単語ベクトルに基づく新たな meaning-frequency law の検証

永田亮 [1] 田中久美子 [2]
[1] 甲南大学 [2] 早稲田大学
nagata-meaningfreqlaw @ ml.hyogo-u.ac.jp.

概要

本稿では，meaning-frequency law として知られる単語頻度と語義数に関する法則を従来とは全く異なるアプローチで検証し，従来より多様な種類の単語について同法則が成り立つことを示す．また，その過程で，同法則を通じて，言語モデル（具体的には BERT）の能力差を計測できることを示す．

1 はじめに

本稿では，Zipf の meaning-frequency law [1] として知られる単語の語義数と頻度に関する法則を，辞書を用いずに，単語ベクトルを通じて検証する方法を提案する．従来は，検証において単語集合に制限を要したが，提案するアプローチをとることで，制限を緩和し，同法則がより多様な種類の単語に対して成り立つことを示す．

meaning-frequency law とは，頻度が高い単語ほど，語義数が多くなるという経験則である．より形式的には，単語の頻度を f，語義数を m としたとき，

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

という冪乗則が成り立つというものである（ただし，δ と c は定数である）．詳細は 2 節で述べるが，従来研究では，基本的には，辞書から語義数 m を得て検証を行うものであった．

しかしながら，辞書に基づいた検証には様々な制限がある．そもそも，ある単語の語義数を決定すること自体が難しいタスクである．例えば，take という単語の語義数は，ジーニアス英和辞典（第 5 版）では 32，プログレッシブ英和辞典（第 5 版）では 6 と大きな差がある [2]（同書には，辞書により語義数が大きく異なる語が多数紹介されている）．語義数が異なれば結果も異なる可能性がある．また，辞書に掲載された語義が全てコーパスに出現するわけではないという問題もある（同様に，逆のケースも生じる）．実際，従来研究では，この問題に対処するために高頻度語や機能語を除外し，単語の原型を対象として検証を行ってきた．したがって，機能語，高頻度語，活用形を含めた場合に meaning-frequency law が成り立つかどうかは明らかではない．

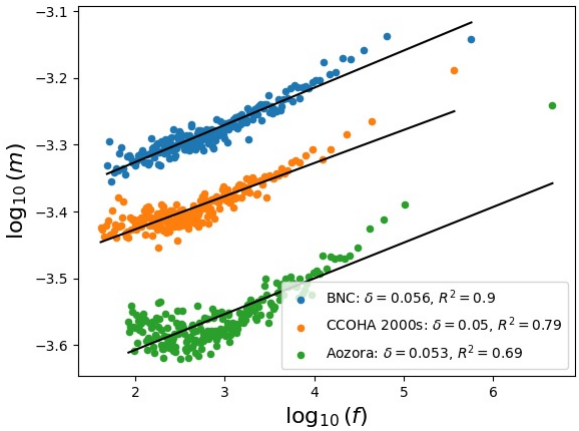

図 1 単語頻度 f と語義の豊富さ m の関係．

そこで，本稿では，辞書を利用せずに，言語モデル（本稿では，BERT）から得られる単語ベクトルを通じて meaning-frequency law を検証する．具体的には，単語ベクトルの散らばり度合いに基づいて語義の豊富さを定量化する．結論を先に述べると，図 1 のように，頻度と語義の豊富さの関係は式 (1) に概ね従うことがわかる．

本研究の貢献は次の 3 点に要約される：(1) 辞書を利用せずに meaning-frequency law を検証する手法を提案する，(2) 提案手法を様々なコーパスに適用して，同法則が従来示されているより多様な種類の単語について成り立つことを示す，(3) 更に，その過程で，同法則を通じて BERT の能力差を計測できることを示す．

2 関連研究

meaning-frequency law は，Zipf 本人による検証以外にも様々な研究者により検証が行われている．例えば，英語 [3]，トルコ語 [4]，複数言語 [5, 6] を対象にした研究がある．また，子供の発話データでも同法則が成り立つという報告 [7] もある．しかしなが

ら，いずれにおいても，語義数が辞書に基づいた検証となる点に問題がある．関連して，活用形を基本形に変換して，また，高頻度語や機能語は対象外にして検証が行われてきた．例えば，従来，用いられることが多い WordNet は，名詞，形容詞，動詞，副詞のみ語義数の情報を収録し，既存研究 [3, 4, 5, 6] では，これらのみを対象とする．

Ilgen ら [4] と Bond ら [5] の検証では，コーパス中の単語に語義を付与することで語義数を求めている．その場合でも，語義集合は既存の辞書に基づくため，情報がない語義，場合によっては単語自体が除外される．また，人手による語義の付与は時間と労力を要するため，検証の規模を大きくすることが難しい．

本稿で提案する方法は，辞書を用いずにコーパスデータのみで検証を可能とする．その結果，検証過程は完全に自動化され大規模なコーパスも対象とすることができる．また，語義数の決定や語義の欠損などの難しい問題も自然に回避できる．

3 検証方法

提案する検証方法では，式 (1) の語義数 m を異なる方法で求める以外は，Bond らの方法 [5] と同じ手順をとる．m は得られているものとし，共通する手順を説明する．

検証は，回帰に基づいて行う．得られたデータ（単語の頻度 f と語義数 m）を式 (1) にあてはめ回帰を行う．ただし，回帰の際に，説明変数と目的変数の bin 平均化を行う．すなわち，頻度の大きい順に，f および m をいくつかの bin にまとめ平均値を求め，更に対数をとってから回帰を行う．本稿では，Bond らの研究 [5] でも用いられている bin サイズ 100 を用いる．Bond ら [5] をはじめ多くの従来研究では，回帰の当てはまり具合を表す決定係数 R^2 により同法則の成り立ち度合いを考える．加えて，傾き δ に対する p 値[1)] を判断材料としている[2)]．

本稿でも，決定係数 R^2 と傾き δ に基づいて meaning-frequency law の検証を行う．加えて，単語の頻度 f と語義数（に対応する量）m の散布図を示す．

残るは，式 (1) の語義数 m をいかに求めるかである．提案手法では，語義数に対応した語義の豊富さという指標を用いる．以降では，m により語義の豊富さを表すこととする．

語義の豊富さは，文脈付き単語ベクトル（以下，単に単語ベクトルと省略）を利用して定量化する．単語ベクトルの方向に基づいて，単語間の意味的な類似度を測る慣習（i.e., ベクトル間の余弦類似度）を考慮すると，単語ベクトルの方向の多様性が語義の豊富さを表すと考えるのは自然である．ベクトルの方向のばらつきは，von Mises-Fisher 分布 [8] を通じて定量化が可能である．この分布は，d 次元の単位ベクトル $\mathbf{x}$ に対して，$f(\mathbf{x};\boldsymbol{\mu},\kappa) \propto \exp\left(\kappa\boldsymbol{\mu}^{\mathsf{T}}\mathbf{x}\right)$ と定義される（ただし，単語ベクトルは単位ベクトルとは限らないため，ノルム 1 となるように事前に正規化しておく）．ここで，$\boldsymbol{\mu}$ ($\|\boldsymbol{\mu}\| = 1$) と κ ($\kappa \geq 0$) は，それぞれ平均方向と集中度と呼ばれるパラメータである．この分布は，超球面上の正規分布に例えられ，単位ベクトル $\mathbf{x}$ が平均方向 $\boldsymbol{\mu}$ を中心に集中度 κ で等方的に分布すると考える．言い換えれば，κ は，ベクトルの方向の集中度合を表している．既存研究 [9] では，集中度 κ が意味変化の検出に有効であることを示している．

したがって，本稿では，語義の豊富さを集中度 κ により捉えることを提案する．本稿で必要となるのはベクトルの方向のばらつきであるので，逆数をとり $m \equiv 1/\kappa$[3)] とする．

集中度 κ の最尤推定は，近似的に，

$$\kappa \approx \frac{l(d - l^2)}{1 - l^2} \tag{2}$$

になることが導かれている [8]．ここで，l は，単語ベクトルを（単語タイプごとに平均した）平均ベクトルのノルムである（d は上述の通り，ベクトルの次元である）．

以上をまとめると，提案する検証方法は次の 6 ステップとなる：(1) 入力コーパス中の単語頻度 f をカウント；(2) 各単語を単語ベクトルに変換；(3) 得られた単語ベクトルのノルムが 1 となるように正規化；(4) 単語タイプごとに平均ベクトルを求め，そのノルム l を算出；(5) 式 (2) を用いて $m = 1/\kappa$ を算出；(6) 得られた f と m に対して回帰を行い，決定係数 R^2，傾き δ を算出．(6) 決定係数 R^2，傾き δ を散布図とともに出力し，meaning-frequency law が成り立つかを吟味．

1) 傾きが零かどうかを t 分布を用いて検定する．
2) Zipf [1] では，$\delta = 0.5$ が提案されているが，2 節で紹介した研究では，$\delta < 0.5$ が報告されており，δ が正かどうかを考察材料としてる．
3) 本稿では，正規分布との類似性を考慮して，逆数で m を定義したが，k の別の関数を考えることも可能である．

4 基礎的検証

3 節で述べた方法を用いて，meaning-frequency law の検証を行う．コーパスは，英語（BNC [10]，CCOHA [11] の 2000 年代の文書），日本語（青空文庫[4]）を用いた（詳細は，付録 A に示す）．単語ベクトルは，bert-large-uncased，cl-tohoku/bert-large-japanese-v2[5] の最終層の出力を用いた．頻度順位 20,000 以上の単語を対象とした（複数のサブワードに分割される単語はあらかじめ除外した）．

結果を（1 ページ目の）図 1 に示す．図 1 より，コーパスを問わず，式 (1) に概ね回帰することがわかる（回帰係数も $p < 0.01$ で有意．以降の回帰分析についても全て $p < 0.01$ で有意）．ただし，高頻度帯で乖離する部分がみられる．

以上は，従来とは異なり，高頻度語，機能語，活用形も対象に含めた検証である．また，表 1 に示す通り，取り扱う語彙サイズも大きい．したがって，より幅広い範囲で meaning-frequency law が成り立つことを新たに示したといえる．

表 1 本研究および従来研究で扱う語彙サイズ．

研究	語彙サイズ
本研究	20,000
Bond ら [5]（英語）	14,500
Bond ら [5]（日本語）	10,000
Cassas ら [6]（英語）	16,200

5 拡張検証

5.1 言語モデルの能力との関係

言語モデルのサイズが小さい場合，細やかな意味の差異が捉えられなくなり，meaning-frequency law が観測されなくなると予想される．図 2 に，サイズの異なる 6 種類の BERT を CCOHA の 2000 年代の文書に適用して得た，頻度と語彙の豊富さの関係図を示す．詳細は付録 B に示すが，モデルのサイズ順に凡例のラベルを記している．

予想通り，モデルサイズが小さくなると meaning-frequency law が観測されない．頻度が高いほど語義が乏しくなるという逆の傾向を示す．一つの説明として，次のような仮説を考えることができる．モデルサイズが小さいと表現力が足りず，語義を十分に峻別できないと仮定する．語義が峻別できないとは，単語ベクトルの世界では二つの語義に対して類似した単語ベクトルが割り当てられることを意味する．その結果，平均単語ベクトルのノルムで決定される語義の豊富さはより小さい値となる[6]．したがって，語義の豊富さを十分に峻別できない場合，頻度が高い単語ほど語義の豊富さが小さく見積もられるということが予想される．

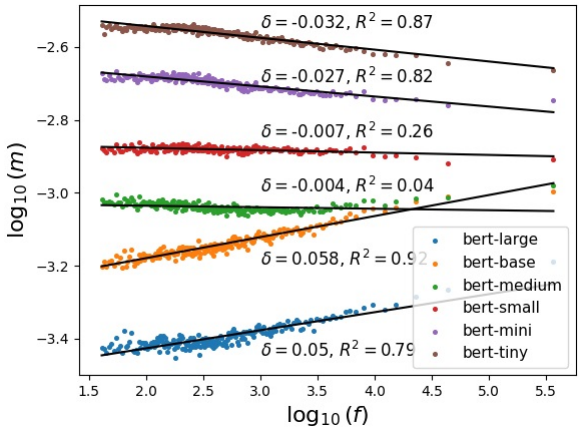

図 2 異なるサイズの言語モデルにおける単語頻度 f と語義の豊富さ m の関係．

このことを別の角度から吟味するため，ランダムシャッフルした CCOHA の 2000 年代文書を用いた検証を行う（シャッフルコーパスの詳細は付録 C に示す）．このシャッフルコーパスから bert-large-uncased を用いて語義の豊富さ m を求めた．

図 3 に結果を示す．図 3 より，シャッフルコーパスでは，おおよそ頻度 10000 未満の単語について，語義の豊富さが増加する傾向が見てとれる．シャッフルコーパスは，完全なランダム単語列ではなく，単語の頻度はシャッフル前のコーパスと同一である．そのため，文脈がシャッフルされても，各単語とも，高頻度語との共起は多くなる傾向となる．更に，語彙サイズも同一のため，高頻度語については，文脈の衝突（たまたま，文脈内に同じ単語が出現すること）が多くなる．これは，モデルサイズが小さく語義が十分に峻別できない状況と似ている．図 3 における，低頻度帯の増加は，このことを反映している可能性もある．

4） https://github.com/aozorahack/aozorabunko_text．2023 年 12 月 3 日アクセス．

5） https://huggingface.co/docs/transformers/model_doc/bert，https://huggingface.co/cl-tohoku/bert-large-japanese-v2

6） 類似したベクトルが多いほど，平均ベクトルのノルムは大きくなる．極端な場合，全ての事例が常に同一のベクトルに変換されと，平均ベクトルも同一となり，ノルムは最大の 1 となる．なお，全ての単語ベクトルは事前にノルム 1 に正規化されていることに注意されたい．

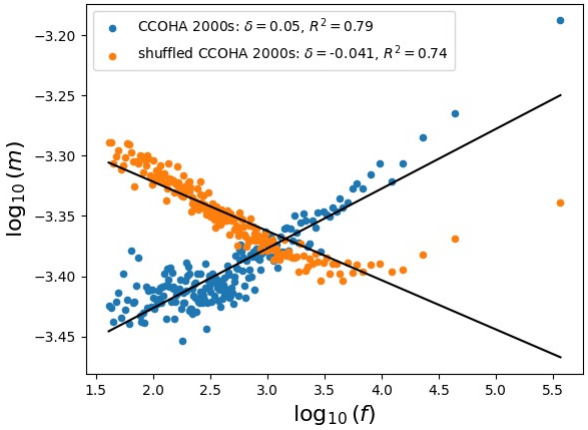

図 3　通常コーパスとシャッフルコーパスにおける単語頻度 f と語義の豊富さ m の関係.

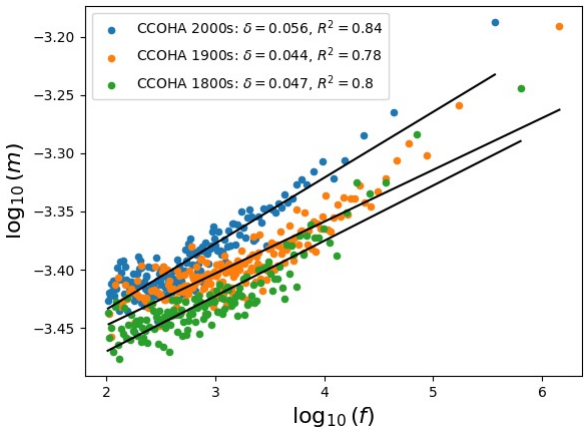

図 4　年代別コーパスにおける単語頻度 f と語義の豊富さ m の関係.

5.2　想定外データの場合

本節では，言語モデルが想定する言語データ以外について吟味する．データとして，古い文書（CCOHA の 1900 年代と 18000 年代）と非母語話者英語コーパス（Lang-8[7] の 2012〜2019 年のデータ）を用いる．使用する言語モデルは bert-large-uncased である．

図 4 に，CCOHA を用いた結果を示す．図 4 より，全ての年代の文書で，meaning-frequency law が観測されることがわかる．主に現代の文書で訓練された BERT でも，少なくとも 1800 年ぐらいまでの文書であれば，語義をある程度峻別できるという一つの証拠となる．実際，BERT が 1800 年代と 2000 年代の文書における意味変化検出に有効であることが報告されている（文献 [9, 12]）．

一方で，古い文書では低頻度帯で傾きが小さくなるという傾向がみられる．また，CCOHA 2000 年代

7)　https://lang-8.jp/

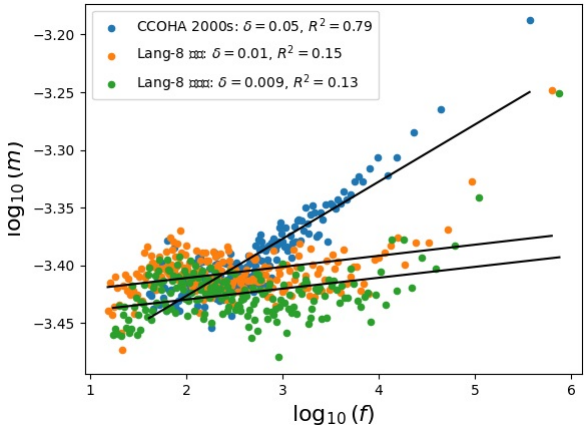

図 5　非母語話者英語コーパスにおける単語頻度 f と語義の豊富さ m の関係.

の文書よりも全体的に語義の豊富さが低い．切片は，使用コーパスのサイズに依存し，かつ，CCOHA では，年代により各レジスタ内の文書数が異なるため慎重に議論を進める必要があるが，主に現代の文書で訓練された BERT が，古い文書における意味を現代の文書と同程度には峻別できていない可能性もある．別の可能性として，年代が進むにつれて，各語における語義数が平均的に増えているという予想もできる．すなわち，個別にみると語義を失う語も，得る語もあるが，平均的にはより多義語になっていく傾向があるという予想である．

図 5 に，非母語話者英語コーパスに対する結果を示す．図 5 は，図 4 と同じような傾向を示すが，CCOHA 2000 年代からの乖離がより大きい．古い文書の場合と同様に，BERT が非母語話者の文書における意味を十分には峻別できていない可能性がある．逆に，非母語話者が母語話者と同程度に語義を使い分けられていない可能性もある．別の非母語話者英語コーパスを用いるなど更なる検証が待たれる．

6　おわりに

本稿では辞書を用いずに meaning-frequency law を検証する方法を提案した．この方法により，従来では難しかった，高頻度語，機能語，活用形も考慮した検証を可能とした．実際に，この方法を用いて，英語と日本語で meaning-frequency law が，従来より幅広い種類の単語に対して成り立つことを示した．また，歴史コーパスと非母語話者コーパスでも同法則がある程度成り立つことも示した．更に，その過程で，同法則を通じて，サイズが異なる BERT の能力差を計測できることを示した．

謝辞

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] 寺澤盾. 英単語の世界. 中央公論新社, 東京, 2016.

[3] Philip Edmonds. **Lexical Disambiguation**, pp. 43–62. Elsevier, Amsterdam, 2005.

[4] Bahar Ilgen and Bahar Karaoglan. Investigation of Zipf's 'law-of-meaning' on turkish corpora. In **Proceedings of the 22nd International Symposium on Computer and Information Sciences**, pp. 1–6, 2007.

[5] Francis Bond, Arkadiusz Janz, Marek Maziarz, and Ewa Rudnicka. Testing Zipf's meaning-frequency law with wordnets as sense inventories. In **Proceedings of the 10th Global WordNet Conference**, pp. 342–352, 2019.

[6] Bernardino Casas, Antoni Hernández-Fernández, Neus Català, Ramon Ferrer-i Cancho, and Jaume Baixeries. Polysemy and brevity versus frequency in language. **Computer Speech and Language**, Vol. 58, No. C, p. 19–50, 2019.

[7] Antoni Hernández-Fernández, Bernardino Casas, Ramon Ferrer i Cancho, and Jaume Baixeries. Testing the robustness of laws of polysemy and brevity versus frequency. In **International Conference on Statistical Language and Speech Processing**, pp. 19–29, 2016.

[8] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. **Journal of Machine Learning Research**, Vol. 6, No. 46, pp. 1345–1382, 2005.

[9] 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.

[10] BNC Consortium. The british national corpus, 2001.

[11] Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. CCOHA: Clean corpus of historical American English. In **Proc. of the 12th Language Resources and Evaluation Conference**, pp. 6958–6966, 2020.

[12] 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.

表 2　使用したコーパスとサイズ．サイズは，英語についてはトークン数，日本語については文字数である．

コーパス	サイズ
CCOHA 1800 年代	111,048,657
CCOHA 1900 年代	262,200,025
CCOHA 2000 年代	68,678,659
BNC	109,369,848
青空文庫	198,755,598
Lang-8 原文	127,864,912
Lang-8 訂正文	152,681,283

表 3　使用した BERT モデルのサイズ．

モデル名	Layer 数	隠れ層サイズ
bert-large	24	1,024
bert-base	12	768
bert-medium	8	512
bert-small	4	512
bert-mini	4	256
bert-tiny	2	128

付録

A　使用コーパスの詳細

表 2 に，検証に使用した各コーパスのサイズを示す．英語については，トークン分割した後のトークン数である．日本語については，文字数である．

CCOHA については，次のような前処理を行った．ノイズと思われる文書は除外した．具体的には，「@@年.txt」（例：@@1525.txt）のように年とファイル名と思われる文字列を含む文書は分析対象外とした．また，文書中のタグ（<P></P> など）は除去した．更に，伏せ字が含まれている文（CCOHA では，著作権の制限により，一定の割合で文章の一部が伏せ字になっている）も除外した．

Lang-8 コーパスは，学習者が作成した文とそれを訂正した文の情報を提供する（ただし，全ての文に訂正情報があるわけではない）．訂正情報がない文については，原文をそのまま訂正文とした．なお，対象言語は英語である．

青空文庫については，Github で公開されているデータ（https://github.com/aozorahack/aozorabunkotext）を使用した．pySBD[8] で文分割した．トークン分割については，cl-tohoku/bert-large-japanese-v2 に従った．

B　モデルサイズが異なる BERT の情報

サイズの異なる 6 種類の英語 BERT を 5 節で使用した．各モデルのサイズを表 3 に示す．なお，表 3 では省略しているが，全て大文字小文字を区別しないモデル（uncased）を用いた．

C　シャッフルコーパスの詳細

シャッフルコーパスの作成方法は次のとおりである．CCOHA の 2000 年代の文書をひとつの長い単語列とみなし，ランダムに 1 単語ずつ取り出して，新たな単語列を作り出すというのを 5 回繰り返し，シャッフルコーパスとした．得られた単語列中の記号.?!を文末位置とみなした．こうして得られるシャッフルコーパスは，単語の並びはランダムにシャッフルされるが，長さおよび各単語の頻度の点で元コーパスと同一となる．

8）https://github.com/nipunsadvilkar/pySBD