文埋め込み方向介入評価における符号一貫性の擬似生成

苅山 湊 [1]
[1] 東京大学
so-kari@g.ecc.u-tokyo.ac.jp

概要

文埋め込み表現に対し，教師なし辞書学習で得られた因子方向に沿った介入を行うと，多数の入力で下流指標が同符号に変化する場合がある．本稿はこの挙動を **符号一貫性** として定式化し，因子方向の意味性とは独立に，埋め込みの異方性と評価設計の相互作用により高く観測され得ることを示す．日本語文埋め込みと MIRACL-ja 検索を用いた介入実験の結果，辞書学習因子方向とランダム方向の符号一貫性は，whitening や主成分除去による幾何的操作によって同水準まで縮退することが確認された．これらの結果は，因子方向の意味性を主張するには，表現幾何および評価設計を制御した対照実験が不可欠であることを示唆する．

1 はじめに

文埋め込みは，情報検索や文類似度推定をはじめとする多くの自然言語処理タスクにおいて，下流モデルの性能を規定する基盤的表現である．近年は性能比較にとどまらず，埋め込み空間における特定の**方向**を概念や属性の軸として解釈し，その方向に沿った表現の摂動（介入）によって下流タスクの挙動を分析・制御しようとする枠組みが提案されてきた．

このような介入評価において，介入後の指標が多数の入力で同符号に変化する挙動は，方向が一貫した意味的効果を持つことの根拠として解釈される場合がある．しかし，本稿は，この「同符号で動く」挙動が必ずしも方向の意味性を反映するとは限らず，方向の意味とは独立に生成され得る可能性に着目する．

具体的には，(1) 埋め込み表現が本質的に持つ**異方性**，および (2) 検索評価における**設計要因**（候補集合の固定やランキング指標の非線形性）の相互作用により，符号の偏りが見かけ上強く観測され得るという仮説を立てる．この仮説の下，本稿では摂動に対する下流指標変化の符号偏りを要約する量として**符号一貫性**指標を定式化し，その振る舞いを制御実験によって検証する．

検証にあたっては，日本語文埋め込みと MIRACL-ja 検索を対象とし，教師なし辞書学習により得られた因子方向への介入を，ランダム方向などの対照条件と比較する．さらに，PCA whitening や主成分除去といった操作により表現幾何を明示的に制御し，符号一貫性指標が方向の意味ではなく，表現幾何や評価設計にどの程度依存するかを分析する．

本稿の貢献は以下の三点に要約される：

- 下流指標変化の符号偏りのみに着目した符号一貫性指標を導入し，介入評価で暗黙的に用いられてきた挙動を明示的な分析対象として定式化する．
- 辞書学習因子方向とランダム対照を同一条件で比較することで，PCA whitening や主成分除去により両者の符号一貫性が同水準に縮退し得る条件を示す．
- 候補集合を条件内で固定した検索評価設定において，Recall を大きく損なわずに符号一貫性が大きく変動し得ることを示し，方向介入評価の解釈上の注意点を整理する．

2 関連研究

文埋め込みの異方性と後処理　文・単語埋め込みが強い異方性を持ち，内積やコサイン類似度に基づく評価を歪め得ることは，これまでに広く指摘されてきた．SIF に基づく文埋め込みでは，共通成分（主成分）を除去する後処理が提案され [1]，All-but-the-Top（ABTT）では，平均ベクトル除去および上位主成分除去が簡便かつ有効な後処理として整理された [2]．また，文脈化表現においても，層ごとに異方性が存在することが報告されている [3]．

これらの研究の主眼は，類似度尺度としての性能改善や表現分布の等方化にあり，後処理操作が**方向**

介入評価における指標応答にどのような影響を与えるかについては，体系的な検証は限られている．本稿は，これらの操作を性能改善手段としてではなく，**介入評価における交絡要因を制御する操作**として位置付ける点に特徴がある．

方向介入と概念方向　埋め込み空間における特定方向を概念や属性と対応づけ，その方向への摂動がモデル出力に与える影響を評価する枠組みとして，TCAV が提案されている [4]．また，線形分類器を用いて表現が情報を保持しているかを診断する probing 手法は，内部表現解析の基本的手法として用いられてきた [5]．

一方で，probing の結果から直接的に行動的・因果的結論を導くことの限界も指摘されており，表現から特定情報を除去する介入を通じて「その情報が実際に用いられているか」を評価する amnesic probing が提案されている [6]．本稿は，この流れを踏まえつつ，**下流指標変化の符号の偏り**という，より単純かつ暗黙的に解釈されがちな量に着目し，その解釈上の注意点を明らかにする．

疎表現・辞書学習　辞書学習およびスパースコーディングは，高次元表現を少数の基底の線形結合として表す枠組みであり，K-SVD はその代表的な学習アルゴリズムである [7]．線形辞書学習は，回転不定性や符号対称性といった性質を本質的に持つため，学習された因子を意味軸として解釈する際には，評価設計や対照条件による検証が重要となる．本稿は，辞書学習因子方向とランダム対照を同一の介入評価枠組みで比較することで，この点を実証的に検討する．

情報検索ベンチマークと評価設計　文埋め込みを用いた検索評価では，BEIR [8] や MIRACL [9] などの大規模ベンチマークが広く用いられている．これらの設定では，候補集合生成と再ランキングの分離や，nDCG や Recall といった複数指標の併用が一般的である．本稿では MIRACL-ja を対象とした評価を行う．

3　手法

3.1　検索設定

クエリ q と文書 x の埋め込みをそれぞれ $\mathbf{e}(q), \mathbf{e}(x) \in \mathbb{R}^m$ とする．コサイン類似度（ℓ_2 正規化後の内積）でランキングを作り，評価指標 $S(\cdot)$ として nDCG@10 を用いる．

未介入のクエリ埋め込みに基づく上位 L 件を候補集合 $C(q)$ として固定し，介入後は $C(q)$ 内で再ランキングする．評価指標 $S(q;\mathbf{e}_q)$ として nDCG@10 を用い，以下では簡単のため $S(\mathbf{e}_q)$ と略記する．候補集合 $C(q)$ は各評価条件（後処理の有無）ごとに，その条件の未介入表現に基づいて作成し，条件内で固定する．

3.2　辞書学習と疎性

埋め込み集合 $\{\mathbf{e}_i\}$ に対し，辞書 $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K] \in \mathbb{R}^{m\times K}$ を学習する．各 $\mathbf{e}_i$ を

$$\mathbf{e}_i \approx \mathbf{D}\mathbf{a}_i \tag{1}$$

で近似し，係数 $\mathbf{a}_i$ は top-k_0 制約で疎に推定する．線形・符号対称な最小条件を維持し，追加の符号制約や非線形変換は導入していない．

3.3　方向介入

因子方向 $\mathbf{d}_j$ に沿ってクエリを加法摂動する：

$$\mathbf{e}_q^{(\pm j)} = \mathrm{norm}\left(\mathbf{e}(q) \pm \lambda \frac{\mathbf{d}_j}{\|\mathbf{d}_j\|_2}\right), \tag{2}$$

$\lambda > 0$ は介入強度，norm は ℓ_2 正規化である．文書側埋め込みは固定し，クエリ側のみを介入する．

3.4　符号一貫性（sign bias）指標

評価クエリ集合 $\mathbb{Q}$ に対し，因子 j の**符号付き**偏りを

$$\overline{\mathrm{Dir}}(j) = \frac{1}{|\mathbb{Q}|}\sum_{q\in\mathbb{Q}} \mathrm{sign}\Big(S(\mathbf{e}_q^{(+j)}) - S(\mathbf{e}(q))\Big), \tag{3}$$

で定義し（$\mathrm{sign}(0) = 0$），符号一貫性指標を

$$\mathrm{Dir}(j) = \left|\overline{\mathrm{Dir}}(j)\right| \tag{4}$$

とする．

3.5　対照実験

比較条件として，以下を用いる：

- **ランダム方向**：ノルムを整合したランダムベクトルで介入する．
- **ランダム因子**：同一辞書から原子を無作為に選ぶ．
- **PC1 除去（ABTT 1 成分版）**：埋め込みから第 1 主成分成分を除去して評価する．
- **PC1 方向介入**：介入方向を PC1 に一致させた単一方向で評価する．

これらを比較し，Dir が幾何や評価設計でどこまで変動するかを検証する．

4 実験設定

埋め込みモデル 文埋め込みには Ruri v3 [10] を用い，モデルの重みは全実験を通じて固定する．検索用入力には，モデルカードで推奨されている prefix（検索クエリ／検索文書）を付与する．

検索タスクと評価設定 主タスクとして MIRACL-ja（dev）検索 [9] を用いる．クエリは dev セットから seed 固定で 200 件をサブサンプルする．各評価条件（後処理の有無）ごとに未介入のクエリ埋め込みに基づいて各クエリの上位 $L = 1000$ 件を候補集合として作成し，条件内で固定する．介入後は同一候補集合内で再ランキングを行い，評価指標として nDCG@10 を算出する．

後処理の忠実性確認 後処理が埋め込みの意味的類似性に与える影響を確認するため，JSTS（JGLUE）[11] を用い，Spearman および Pearson 相関を報告する．

辞書学習と介入設定 辞書学習では辞書サイズを $K = 256$，反復回数を 15，乱数 seed を 0 に固定する．疎性は活性因子数 $k_0 \in \{4, 8, 16\}$ を比較し，一部の実験では dense 表現（$k_0 = K$）も評価する．介入強度は全実験で $\lambda = 0.5$ に固定する．

評価対象の因子は，評価クエリ集合 $\mathbb{Q}$ 上での平均活性度が高い順に上位 80 因子とする．ここで活性度は，各クエリ q に対する疎係数の絶対値 $|a_{q,j}|$ を用い，$\mathbb{E}_{q\in\mathbb{Q}}[|a_{q,j}|]$ により因子 j を順位付けする．対照条件として，辞書から無作為に選択したランダム因子 80，およびノルムを整合したランダム方向 80 を用いる．

Whitening の推定 PCA whitening は，辞書学習用コーパスから最大 30000 文を用いて推定し，得られた線形変換をクエリおよび文書の埋め込みの双方に適用する．

5 結果

5.1 Whitening の忠実性（Recall を維持）

表 1 に示す通り，PCA whitening は JSTS 相関および MIRACL の nDCG@10 を低下させた一方，Recall@100 は本設定では同値であった．なお，本節の IR 指標（nDCG@10, Recall@100）は全文書ランキングで算出している．すなわち PCA whitening は，正解文書の包含（Recall）を保ったまま，ランキング（nDCG）に影響し得る．以降ではこの性質を利用し，表現幾何の制御により符号一貫性がどの程度変化するかを検証する．

Metric	後処理なし	PCA whitening 後
JSTS Spearman	0.834	0.796
JSTS Pearson	0.871	0.855
MIRACL nDCG@10	0.130	0.110
MIRACL Recall@100	0.1850	0.1850

表 1 Whitening 前後における埋め込みの忠実性評価（JSTS valid, MIRACL-ja dev）．

5.2 Whitening 後は辞書因子と対照が同水準に縮退

図 1 に，PCA whitening 後・$k_0 = 8$ における符号一貫性 Dir(j) の分布を示す．平均は dict 0.020（std 0.013），rand_factor 0.019（std 0.012），rand_dir 0.015（std 0.010）であり，辞書因子と対照条件の分布が近い．

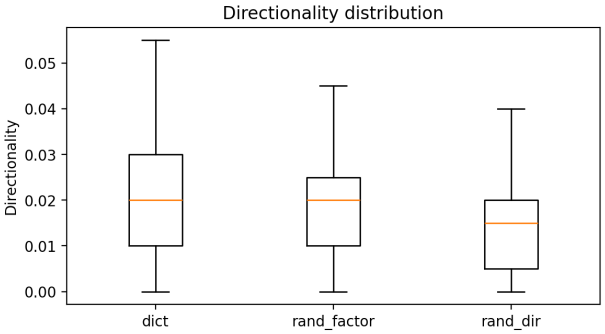

図 1 符号一貫性 Dir(j) = $|\mathbb{E}_q[\mathrm{sign}(\Delta S)]|$ の分布（$\Delta S := S(\mathbf{e}_q^{(+j)}) - S(\mathbf{e}(q))$；PCA whitening 後，$k_0 = 8$）．dict：活性上位 80 因子（$\mathbb{Q}$ 上の平均活性度に基づく），rand_factor：辞書から無作為選択，rand_dir：ノルム整合ランダム方向．

5.3 疎性と幾何制御：whitening による縮退

Condition	k_0	μ	σ
後処理なし	4	0.0581	0.016
後処理なし	8	0.0714	0.018
後処理なし	16	0.0774	0.019
後処理なし (dense)	256	0.0603	0.016
PCA whitening 後	4	0.0201	0.014
PCA whitening 後	8	0.0203	0.013
PCA whitening 後	16	0.0183	0.015
PCA whitening 後 (dense)	256	0.0199	0.014

表 2 符号一貫性指標のアブレーション結果（MIRACL-ja dev；dense は辞書を $k_0 = 4$ で学習し，符号化のみを $k_0 = K$ とした条件）

図 2 および表 2 に示す通り，後処理なしでは k_0 の増加に伴い Dir が上昇する一方，PCA whitening 後では k_0 に依らず Dir ≈ 0.02 に留まる．また，辞書を $k_0 = 4$ で学習して固定し，符号化のみを dense

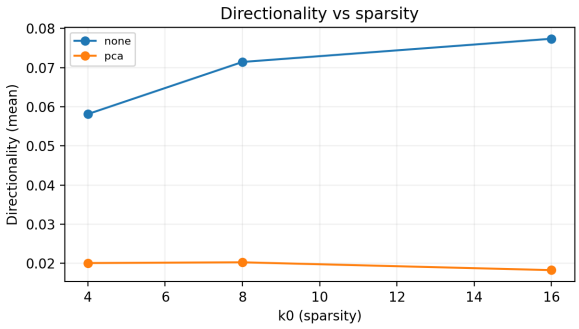

図 2　符号一貫性 Dir と疎性（活性因子数 k_0）の関係．後処理なしおよび PCA whitening 後の条件を示す．

（$k_0 = K$）にした条件でも，Dir の傾向は大きく変化しない（表 2）．

5.4　PC1 操作（1 成分除去でランダム水準）

後処理なし（$k_0 = 8$）における辞書因子の平均 Dir は 0.0714 であるが，PC1 を 1 成分除去すると 0.0241 まで低下し，ランダム方向平均 0.0244 と同水準になった（表 3）．一方で，PC1 方向介入は 0.0950 を示した．なお，因子ごとの $\cos(\mathbf{d}_j, \mathrm{PC1})$ と $\mathrm{Dir}(j)$ の相関は小さい（Pearson 0.013, Spearman 0.031）．

条件（後処理なし，$k_0 = 8$）	符号一貫性 Dir
辞書因子（baseline）	0.0714
PC1 除去（ABTT-1）	0.0241
ランダム方向（平均）	0.0244
PC1 方向介入	0.0950

表 3　PC1 に関する対照実験（MIRACL-ja dev；介入評価は候補集合を条件内で固定した再ランキング）

6　考察

本節では，観測された符号一貫性が，特定の方向（例：PC1）への整列を直接反映したものではなく，埋め込み空間の異方性と候補集合を条件内で固定した再ランキング評価という設計の下で擬似的に生成され得る機構について考察する．

後処理なし（$k_0 = 8$）では辞書因子の平均 Dir = 0.0714 が観測された一方，PC1 を 1 成分除去すると Dir = 0.0241 まで低下し，ランダム方向平均 0.0244 と同水準となった（表 3）．さらに PCA whitening 後には k_0 に依らず Dir ≈ 0.02 に留まり，辞書因子と対照条件の分布も近い（図 1，表 2）．

ここで，正規化内積に基づくスコア

$$s(q,x) = \langle \hat{\mathbf{e}}(q), \hat{\mathbf{e}}(x) \rangle, \quad \hat{\mathbf{e}}(q) = \mathbf{e}(q)/\|\mathbf{e}(q)\|_2 \tag{5}$$

を用い，クエリへの介入を $\hat{\mathbf{e}}'(q) = \mathrm{norm}(\mathbf{e}(q) + \lambda \mathbf{u})$ とする．λ が十分小さいとき，文書 x に対するスコア変化は一次近似により

$$\Delta s(q,x) \approx \frac{\lambda}{\|\mathbf{e}(q)\|_2} \langle \mathbf{u}_\perp(q), \hat{\mathbf{e}}(x) \rangle, \qquad \mathbf{u}_\perp(q) := \mathbf{u} - \langle \mathbf{u}, \hat{\mathbf{e}}(q) \rangle \hat{\mathbf{e}}(q) \tag{6}$$

と表される（高次項は $O(\lambda^2)$）．

これは，正規化制約下で介入方向の接空間成分のみがスコア変化に寄与することを示している．

異方性が強い場合，文書埋め込みは支配的成分方向 $\mathbf{v}_1$（例：PC1）を含み，

$$\hat{\mathbf{e}}(x) = \alpha_x \mathbf{v}_1 + \boldsymbol{\varepsilon}_x \tag{7}$$

と分解できる（α_x の分散が相対的に大きい状況を想定する）．このとき (6) より，$\langle \mathbf{u}_\perp(q), \mathbf{v}_1 \rangle$ に比例する成分が候補集合内の多数の文書に対して系統的に寄与し得る．

候補集合を条件内で固定した再ランキング評価では，この系統的なスコア変化が順位境界付近の入替を通じて ΔnDCG@10 の符号として集計されやすく，その結果として符号一貫性が高く観測され得ると考えられる．重要なのは，この現象が「各因子が PC1 に強く整列している」ことを必ずしも意味しない点である．実際，因子方向と PC1 の整列度 $\cos(\mathbf{d}_j, \mathbf{v}_1)$ と $\mathrm{Dir}(j)$ の相関は小さく（Pearson 0.013, Spearman 0.031），符号一貫性の高さを単一主成分への整列のみで説明することは困難である．

PCA whitening は共分散を等方化する線形変換であり，(7) における α_x の偏りを相対的に低減すると考えられる．その結果，(6) に基づく内積変化の符号が対称化され，辞書因子・ランダム方向のいずれに対しても符号一貫性が同水準へ縮退し得るという観測（図 1，表 2）と整合的である．

7　結論

本稿は，下流検索指標変化の符号偏りを要約する符号一貫性指標を導入し，その値が方向の意味性そのものを保証しないことを示した．MIRACL-ja 検索における介入実験では，後処理なしで辞書因子が Dir ≈ 0.07 を示す一方，PC1 を 1 成分除去すると Dir ≈ 0.02 まで低下してランダム方向と同水準となり，PCA whitening 後も Dir ≈ 0.02 に縮退した．

以上より，本設定で観測された符号一貫性の高さは，特定方向（PC1）への整列によって一意に説明されるというより，埋め込み空間の異方性と候補集合を条件内で固定した再ランキング評価という設計の下で擬似的に高く観測され得ることが示唆される．

参考文献

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

[2] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In **International Conference on Learning Representations (ICLR)**, 2018. arXiv:1702.01417.

[3] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In **Proceedings of EMNLP-IJCNLP**, 2019. arXiv:1909.00512.

[4] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In **Proceedings of the 35th International Conference on Machine Learning (ICML)**, 2018. arXiv:1711.11279.

[5] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2016.

[6] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 160–175, 2021.

[7] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. **IEEE Transactions on Signal Processing**, Vol. 54, No. 11, pp. 4311–4322, 2006.

[8] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In **NeurIPS Datasets and Benchmarks Track**, 2021.

[9] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. **Transactions of the Association for Computational Linguistics**, Vol. 11, pp. 1114–1131, 2023.

[10] Hayato Tsukagoshi and Ryohei Sasano. Ruri: Japanese general text embeddings, 2024.

[11] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)**, pp. 2957–2966, 2022.