性能値をプロットした図を吟味する．図 3 は意味変化検出精度，図 4 は意味変化の度合いの推定性能である．両図とも，横軸は層に対応し，縦軸は傾き δ と性能値に対応する．

両図より，最適層はモデル，言語，対象手法，タスクにより，かなり変化することがわかる．とりわけ，手法 [12]（図 3，図 4 中では，“average pairwise distance”）は，低〜中層で性能が最大になっており興味深い．以上の結果より，最終層や前最終層など決め打ちで層を選択するのではなく，提案手法のように適応的に層を選択することが重要といえる．ただし，提案手法は言語モデルと対象言語については

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

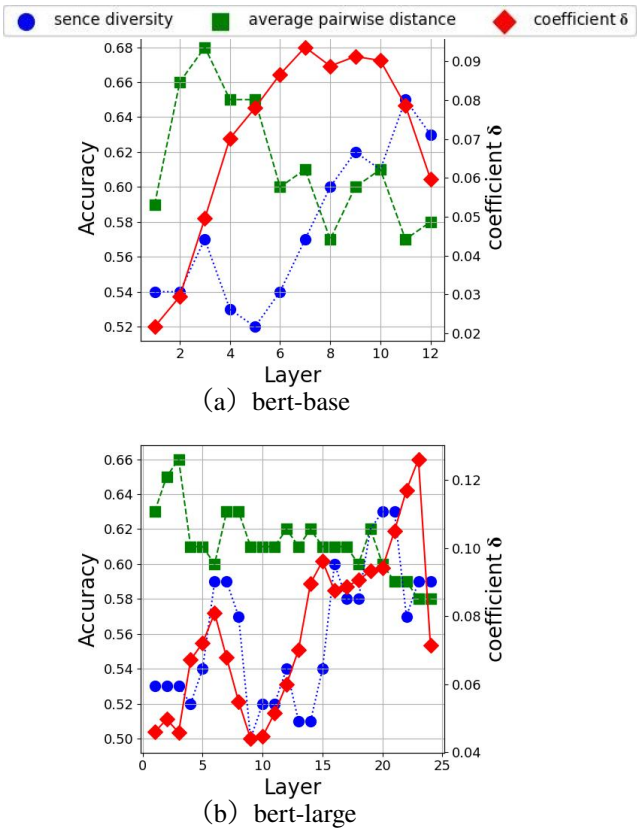


図 3　層別の傾き δ と検出精度の関係.

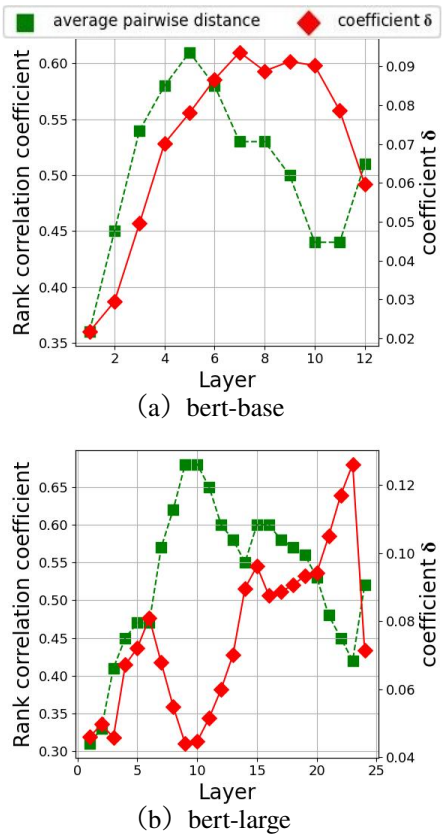


図 4　層別の傾き δ と意味変化の度合い推定性能の関係.

適応的であるが，対象手法についてはその限りでない．対象手法に関しても適応できることが望ましいが，汎用性とのトレードオフである．

また，図 3，4 から，傾き δ の大きさが，性能傾向を大まかに捉えているようにみえる．実際，傾き δ と性能値のスピアマンの順位相関係数は，手法 [13] では，bert-base と bert-large でそれぞれ 0.36, 0.63 と中程度の相関がみられる（全ての相関係数は付録 A.3 に示す）．一方，bert-large を用いた手法 [12] などでは，−0.61 など負の相関もみられる．図 4 でも，第 8〜12 層目で傾き δ と性能値のグラフが線対称のようになっており負の相関を示唆する．これらの層では，相対的に決定係数が低く，傾き δ の値の信頼度も低いことが理由として考えられる（付録 A.1 の図 5 に全ての決定係数の値を示す）．このことから，傾き δ と併せて決定係数 R^2 も考慮するすることで更なる性能向上が期待できる．

以上の結果は次の通りまとめられる．従来の知見と異なり，最終層や前最終層の出力が単語ベクトルとして最適ではないことがある．そもそも，最適な層を決定することは難しい問題で，究極的には検証データを用意して性能を評価する必要がある．ただし，言語研究では適当な検証データが存在しないことが多い．また，仮説の検証などそもそも性能を測ることができない場合もある．そのような状況において，提案手法は一つの有益な指針となる．提案手法は常に最適層を選択できるわけではないが，今回の結果を見る限り，最適層からの大幅な性能低下は抑えられる可能性が高い．

6　おわりに

本研究では，meaning-frequency law による傾き δ を利用して単語ベクトルを得るための言語モデルの層を選択する手法を提案した．提案手法の特徴は，生コーパスさえ利用可能であれば適応的に層選択ができるという汎用性にある．実データを用いて提案手法の評価を行ったところ次の三つの知見が得られた：(1) 最適な層は最終層や前最終層であるとする従来の知見は必ずしも正しくなく，層選択は適応的にされるべきである：(2) 提案手法は常に最適層を選択できるわけではないものの，汎用的に層選択が行える；(3) 提案手法で選択された層は最大性能からの性能低下を抑え，層選択のための指針となる．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

謝辞

産総研及び AIST Solutions が提供する ABCI 3.0 を「ABCI 3.0 開発加速利用」を支援を受けて利用した．

参考文献

[1] George Kingsley Zipf. The meaning-frequency relationship of words. **The Journal of General Psychology**, Vol. 33, No. 2, pp. 251—256, 1945.

[2] Taichi Aida and Danushka Bollegala. Unsupervised semantic variation prediction using the distribution of sibling embeddings. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 6868–6882, 2023.

[3] Andrey Kutuzov and Mario Giulianelli. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In **Proceedings of the Fourteenth Workshop on Semantic Evaluation**, pp. 126–134, 2020.

[4] Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. Explaining and improving BERT performance on lexical semantic change detection. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop**, pp. 192–202, 2021.

[5] 大谷直輝, 永田亮, 高村大也, 川崎義史. 深層学習を用いた構文文法の実証的な研究の可能性を探る — better off 構文を例にして —. 言語研究, Vol. 166, pp. 59–86, 2024.

[6] Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **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. 55–65, 2019.

[7] Ryo Nagata and Kumiko Tanaka-Ishii. A new formulation of Zipf's meaning-frequency law through contextual diversity. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15323–15335, 2025.

[8] Ryo Nagata, Yoshifumi Kawasaki, Naoki Otani, and Hiroya Takamura. A computational approach to quantifying grammaticization of English deverbal prepositions. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 211–220, 2024.

[9] Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. CCOHA: Clean corpus of historical American English. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 6958–6966, 2020.

[10] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. **Language Resources and Evaluation**, Vol. 48, pp. 345–371, 2014.

[11] Dominik Schlechtweg, Pierluigi Cassotti, Bill Noble, David Alfter, Sabine Schulte Im Walde, and Nina Tahmasebi. More DWUGs: Extending and evaluating word usage graph datasets in multiple languages. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 14379–14393, 2024.

[12] Francesco Periti and Nina Tahmasebi. A Systematic Comparison of Contextualized Word Embeddings for Lexical Semantic Change. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 4262–4282, 2024.

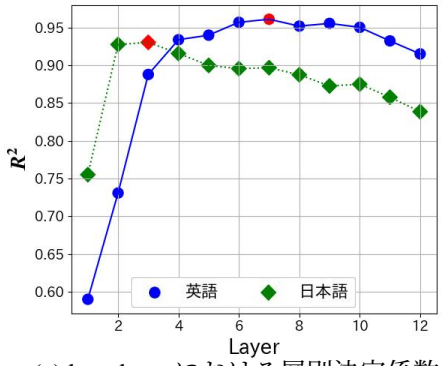
[13] Ryo Nagata, Hiroya Takamura, Naoki Otani, and Yoshifumi Kawasaki. Variance matters: Detecting semantic differences without corpus/word alignment. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 15609–15622, 2023.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

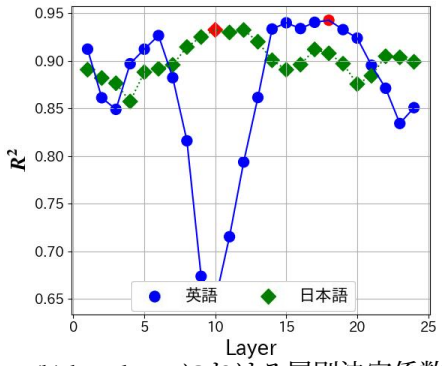
A　付録

A.1　層別の決定係数の値

図 5 に，層別の決定係数 R^2 の値を示す．英語 bert-large では第 8〜12 層目に決定係数の大きな落ち込みがみられる．



(a) bert-base における層別決定係数



(b) bert-large における層別決定係数

図 5　層別の決定係数の値．コーパス：CCOHA 2000 年代文書．

A.2　英語における頻度と語義の豊富さの関係の詳細

図 6 に，bert-large における単語頻度 f と語義の豊富さ v のプロットを示す．

A.3　層別の傾きとタスク性能との関係

表 2 に，傾き δ と各手法における性能値とのスピアマンの順位相関係数を示す．

表 2　層別の傾きとタスク性能の間の順位相関係数．

モデル	手法	意味変化検出	変化度合いの推定
bert-base	手法 [12]	−0.25 (0.44)	0.12 (0.72)
	手法 [13]	0.36 (0.26)	NA
bert-large	手法 [12]	−0.61 (0.00)	−0.20 (0.36)
	手法 [13]	0.63 (0.00)	NA

カッコ内の数値は p 値．

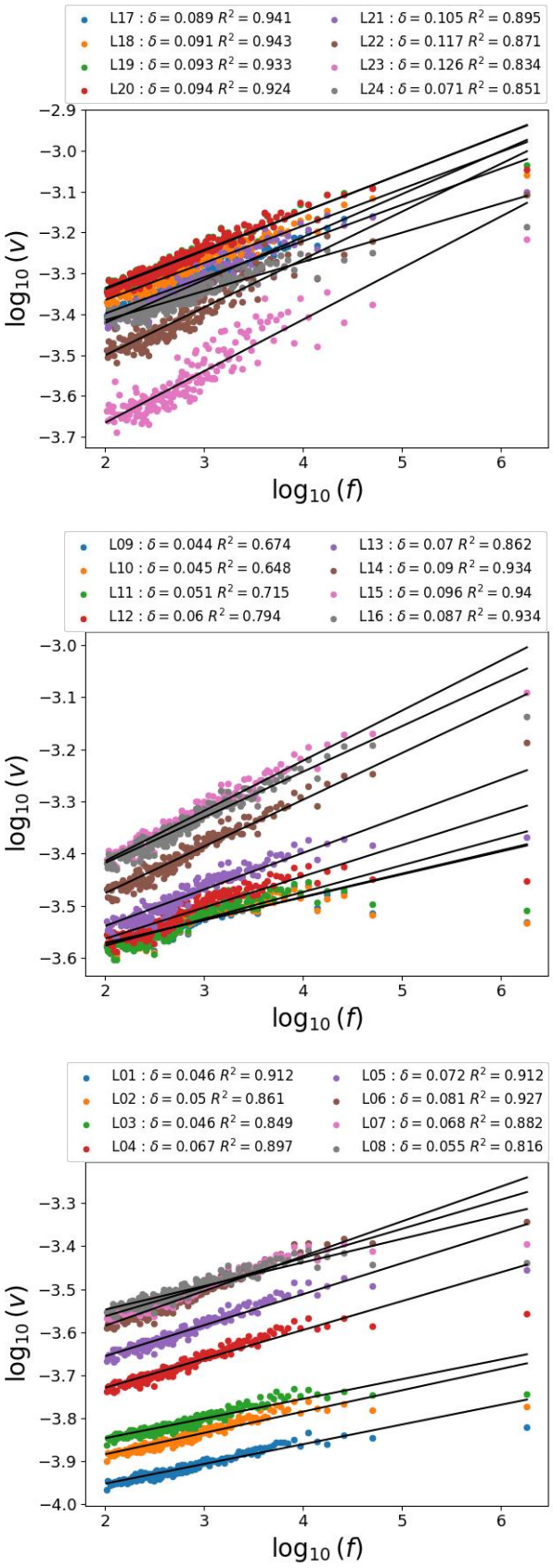


図 6　英語における bert-large の頻度と語義の豊富さの関係．コーパス：CCOHA 2000 年代文書．L01 などは層番号に対応．L24 が最終層．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).