語間の「意味の広さ」の差だけを統計的に検定する手法

江原　遥
東京学芸大学
ehara@u-gakugei.ac.jp

概要

本研究では，語の意味の差に影響されず，語の「意味の広さ」の差だけを検定する手法を提案する．球面上の文脈化埋め込みベクトルから語集合の意味の広さを測る画期的な手法として [1] が提案されているが，具体的に語間の意味の広さの差の検定手法は提案されていない．まず，単純な置換検定法では第一種の過誤が生じる問題があることを示す．本研究では Householder 変換により平均ベクトルを整列することでこの問題を解決できることを理論的に示す．GPU を用いた高速な置換検定法も示す．実データで第一種の過誤が 62.2%減少し，GPU により 23 倍の高速化を達成した．さらに平均方向の整列により，日本語と英語で意味の広さに統計的有意な差がある語ペア (油-oil など) を初めて抽出した．

1　はじめに

文脈化埋め込みは，NLP における語彙意味の標準的表現となっており，文脈やコーパスに伴う意味の変化を精密に分析するために広く用いられている [2, 3, 4]．従来研究の多くは，たとえば 2 つの用法が同一語義かどうか，あるいは領域や時間によって語の意味がどう変化するかといった**意味差**の検出に焦点を当ててきた．こうした課題は，母語話者に意味類似度の判断を求めるなど，直接アノテーションしやすい場合が多い．

一方で，**意味の広がり**（語が多様な文脈実現にどれほど広くまたがるか）を即時に見抜くことは難しい．話者が用法間の差異を感じ取れても，語が「広い」かどうかを判断するには，多数の自然用例を見比べる必要があることが多い．このため近年は，BERT[5] などのマスク言語モデルを用いて，文脈化トークンベクトルの幾何学（トークン点群）から得られる文脈多様性を，意味の広がりのコーパス駆動な近似量として用いる研究が提案されている [1]．語・コーパス・モデル間でこうした広がり指標を比較するには，差が小さい領域でも信頼できる統計検定が必要である．

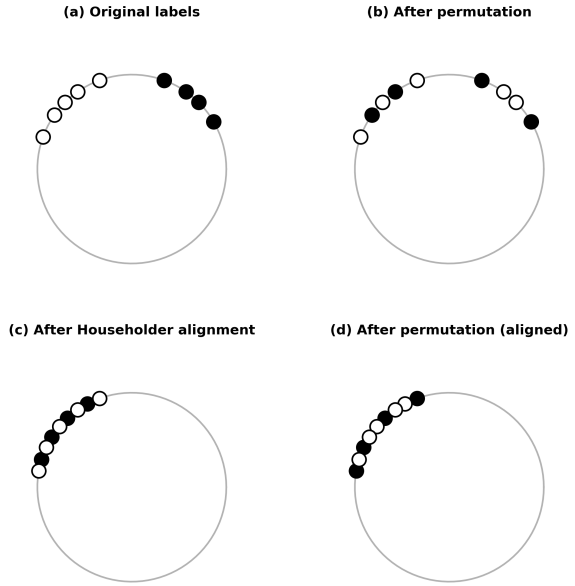

図 1　素朴な置換検定における第一種過誤の増大と，Householder 整列化による解消の模式図．各円は単位球面（2 次元に投影して表示）を表し，黒点・白点は 2 つの語タイプから得た ℓ_2 正規化埋め込みである．**上段（素朴な検定）：**(a) 2 群は球面上の異なる領域（平均方向が異なる）を占めるが，分散は同程度である．(b) ラベルを無作為に入れ替えると，黒点が両領域に跨って配置され，見かけ上分散が過大になる．この見かけの分散増大が偽陽性（第一種過誤）を引き起こす．**下段（Householder 整列化検定）：**(c) Householder 反射を適用して平均方向を一致させると，2 群は同一領域に整列する．(d) 整列後のデータでラベルを入れ替えると真の分散が保たれ，較正された p 値が得られる．

文脈化トークン埋め込みの分布は不明であるため，非パラメトリックな検定を用いたい．このための方策として NLP でも置換検定を用いることが行われてきたが，計算コストが大きい問題が指摘されている [6]．さらに，文脈化トークンベクトルの 2 つの埋め込み集合の散らばりを比較する目的で置換検定をそのまま適用すると，第一種過誤が増大し，実際には広がりが等しい場合でも「有意差あり」と

判定してしまう問題があることを指摘する．

直感的説明: 図 1 に，この増大の直観的理由を示す．単位正規化された埋め込みベクトルが球面上に分布すると考え，黒点・白点で 2 つの単語タイプのトークン点群を表す（図 (a)）．両群の散らばり自体は同程度でも，平均方向が異なるため球面上の領域がずれている．ここで，黒点・白点の散らばり度合いが等しいかの統計的検定を行いたい．置換検定は，単純にこの黒丸と白丸をランダムに入れ替えた点群を作り（置換），入れ替えた点群の中で目的の統計量（本研究の場合は後に定義する黒丸・白丸の各「散らばり」度合い）を計算することを多数繰り返し，その中で当初のデータから取った統計量が得られることが稀であるかを確認するものである．黒丸と白丸をランダムに入れ替えたものを母集団のようにみなして，その中での稀さを見るものである．

さて，ここで平均ベクトルを動かさずにラベルを素朴に入れ替え（図 (b)），黒点の散らばり度合いだけを見てみよう．黒点が左右の 2 領域に配置されるため散らばりが増大し，帰無分布の構成が歪む．その結果，真に散らばりが等しい状況でも帰無仮説が棄却され，偽陽性（Type-I Error）が増えてしまう．黒丸と白丸の散らばり度合いを真に確認ずるためには，まず図 1 下段の (c) のように 2 群を整列させた後，次に黒と白のラベルを入れ替えて図 1(d) のように検定を行う必要があることがわかる．

本研究では，標準的な分散統計量に対する置換検定を行う前に，2 つの点群を**平均方向だけ**整列化することでこの問題に対処する．具体的には，2 群の単位平均方向を $\hat{\mu}_X = \mu_X/\|\mu_X\|_2$，$\hat{\mu}_Y = \mu_Y/\|\mu_Y\|_2$ とし，次で定まる Householder 行列を構成する：

$$H = I - 2\frac{vv^\top}{v^\top v}, \qquad v = \hat{\mu}_X - \hat{\mu}_Y, \tag{1}$$

このとき $H\hat{\mu}_X = \hat{\mu}_Y$ が成り立つ．H を X の全ベクトルに適用すると，ノルムと X 郡内の相対的幾何関係を保ったまま平均方向の不一致だけを除去できる．図 1(c) のように変換後は両群が同一領域に重なるため，ラベル入替え（図 (d)）は同一領域内での交換となり「散らばり」の増大が起きにくい．結果として，置換検定は意味差（平均方向差）ではなく広がり（分散差）をより直接に標的化し，名目有意水準を保つ較正された p 値が得られる．さらに置換検定を現実的にするため「散らばり」度合いが埋め込みベクトルの線形演算で計算できることを活かし，置換検定と線形代数演算をバッチ化した GPU 実装を導入し，CPU に対して 23 倍の高速化を得た．

本研究の主な貢献は以下の通りである：

- 平均方向により Householder 整列化置換検定を提案し，理論的性質と実験により第一種過誤が 62.2%低減することを示した．
- 埋め込み集合統計量に対する置換検定を GPU で高速化する実装戦略を提示し，素朴な CPU 実装に対して 23 倍の高速化を達成した．
- 平均ベクトルを揃えることで言語を超えて「意味の広さ」の差を初めて検定し，有意に差があるものとそうでないものがあることを示した．

2 関連研究

文脈化埋め込みに基づく文脈多様性 語の意味的広さを考慮して文脈化「されていない」GloVe などの単語埋め込みに意味の広さを与えようとする試みはあるが [7]，結局，「広さ」と言われている空間上の点が具体的にコーパス上のどの用例に対応しているのかわからない問題がある．BERT [5] に代表される文脈化埋め込みは，トークンごとに異なる埋め込みを生成できるので，埋め込み空間上の点とコーパス上の用例を対応付けながら語タイプ統計量を定義することを可能にした．2025 年には BERT に最新の技術を詰め込みより高精度な ModernBERT[8] が提案されている．ACL2025 の Outstanding Paper となった [1] では，こうした文脈多様性を意味の近似量として用いる Zipf 則の拡張を提案している．もちろん，文脈化単語埋め込みは広がり方が方向によって違う異方性を持つなど，単純に「散らばり」を 1 つの値で表すことが難しい [9]．そこで，語単位ではなく頻度（あるいはランク）でビン分けし，ビン内平均を用いる集約的分析が用いられ [10]，[1] もこの方法を踏襲している．多言語においてビン平均では法則が支持される一方，個々の語の埋め込みベクトル集合からの語義数予測は著しく困難であると報告 [11] されている．本研究では，この困難とされているビン化なしの設定で，2 語の広さの差を埋め込みベクトルから検定する手法を提案するものである．

なお，置換検定については自然言語処理で近年厳密な方法 [12] が提案されているが，**これは結果が離散値の場合のみ適用可能な手法であり，埋め込みベクトルという連続値を扱う本研究の目的には適用できない**ことに注意が必要である．

語の散らばりの指標: Mean Resultant Length

[1] では，語タイプ w のコーパス中の d 次元文脈化埋め込みベクトル $\mathbf{h}_{w,i} \in \mathbb{R}^d$ 集合に対し，まず各ベクトルをノルム 1 になるよう正規化したベクトル集合 $\mathbf{x}_{w,i} = \mathbf{h}_{w,i}/\|\mathbf{h}_{w,i}\|_2 \in \mathbb{S}^{d-1}$ を考える．そして，この語の平均ベクトル $\bar{\mathbf{x}} = \frac{1}{n}\sum_i \mathbf{x}_i$ のユークリッドノルム $l := \|\frac{1}{n}\sum_i \mathbf{x}_i\|$ を意味の広がりの近似量とみなす．l は Mean Resultant Length (MRL) と呼ばれ，一般に方向ベクトルの散らばり度合いを与える方向統計の基礎的な統計量であり，分散と異なり小さいほど散らばり度合いが大きい．[1] では，頻度や語義数との相関を取るため，von-Mises Fisher(vMF) 分布の集中度パラメタの近似値 $\kappa \approx \frac{l(d-l^2)}{1-l^2}$ の逆数 $v = \frac{1}{\kappa}$ が平均的には WordNet 語義数と相関することを多数の実験を通じて示している．しかし，初頭的な計算により $[0,1]$ の範囲で d 固定の v は l の単調関数であることが示せるので，l と v の大小関係は逆だけなので **2 語の検定の目的では l の大小さえわかれば良い**．重要な点として l と大小関係が一致する l^2 は各埋め込みベクトルの線形計算で求められるので GPU を活用することができる．この点が，MRL に基づく「意味の広さ」の指標について，GPU を用いて置換検定を高速化できる理由である．

3 提案手法

Householder 整列化置換検定　語タイプ w の文脈化埋め込みベクトル $\mathbf{h}_{w,i} \in \mathbb{R}^d$ から，$\mathbf{x}_{w,i} = \mathbf{h}_{w,i}/\|\mathbf{h}_{w,i}\|_2 \in \mathbb{S}^{d-1}$ を得て，方向分散を意味の広がりの近似量とみなす．2 語 u, k の標本を $X = \{\mathbf{x}_i\}_{i=1}^n$, $Y = \{\mathbf{y}_j\}_{j=1}^m$（必要なら共通サイズにサブサンプル）とし，帰無は「分散は等しいが平均方向は異なり得る」とする．平均方向差で置換の交換可能性が崩れるため，Householder 反射で X の平均方向を Y へ整列化する：$\bar{\mathbf{x}} = \frac{1}{n}\sum_i \mathbf{x}_i$, $\bar{\mathbf{y}} = \frac{1}{m}\sum_j \mathbf{y}_j$, $\hat{\boldsymbol{\mu}}_x = \bar{\mathbf{x}}/\|\bar{\mathbf{x}}\|_2$, $\hat{\boldsymbol{\mu}}_y = \bar{\mathbf{y}}/\|\bar{\mathbf{y}}\|_2$, $\mathbf{u} = (\hat{\boldsymbol{\mu}}_x - \hat{\boldsymbol{\mu}}_y)/\|\hat{\boldsymbol{\mu}}_x - \hat{\boldsymbol{\mu}}_y\|_2$, $\mathbf{H} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^\top$, $\mathbf{x}'_i = \mathbf{H}\mathbf{x}_i$（$Y$ は不変）．分散統計は $\mathrm{MRL}r(A) = \|\frac{1}{|A|}\sum_{\mathbf{a}\in A}\mathbf{a}\|_2$ を用い，単調写像 $\kappa = g_d(r)$ を介して広がり $v(A) = 1/g_d(r(A))$ を定義，統計量は $T_{\mathrm{obs}} = \log v(X') - \log v(Y)$（片側は $T_{\mathrm{obs}} > 0$）．置換では $Z = X' \cup Y$ を B 回ランダムに (n, m) へ再割当てして $T^{(b)}$ を計算し，$p = (1 + \sum_{b=1}^B \mathbb{I}[T^{(b)} \geq T_{\mathrm{obs}}])/(B+1)$ を返す．整列化のラベル依存はクロスフィット（半分で $\mathbf{H}$ 推定→残り半分で検定）で緩和する．

GPU による置換推論の高速化　置換は素朴に $O(BNd)$（$N = n + m$）だが，プール行列 $\mathbf{X} \in \mathbb{R}^{N\times d}$ と，置換 b を表す符号 $\mathbf{s}^{(b)} \in \{+1, -1\}^N$（+1 が n 個）で一括化できる．$\mathbf{S} \in \{+1, -1\}^{B\times N}$ を用意し，$\mathbf{t} = \mathbf{1}^\top\mathbf{X}$，$\mathbf{U} = \mathbf{S}\mathbf{X}$（単一 GEMM）を計算すると，各置換の群和は $\boldsymbol{\sigma}_{1,2}^{(b)} = (\mathbf{t} \pm \mathbf{U}_{b,:})/2$ で復元でき，平均・MRL・$T^{(b)}$ はバッチ要素演算で計算できる．$\mathbf{S}$ は B_0 行ずつ生成して超過カウントのみ更新すればメモリは $O(Nd + B_0 d)$．Householder は行列を明示せず $\mathbf{x}' = \mathbf{x} - 2\mathbf{u}(\mathbf{u}^\top\mathbf{x})$（バッチなら $\mathbf{X}' = \mathbf{X} - 2(\mathbf{X}\mathbf{u})\mathbf{u}^\top$）で高速に適用できる．

4 実験

主実験　**検定はそもそも意味の広がりの差が微妙であるときに用いられるものであるので，2 語の間の「意味の広さ指標」の差が小さい語のペアの間での性能を計測する必要がある**．本研究では次の方法でこのペアを作成する．まず，すべての語について前述の MRL を求め，その昇順のランキングを作る．そして，順位差 (gap) を用いて近い順位の語のペアを測る．例えば [1] では 100 語でビン化している．

まずは WordNet 語義数を意味の広さの真の指標とし，WordNet 語義数が同じ語ペアを誤って棄却してしまう率（第一種の過誤）と，WordNet 語義数が異なるものをどれだけ検出できるか（検出率）が基本的な検定の性能評価となる．BNC[13] に ModernBERT[8] を適用し，MRL 指標順位差 10 位以内かつ WordNet[14] の語義数に差が無い 500 ペアを「意味の広さに差がない」ペアとして第一種の過誤を，語義数に差がある 500 ペアで「検出力」を評価した．比較は (i) 整列化なし（Baseline）と (ii) Householder 反射で整列化後に検定（Proposed）である．表 1 に結果を示す．差があるかを見る検定である以上，第一種過誤大きいものは使いづらい．提案手法により，$p < 0.01$ のとき第一種の過誤が 2.09% から 0.79% まで 62%削減できていることがわかる．検出力は $p < 0.01$ のときには既存手法と変わらない．これにより提案手法の有用性が示された．

適合率実験　更に順位差 1-10 において詳細な分析を行った．既存法と提案法で検定で有意になった語ペアのうち，実際に WordNet 語義数が異なっているものの比率を正解だと思うと，NLP に馴染み深い「適合率 (precision)」で評価することができる．やはり BNC と ModernBERT で $\alpha = 0.01$，各 gap で 300 ペアについて実験した結果を表 2 に示す．

Proposed は概ね第一種過誤を悪化させず，Gap=5,10 では第一種過誤を下げつつ適合率も改善した（例：Gap=10 で第一種過誤 0.040→0.027,

有意水準	手法	第一種過誤	検出力
$p \le 0.05$	Baseline	3.40%	1.91%
	Proposed	**1.57%**	1.50%
$p \le 0.01$	Baseline	2.09%	1.29%
	Proposed	**0.79%**	1.29%

表 1　BNC × ModernBERT における第一種過誤と検出力

	第一種過誤（↓）		適合率（↑）	
Gap	Baseline	Proposed	Baseline	Proposed
1	0.013	**0.013**	0.250	**0.250**
2	0.010	**0.010**	0.667	**0.667**
3	0.007	**0.007**	0.500	**0.500**
4	**0.020**	0.030	0.333	**0.444**
5	0.017	**0.013**	0.600	**0.750**
6	0.013	**0.013**	0.750	**0.750**
7	0.020	**0.020**	0.500	**0.500**
8	0.013	**0.013**	0.500	**0.500**
9	0.027	**0.023**	**0.375**	0.286
10	0.040	**0.027**	0.417	**0.625**

表 2　分散順位差 1–10 における第一種過誤と適合率の比較（$\alpha \le 0.01$）．太字は良い方（第一種過誤は低い方，適合率は高い方）を示す．

適合率 0.417→0.625）．Gap=1–3 では両手法とも名目水準付近である．

ModernBERT 以外の埋め込みベクトル実験　他言語・他モデルでも同傾向かを確認するため，BNC ＋ BERT-tiny，BNC ＋ ModernBERT，BCCWJ[15] ＋ BERT-large の 3 条件を評価した．各条件で 500 ペアをサンプルし，置換回数 $B = 5{,}000$，$\alpha = 0.05$ で検定し，棄却率を比較した．表 3 より，Gap=50 では Proposed の棄却率が一貫して低く（2.2–2.4%），Baseline（3.2–4.0%）より保守的である．一方 Gap=100 では Proposed がやや低く（4.4–6.2%），整列化は第一種過誤低減と引き換えに検出力がやや小さくなる．

計算速度実験　BCCWJ ＋ BERT-large，Gap=50，$B = 20{,}000$ で CPU/GPU 実装を比較した（表 4）．GPU は 1 置換あたり 0.069 ms で，CPU（1.588 ms）に対して約 23 倍高速である．GPU は H100 を 2 台搭載した環境で実験した．20,000 回の置換検定で GPU メモリの使用量は 750MB 程度だった．

言語・コーパスを超えた意味の広さの検定　**本研究の手法は平均ベクトルの影響を捨象するため多言語埋め込みを複数のコーパスに適用することで言語・コーパスを超えた意味の広さの違いを置換検定できる．**言語が違えば当然，ベクトル空間が同じでも当然，語のベクトル集合が向いている方向が異なるため，置換検定を行っても図 1 の説明のように，この置換検定は語の平均を整列させる効果なしでは意味をなさない．著者の知る限り同様の研究はない．BNC と BCCWJ の全体に対して bert-base-multilingual-cased を共通して用い，簡単な 20 語の単語ペアで検定を行った．その結果，$p < 0.01$ で統計的有意に日本語のほうが意味が広いペアとして**山-mountain**，**木-tree**，**秋-autumn**，**油-oil**，**国-country** が抽出され，逆に統計的有意に英語のほうが意味が広いペアとして**業-work** が抽出された．次の語ペアは文脈化埋め込みベクトル集合の MRL 指標に言語間で差があったが統計的有意にならなかった：年-year, 月-month, 夏-summer, 外-outside, 塔-tower, 町-town, 村-village, 氷-ice, 恐-fear, 罪-sin.

コーパス	**モデル**	**Gap**	**Baseline**	**Proposed**
BNC	BERT-tiny	50	0.034	0.024
		100	0.064	0.044
BNC	ModernBERT	50	0.040	0.022
		100	0.090	0.062
BCCWJ	BERT-large	50	0.032	0.022
		100	0.090	0.046

表 3　帰無（gap=50）および対立（gap=100）条件における棄却率．Baseline は整列化なしの標準置換検定，Proposed は提案整列化を適用した検定である．

実装	1 置換あたり時間	総時間（$B = 20{,}000$）
CPU	1.588 ms	31,750 ms
GPU	0.069 ms	1,377 ms

表 4　CPU 実装と GPU 実装の実行時間比較．

5　結論

本研究は，平均方向を揃えて 2 群の語の意味の広がりだけを GPU で高速に置換検定する手法を提案した．これにより 62.2%の第一種の過誤低減，23 倍の高速化を実現した．また，言語を超えた意味の広がりの検定を知る限り初めて行った．

本研究は文脈化トークン埋め込みベクトル集合を扱ったが文ベクトル集合にも技術的にはそのまま適用できる．文ベクトルでは異方性の問題が大きい[9] ことが指摘されているが，この問題を解決し知識の広がりを人手判断に依らずに定量化する基盤を構築することが将来の課題である．

謝辞

本研究は，科学技術振興機構さきがけ研究費 (JPMJPR2363)，JSPS 科研費 22K12287 の支援を受けた．

参考文献

[1] Ryo Nagata and Kumiko Tanaka-Ishii. A new formulation of Zipf's meaning-frequency law through contextual diversity. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proc. of ACL**, July 2025.

[2] Francesco Periti and Nina Tahmasebi. A systematic comparison of contextualized word embeddings for lexical semantic change. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **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, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[3] Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3960–3973, Online, July 2020. Association for Computational Linguistics.

[4] Andrey Kutuzov and Mario Giulianelli. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, **Proceedings of the Fourteenth Workshop on Semantic Evaluation**, pp. 126–134, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proc. of NAACL-HLT**, June 2019.

[6] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In Iryna Gurevych and Yusuke Miyao, editors, **Proc. of ACL**, July 2018.

[7] Andrea Vallebueno, Cassandra Handan-Nader, Christopher D Manning, and Daniel E. Ho. Statistical uncertainty in word embeddings: GloVe-V. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proc. of EMNLP**, November 2024.

[8] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proc. of ACL**, July 2025.

[9] 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, **Proc. of EMNLP-IJCNLP**, November 2019.

[10] Francis Bond, Arkadiusz Janz, Marek Maziarz, and Ewa Rudnicka. Testing Zipf's meaning-frequency law with wordnets as sense inventories. In Piek Vossen and Christiane Fellbaum, editors, **Proc. of GWC**, July 2019.

[11] Francis Bond. Adding audio to wordnets. In Chiara Zanchi, Luca Brigada Villa, Erica Biagetti, Alexandre Rademaker, Francis Bond, and German Rigau, editors, **Proceedings of the 13th Global Wordnet Conference**, pp. 185–191, Pavia, Italy, January 2025. Global Wordnet Association.

[12] Ran Zmigrod, Tim Vieira, and Ryan Cotterell. Exact paired-permutation testing for structured test statistics. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proc. of NAACL-HLT**, July 2022.

[13] BNC Consortium. The British National Corpus, XML edition, 2007.

[14] George A. Miller. WordNet: A lexical database for English. **Communications of the ACM**, Vol. 38, No. 11, pp. 39–41, 1995.

[15] 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.

[16] Christos Xypolopoulos, Antoine Tixier, and Michalis Vazirgiannis. Unsupervised word polysemy quantification with multiresolution grids of contextual embeddings. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, **Proc. of EACL**, April 2021.

[17] Hiroaki Yamagiwa and Hidetoshi Shimodaira. Norm of mean contextualized embeddings determines their variance. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, **Proc. of COLING**, January 2025.

[18] Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, **Proc. of EMNLP**, July 2004.

[19] Yvette Graham, Nitika Mathur, and Timothy Baldwin. Randomized significance tests in machine translation. In Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, editors, **Proc. of WMT**, June 2014.

[20] Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In Jun'ichi Tsujii, James Henderson, and Marius Paşca, editors, **Proc. of EMNLP-CoNLL**, July 2012.

[21] Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. HyperLex: A large-scale evaluation of graded lexical entailment. **Computational Linguistics**, Vol. 43, No. 4, pp. 781–835, December 2017.

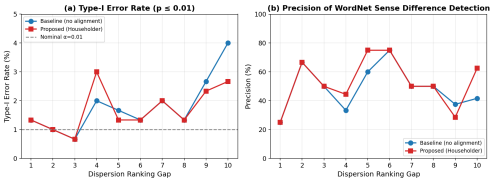

図 2　分散順位差 1–10 における (a) 第一種過誤率と (b) 適合率．破線は $\alpha = 0.01$．

A　提案手法のアルゴリズム表記

提案手法は本文内に完全に記述されているが，アルゴリズム形式で記述する．

B　その他の関連研究

多義性を幾何的に捉える試み [16] や，平均ノルムと分散の関係分析 [17] があるが，較正された仮説検定を主眼とするものではない．

置換検定や近似ランダム化検定は，機械翻訳を中心に NLP 評価で長く用いられてきた [18, 19]．検定選択や報告に関する実務的指針が整理されている [6, 20]．語義数ではなく単語間の含意の強さを評価するタスクは提案されてきた [21]．しかし，結局こうしたタスクでは文中の語の含意関係を正確に捉えることが目的であり，検定などを通じた語の「広がり」を論じるものではない．

Algorithm 1 ハウスホルダー反射による整列付き置換検定

Require: 単位ノルムに正規化された埋め込み $X = \{\mathbf{x}_i\}_{i=1}^{n}$，$Y = \{\mathbf{y}_j\}_{j=1}^{m}$；置換回数 B；有意水準 α

Ensure: 帰無仮説 $H_0:\ \mathrm{disp}(X) = \mathrm{disp}(Y)$ を検定するための p 値

【ステップ 1】標本平均ベクトルと平均方向の計算

$\bar{\mathbf{x}} \leftarrow \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i;\quad \bar{\mathbf{y}} \leftarrow \frac{1}{m}\sum_{j=1}^{m}\mathbf{y}_j$

$\hat{\boldsymbol{\mu}}_x \leftarrow \bar{\mathbf{x}}/\|\bar{\mathbf{x}}\|_2;\quad \hat{\boldsymbol{\mu}}_y \leftarrow \bar{\mathbf{y}}/\|\bar{\mathbf{y}}\|_2$

【ステップ 2】ハウスホルダー反射の構成

$\mathbf{u} \leftarrow (\hat{\boldsymbol{\mu}}_x - \hat{\boldsymbol{\mu}}_y)/\|\hat{\boldsymbol{\mu}}_x - \hat{\boldsymbol{\mu}}_y\|_2$

【ステップ 3】X を Y の平均方向に整列

for $i = 1$ から n まで **do**

　$\mathbf{x}'_i \leftarrow \mathbf{x}_i - 2\mathbf{u}(\mathbf{u}^{\top}\mathbf{x}_i)$

end for

$X' \leftarrow \{\mathbf{x}'_i\}_{i=1}^{n}$

【ステップ 4】観測検定統計量の計算

$r_{X'} \leftarrow \left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}'_i\right\|_2;\quad r_Y \leftarrow \left\|\frac{1}{m}\sum_{j=1}^{m}\mathbf{y}_j\right\|_2$

$T_{\mathrm{obs}} \leftarrow \log(1/g_d(r_{X'})) - \log(1/g_d(r_Y))$

【ステップ 5】整列後データに対する置換検定

$Z \leftarrow X' \cup Y$　　▷ 整列後の標本をプール

$c \leftarrow 0$　　▷ $T^{(b)} \geq T_{\mathrm{obs}}$ となった回数

for $b = 1$ から B まで **do**

　Z を大きさ (n, m) となるようにランダムに二分割し，$(X^{(b)}, Y^{(b)})$ を得る

　$r_1^{(b)} \leftarrow \left\|\frac{1}{n}\sum_{\mathbf{z}\in X^{(b)}}\mathbf{z}\right\|_2;\quad r_2^{(b)} \leftarrow \left\|\frac{1}{m}\sum_{\mathbf{z}\in Y^{(b)}}\mathbf{z}\right\|_2$

　$T^{(b)} \leftarrow \log(1/g_d(r_1^{(b)})) - \log(1/g_d(r_2^{(b)}))$

　if $T^{(b)} \geq T_{\mathrm{obs}}$ **then**

　　$c \leftarrow c + 1$

　end if

end for

【ステップ 6】p 値の計算

$p \leftarrow (1 + c)/(B + 1)$

p を返す