低ランク幾何変換によるテキスト埋め込みベクトルの日本語クラスタリング適応と評価

須賀 圭一 [1]　市川 佳彦 [1]
[1] 株式会社 Insight Edge
{keichi.suga,yoshihiko.ichikawa}@insightedge.jp

概要

本研究では，商用 API の汎用埋め込みにおける日本語クラスタリング性能を高める軽量な後処理の評価を実施した。手法としては，「恒等変換に近い低ランク線形変換」を学習し，空間の幾何構造を微調整する。実験の結果，少数のパラメータでクラスタリング精度をある程度向上させつつ，他タスクの性能低下を抑制できることが示された。また，アブレーション分析を通じ，表現力と汎用性のトレードオフや埋め込み空間の幾何学的特性に関する考察も行った。

1　はじめに

商用テキスト埋め込み API は高品質な汎用表現を提供するが，特定タスクに対しては最適化されていない。モデル内部へのアクセスが制限される商用 API では，ファインチューニングは不可能であり，軽量な後処理による適応が求められる。本研究では，ベース埋め込み e に対し，恒等行列への低ランク更新 $W = I + UV^\top$（$U, V \in \mathbb{R}^{d\times r}, r \ll d$）による幾何変換 $z = We$ が提案されていることを活用し [1, 2]、直交性正則化 $\mathcal{R}_{\text{orth}} = \|W^\top W - I\|_F^2$[3] により，汎用性を維持しつつ日本語クラスタリング性能を向上させたい場合の評価を行った。

2　手法：低ランク幾何変換

2.1　低ランク線形変換の定式化

商用埋め込み API $f_{\text{base}} : \mathcal{X} \to \mathbb{R}^d$ の出力 $e = f_{\text{base}}(x)$ に対し，変換行列 $W = I + UV^\top$（$U, V \in \mathbb{R}^{d\times r}$）を適用し，$z = We = e + UV^\top e$ を得る。ここで $r \ll d$ はランクであり，パラメータ数は $2dr$ に抑えられる。この変換は，$h = V^\top e \in \mathbb{R}^r$ による射影と，$z = e + Uh$ による補正の 2 段階と解釈でき，埋め込み空間の少数の重要な方向に沿って幾何構造を微調整する。計算量は $O(dr)$ であり，フルランクの変換（$O(d^2)$）に比べ大幅に軽量である。

2.2　損失関数と正則化

クラスタリング・分類タスクに対して、教師ありコントラスト損失 (Supervised Contrastive Loss)[4] を用いる。コントラスト学習によるクラスタリング手法は近年活発に研究されており [5, 6]，本研究でもこのアプローチを適用する：

$$\mathcal{L}_{\text{supcon}} = \sum_{i=1}^{N} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp\left(\tilde{z}_i^\top \tilde{z}_p/\tau\right)}{\sum_{a\neq i} \exp\left(\tilde{z}_i^\top \tilde{z}_a/\tau\right)}, \quad (1)$$

ここで $P(i) = \{j \mid j \neq i, y_j = y_i\}$ は同一ラベルを持つ正例の集合，$\tilde{z}$ は L2 正規化済み埋め込み，τ は温度パラメータである。この損失は，同一クラスのサンプルを埋め込み空間内で近づけ，異なるクラス間を遠ざけることで，クラスタリング性能の改善を意図したものである。

さらに、汎用性を維持するため，以下の正則化項を導入する：

$$\mathcal{R}_{\text{id}} = \|U\|_F^2 + \|V\|_F^2, \quad (2)$$

$$\mathcal{R}_{\text{orth}} = \|W^\top W - I\|_F^2. \quad (3)$$

恒等性正則化 $\mathcal{R}_{\text{id}}$ はパラメータのノルムを抑制し，直交性正則化 $\mathcal{R}_{\text{orth}}$ は W がほぼ直交変換となるよう誘導する。これにより変換がほぼ恒等近傍になるようにする。

検索タスク向け補助損失　提案変換の主目的はクラスタリング性能の向上であるが，検索 (Retrieval) タスクの性能維持も必要となる。そこで，オプションとして検索タスクに対する InfoNCE(Informative Neighborhood Contrastive Estimation) 損失 [7] を導入する。検索タスクにおける訓練例を，クエリ q，正例文書 d^+，負例文書 $\{d_1^-, \ldots, d_K^-\}$ から構成される三つ

組とすると，InfoNCE 損失は次式で定義される：

$$\mathcal{L}_{\text{retr}} = -\log \frac{\exp\left(\tilde{z}_q^\top \tilde{z}_{d^+}/\tau\right)}{\sum_{d\in\{d^+,d_1^-,\ldots,d_K^-\}} \exp\left(\tilde{z}_q^\top \tilde{z}_d/\tau\right)}. \quad (4)$$

この損失は，クエリ埋め込みと正例文書埋め込みの類似度が，負例文書との類似度に比べて大きくなるよう学習する。

総合損失は，検索損失の重み $\alpha \geq 0$ を用いて次式で与えられる：

$$\mathcal{L} = \mathcal{L}_{\text{supcon}} + \alpha\,\mathcal{L}_{\text{retr}} + \lambda_{\text{id}}\,\mathcal{R}_{\text{id}} + \lambda_{\text{orth}}\,\mathcal{R}_{\text{orth}}. \quad (5)$$

本研究では，まず $\alpha = 0$ としてクラスタリング特化設定で提案変換の効果を評価し，その後、検索タスクへの汎化能力を検証する。さらに、$\alpha > 0$ となるマルチタスクに対応するための評価をする。

3 実験設定

3.1 評価タスクおよび指標

日本語テキスト埋め込みベンチマークである JMTEB[8, 9] を用いて評価を行う。JMTEB には，クラスタリング (Clustering)，分類 (Classification)，検索 (Retrieval)，文間類似度 (STS)，リランキング (Reranking) のタスクが含まれる。本研究では，クラスタリングと分類を学習に用い，教師ありコントラスト損失により低ランク変換パラメータを最適化する。評価時には，学習に用いたタスクに加え，検索，STS，リランキングのタスクを評価専用とする。

3.2 ベースライン設定

ベースラインとして，Azure OpenAI Service が提供する多言語汎用埋め込みモデル `text-embedding-3-large`[10]（埋め込み次元 $d = 3072$）の素のベクトルを用いる。実験条件を明確にするために、ベースラインは上記のモデルに絞ったが、本研究の枠組みはオープンソースの埋め込みモデルを含む他モデルにも同様に適用可能であると考えている。

3.3 アブレーション設計

手法の各設計要素が性能に与える影響を系統的に評価するため，5 段階のフェーズに分けて実験を実施した。

Phase1：ベースライン評価 (Baseline) まず，変換を適用しないベース埋め込み（`text-embedding-3-large`）の性能を測定し，比較基準を設定した。これにより，提案変換による性能変化を定量的に評価可能とする。ベースライン性能は，商用 API が提供する汎用埋め込みの現状性能を表しており，クラスタリング，分類，検索，STS，リランキングの各タスクタイプにおける性能の参考値となる。また、単純な前処理手法として，以下も評価する。mean-centering：全データにわたる埋め込みベクトルの平均を算出し，各ベクトルからこの平均を減算することで，分布を原点周りに中心化する。L2 正規化の有無：各ベクトルを L2 ノルムで正規化したもの。

Phase2：基本設定での初期検証 低ランク変換の基本的な有効性を検証するため，中程度のランク（$r = 64$）と標準的な正則化設定（$\lambda_{\text{id}} = 10^{-4}, \lambda_{\text{orth}} = 0$）で初期実験を実施した。この結果は後続の Phase3 における基準モデル（Phase3_r64）として機能し，ランク選択の基準点を提供する。

Phase3：ランクスイープ 最適なランク r を探索するため，$r \in \{4, 16, 64, 128, 256\}$ の 5 種類のモデルを学習・評価した。全モデルで恒等性正則化のみを適用し（$\lambda_{\text{id}} = 10^{-4}, \lambda_{\text{orth}} = 0$），正則化の影響を同一にした。$r = 4$ では約 25K パラメータ，$r = 64$ では約 393K，$r = 256$ では約 1.6M と，パラメータ数は $2dr$ に比例して増加する。この実験により，(1) ランクが小さすぎる場合の表現力不足，(2) ランクが大きすぎる場合の過学習，(3) クラスタリング性能と他タスク性能のトレードオフ，という 3 つの観点から最適なランクを特定する。

Phase4：正則化 Phase3 で特定した最適ランクを用いて，正則化項の効果を検証した。具体的には，以下の 3 モデルを比較した：

- **Phase4_NoReg**：$\lambda_{\text{id}} = 0, \lambda_{\text{orth}} = 0$（正則化なし）変換の自由度が最大であり，クラスタリング性能の理論的上限に近いが，他タスクへの悪影響も最大となる可能性がある。
- **Phase4_OrthOnly**：$\lambda_{\text{id}} = 0, \lambda_{\text{orth}} = 10^{-4}$（直交性正則化のみ）$W^\top W \approx I$ を促すことで距離保存特性を付与し，汎用性の維持を図る。
- **Phase4_RegBest**：$\lambda_{\text{id}} = 10^{-4}, \lambda_{\text{orth}} = 10^{-4}$（両正則化）恒等性と直交性の両方の制約を課す場合。

この比較により，(1) 正則化の必要性，(2) 恒等性と直交性のどちらが重要か，(3) 両正則化の併用効果，を明らかにする。特に，直交性正則化が距離保存を通じて他タスクの汎化性能に寄与するかを検証

する。

Phase5：マルチタスク学習（α の調整） Phase4 までの実験では，クラスタリング・分類タスクのみを学習対象とし（$\alpha = 0$），検索タスクへの汎化能力をゼロショット設定で評価した。この設定は，クラスタリング性能を最大化する一方で，検索タスクへの適応は距離保存特性に依存するため，性能劣化が避けられない。そこで，クラスタリング性能と検索性能のトレードオフを制御するため，総合損失（式 5）の検索損失重み α を調整したマルチタスク学習を検討する。

具体的には，Phase4 で最良の設定（$r = 64, \lambda_{\text{id}} = 0, \lambda_{\text{orth}} = 10^{-4}$）を固定し，$\alpha \in \{0.1, 0.5, 1.0\}$ の複数値でモデルを学習する。これにより，以下の 2 つの極端な設定の間の連続的なトレードオフを評価できる：

- **クラスタリング特化**（$\alpha = 0$）：検索損失がない場合であり、クラスタリング性能は最大化されるが，検索性能は劣化する可能性が高い。
- **マルチタスク**（$\alpha > 0$）：検索損失を追加。検索性能は維持・改善される一方，クラスタリング性能は若干犠牲になる可能性がある。

評価では，クラスタリング指標（v-measure）を x 軸，検索指標（nDCG@10）を y 軸としたトレードオフの分析をして，実用上の要件に応じた最適な α を特定する。

3.4 学習手順

各モデルの学習は以下の手順で行った：(1) クラスタリング・分類タスクのデータをクラスごとにサンプリング（各バッチに複数クラスが含まれるよう調整），(2) 各サンプルをベース埋め込み API に通し，キャッシュされた e を取得，(3) 低ランク変換 $z = e + UV^{\top}e$ を適用後 L2 正規化，(4) SupConLoss（式 1）と正則化項（式 2, 3）を計算し，AdamW で更新。ハイパーパラメータは温度 $\tau = 0.07$，学習率 10^{-3}，バッチサイズ 128，エポック数 50 とした。

4 実験結果と考察

4.1 ベースライン性能

表 1 に，全モデルのタスクタイプ別平均スコアを示す (なお、提供される API の違いにより Web 掲載のものと若干の誤差を含んでいる場合がある)。Phase1 では，ベース埋め込み（Baseline）に加え，単純な前処理手法として mean-centering（Phase1_mc）と L2 正規化との組み合わせ（Phase1_L2）を評価した。これらの前処理は，クラスタリング（+1.35%）と分類（+0.46%）で僅かな改善を示し，検索タスクでも性能を維持した（+0.31%）。総合性能はベースラインから 3.11%向上し（0.7420 → 0.7666），単純な前処理でも一定の効果があることが確認された。Phase2 では，低ランク変換（$r = 64, \lambda_{\text{id}} = 10^{-4}, \lambda_{\text{orth}} = 0$）により，クラスタリング性能が 8.43%向上した（0.5056 → 0.5482）。一方で，検索タスクの性能は大幅に低下し（0.8128 → 0.1664, -79.52%），総合性能はベースラインから 40.37%劣化した（0.7420 → 0.4424）。これは，教師ありコントラスト損失のみでは検索タスクへのゼロショットでの汎化が困難であることを示している。

表 1 タスクタイプ別平均スコア（各タスクタイプ内の全タスクの平均）

モデル	Clust.	Class.	Retr.	STS	Rerank.	Overall
Baseline	0.5056	0.7728	0.8128	0.8254	0.8913	0.7420
Phase1_mc	0.5056	0.7731	0.8153	0.8250	0.8910	0.7651
Phase1_L2	0.5124	0.7764	0.8152	0.8233	0.8872	0.7666
Phase2	0.5482	0.6729	0.1664	0.7954	0.8733	0.4424

表 2 ランクの影響（$\lambda_{\text{id}} = 10^{-4}, \lambda_{\text{orth}} = 0$）

Rank	Clust.	Class.	Retr.	Overall	vs $r = 64$
4	0.3798	0.6567	0.1474	0.3874	-12.42%
16	0.5379	0.6721	0.1596	0.4456	+0.74%
64	0.5482	0.6729	0.1664	**0.4424**	—
128	0.5376	0.6723	0.1601	0.4395	-0.65%
256	0.5452	0.6693	0.1632	0.4401	-0.51%

4.2 ランクスイープ

表 2 にランクの影響を示す。$r = 64$ が最適であり，$r = 4$ では不十分（-12.42%），$r \geq 128$ では過剰パラメータ化により性能が劣化した（-0.65%）。これは $r \approx d/48$ が適切であることを示唆する。$r = 16$ から $r = 64$ への増加により性能がわずかに向上する（+0.74%）が，$r = 64$ 以降は性能が横ばいまたは低下する。これは，クラスタリング・分類タスクの本質的な複雑さが $r = 64$ 程度の部分空間で表現可能であり，それ以上のパラメータは過学習を引き起こすことを示している。また，全てのランク設定で検索タスクの性能が低い（0.15 前後）ことから，ランクの調整だけでは検索タスクへのゼロショットでの汎化は困難であることが確認された。

表 3　正則化項の有無による性能比較（$r = 64$）

モデル	λ_{id}	λ_{orth}	Overall	vs NoReg
Phase4_NoReg	0	0	0.3780	—
Phase3_r64	10^{-4}	0	0.4424	+17.04%
Phase4_OrthOnly	**0**	$\mathbf{10^{-4}}$	**0.6555**	**+73.41%**
Phase4_RegBest	10^{-4}	10^{-4}	0.6530	+72.75%

表 4　マルチタスク学習（$r = 64$，$\lambda_{\text{orth}} = 10^{-4}$）

α	Clustering	vs Baseline	Retrieval	vs Baseline
Baseline	0.5143	—	0.7322	—
0.0 (Phase4)	0.7281	+41.6%	0.6593	-10.0%
0.1	0.7267	+41.3%	0.6730	-8.1%
0.5	0.7401	+43.9%	0.6864	-6.3%
1.0	**0.7456**	**+44.9%**	**0.6913**	**-5.6%**

4.3　正則化

表 3 に正則化の効果を示す。直交性正則化のみ（Phase4_OrthOnly）が，恒等性正則化のみ（Phase3_r64）に対して 48.19%，正則化なし（Phase4_NoReg）に対して 73.41%の改善を示した。両正則化を併用（Phase4_RegBest）しても性能は向上せず（0.6530），**直交性正則化のみで十分**であることがわかる。これは、恒等性正則化 $\mathcal{R}_{\text{id}}$ はパラメータのノルム $\|U\|_F^2 + \|V\|_F^2$ を抑制するが，これは変換の大きさを制限するだけで，変換の性質（距離保存性等）を保証しない。一方，直交性正則化 $\mathcal{R}_{\text{orth}}$ は $W^\top W \approx I$ を促し，これは自動的に $\|W - I\|_F^2$ の抑制も含意する。したがって，直交性正則化は恒等性正則化の効果を包含しつつ，さらに距離保存という強い制約を課すため，単独で十分な効果を発揮すると考えられる。

4.4　マルチタスク学習

Phase4 で観測された検索タスクの性能劣化に対し，Phase5 では教師ありコントラスト損失と InfoNCE の重み付き結合によるマルチタスク学習を導入した。表 4 に異なる α 値での性能を示す。

$\alpha = 1.0$（InfoNCE のみ）では，クラスタリング性能を 44.9%向上させつつ，検索タスクの劣化を 5.6%に抑制できた。これは，Phase4（$\alpha = 0.0$）の検索性能劣化（-10.0%）を低減させる一方，クラスタリング性能も維持している。α の増加に伴い，検索性能が単調に改善（0.6593 → 0.6913）する一方，クラスタリング性能も向上（0.7281 → 0.7456）している。これは，InfoNCE による意味的類似度の学習が，クラスタリングタスクにも有効であることを示していると考えられる。

表 5　正則化設定別のタスク性能（$r = 64$）

モデル	Clust.	Retr.	Overall	vs Base
Baseline	0.5056	0.8126	0.7420	—
Phase3_r64	0.5482	0.1664	0.4424	-40.4%
Phase4_NoReg	0.4639	0.0830	0.3780	-49.1%
Phase4_OrthOnly	**0.5830**	0.5842	**0.6555**	**-11.7%**

4.5　タスク適応と汎化性能の考察

表 5 に，Phase4 の正則化設定別の詳細な性能を示す。Phase4_OrthOnly は，クラスタリングで 15.32%改善した（0.5056 → 0.5830）。一方，検索タスクでは 28.11%劣化（0.8126 → 0.5842）したが，これは教師ありコントラスト損失がクラスタリング向けに同一ラベルのサンプルを集約する一方，検索タスク特有のクエリ-文書ペア構造を学習できないためであると考えられる。

表 5 から，直交性正則化の効果が確認できる。Phase3_r64（恒等性正則化のみ）では検索性能が-79.5%劣化したのに対し，Phase4_OrthOnly（直交性正則化のみ）では-28.1%に抑制され，51 ポイント改善している。これは，$W^\top W \approx I$ を満たす変換がベクトル間の距離を保存するため，未学習タスクの意味的距離構造を破壊しないためであると思われる。しかし，依然としてベースラインには及ばず，**検索タスクには専用の学習が必要**であった。Phase5 で導入したマルチタスク学習により，この限界を緩和できることを確認した。

5　まとめ

本研究では，商用埋め込み API に対する低ランク幾何変換による日本語クラスタリング適応手法を評価した。まとめとしては以下の通りである：(1) 本実験でのタスクでは、直交性正則化のみで 73%の性能改善を達成し，恒等性正則化との併用は不要であることがわかった。(2) 最適ランク $r = 64$（$\approx d/48$）を特定し，過小・過大いずれも性能を劣化させることがわかった。(3) クラスタリングで 15%改善しつつ，他タスクでは 2%程度の劣化に留まり，実用的なトレードオフを評価した。(4) 検索タスクでは 28%劣化したが，マルチタスク学習により劣化を 5.6%に抑制できることがわかった。今後の課題として，より広範なタスクを含むマルチタスク学習と，特定タスクでの性能劣化の原因分析が挙げられる。

参考文献

[1] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. **arXiv preprint arXiv:2106.09685**, 2021.

[2] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In **Proceedings of the AAAI Conference on Artificial Intelligence**, 2018.

[3] Qiang Qiu and Guillermo Sapiro. Orthogonal low-rank transformation. In **Handbook of Robust Low-Rank and Sparse Matrix Decomposition**. Chapman and Hall/CRC, 2014.

[4] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2020. https://arxiv.org/abs/2004.11362.

[5] Denghui Zhang, Zixuan Yuan, Yanchi Liu, and Haifeng Chen. Supporting clustering with contrastive learning. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics**, 2021.

[6] Yujie Qian, Yiming Cui, Ruoyu Li, and Zheng Lin. Contrastive learning subspace for text clustering. **arXiv preprint arXiv:2408.14119**, 2024.

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. https://arxiv.org/abs/1807.03748.

[8] Shengzhe Li, Masaya Ohagi, and Ryokan Ri. Jmteb: Japanese massive text embedding benchmark, 2024. https://huggingface.co/datasets/sbintuitions/JMTEB.

[9] SB Intuitions. 日本語テキスト埋め込みベンチマーク jmteb の構築, 2024. https://www.sbintuitions.co.jp/blog/entry/2024/05/16/130848.

[10] OpenAI. text-embedding-3-large, 2024. https://platform.openai.com/docs/guides/embeddings.

[11] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In **Proceedings of the 5th International Conference on Learning Representations (ICLR)**, 2017. https://dblp.org/rec/conf/iclr/AroraLM17.html.

[12] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval, 2021. https://arxiv.org/abs/2103.15316.

[13] 佐々木翔大, Benjamin Heinzerling, 鈴木潤, 乾健太郎. 白色化が単語埋め込みに及ぼす効果の検証. 言語処理学会第 29 回年次大会発表論文集, 2023. https://www.anlp.jp/proceedings/annual_meeting/2023/pdf_dir/Q11-3.pdf.

[14] Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. Whiteningbert: An easy unsupervised sentence embedding approach. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 238–244, 2021. https://aclanthology.org/2021.findings-emnlp.23/.

[15] Andor Diera, Lukas Galke, and Ansgar Scherp. Isotropy matters: Soft-zca whitening of embeddings for semantic code search. In **Proceedings of the European Symposium on Artificial Neural Networks (ESANN)**, 2025. https://www.esann.org/sites/default/files/proceedings/2025/ES2025-58.pdf.

[16] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. 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. 55–65, 2019. https://aclanthology.org/D19-1006/.

[17] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In **Proceedings of the 37th International Conference on Machine Learning**, pp. 9929–9939, 2020. https://proceedings.mlr.press/v119/wang20k.

[18] Lukas Stankevičius and Mantas Lukoševičius. Extracting sentence embeddings from pretrained transformer models, 2024. https://arxiv.org/abs/2408.08073.

[19] Kevin Zielnicki, Michael C. Mozer, and Rishabh Mehrotra. Orthogonal low rank embedding stabilization. **arXiv preprint arXiv:2508.07574**, 2023.

[20] 塚越駿, 笹野遼平, 武田浩一. 定義文を用いた文埋め込み構成法. 自然言語処理, Vol. 30, No. 1, pp. 125–155, 2023. https://doi.org/10.5715/jnlp.30.125.

[21] Hirofumi Shimizu and Daisuke Kawahara. 非言語データを用いた対照学習による文埋め込み学習の日本語における効果検証. 人工知能学会全国大会論文集（第 37 回）, pp. 3Xin4–65, 2023. https://doi.org/10.11517/pjsai.JSAI2023.0_3Xin465.

[22] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, 2020.

A　幾何学的解釈と設計意図

低ランク幾何変換の重要な特性は，**恒等写像からの微調整**である。変換行列を $W = I + UV^\top$ と定義することで，恒等変換からのずれが明示的に分離され，正則化により W の逸脱を制御できる。特に，直交性正則化 $\mathcal{R}_{\text{orth}}$ は，$W^\top W \approx I$ を満たす変換を促し，これは距離保存の観点から重要である。すなわち，

$$\|Wx - Wy\|^2 = (x-y)^\top W^\top W(x-y) \approx \|x-y\|^2$$

となり，ベースである埋め込み空間の距離構造を大きく変えずに，クラスタリングに有利な方向への調整が可能となる。この性質により，ベース埋め込みが持つ多タスク対応能力を維持しつつ，特定タスクへ適応できると考える。

また，ランク r は表現力と汎用性のトレードオフを制御するハイパーパラメータとなる。r が小さすぎると変換の自由度が不足し，クラスタリング性能の改善が限定的となる。一方，r が大きすぎると過学習が生じ，他タスクへの悪影響が増大する。本研究では，このパラメータの最適なバランス点を調査した．

B　関連研究

埋め込みの後処理に関する既存研究として，SIF (Simple but tough-to-beat)[11] は主成分除去による軽量な文埋め込み手法を提案した。その後，BERT-whitening[12, 13, 14] や Soft-ZCA whitening[15] は共分散の逆平方根行列を用いた変換により，英語 STS タスクやコード検索で性能改善を示した。しかし，フルランク変換は特定タスクに過適応し，他タスクの性能を損なうリスクがある。Ethayarajh[16] は，BERT 等の文脈化埋め込みの幾何学的性質を分析し，異方性が性能に与える影響を明らかにした。Wang & Isola[17] は，コントラスト学習における整列性 (alignment) と均一性 (uniformity) の重要性を示し，これは本研究の直交性正則化とも関連する。さらに，事前学習モデルからの文埋め込み抽出手法 [18] も研究されているが，これらは主に特定の埋め込み層の選択や平均化手法に焦点を当てており，幾何変換とは異なるアプローチである。

本研究の手法として、低ランク幾何変換 [1, 2] は，(1) 恒等写像からの微調整により汎用性を維持，(2) 低ランク制約によりパラメータ数を $O(dr)$ に抑制，(3) 直交性正則化 [3, 19] により距離構造を保存，という点が特徴であり、特に，日本語クラスタリングタスクに特化した評価と，汎化能力の定量的分析を行っている．

C　日本語特有の要因分析

日本語テキスト埋め込みにおける本手法の効果は，以下の言語的要因に影響される可能性がある。日本語における文埋め込み研究としては，定義文を用いた手法 [20] や非言語データを用いた対照学習 [21] 等が提案されているが，本研究は後処理による適応という点で異なるアプローチを取る：

表記揺れと語彙多様性　日本語は，漢字・ひらがな・カタカナの混在により，同一概念の表記揺れ（例：「サーバー」「サーバ」「鯖」）が多い．商用埋め込みは多言語学習により表記揺れを吸収しているが，クラスタリングタスクでは同一ラベル内の表記揺れサンプルを近づける必要がある．教師ありコントラスト損失による低ランク変換は，こうした細かな変動を平滑化し，クラスタリング性能を向上させたと考えられる．

ドメイン差と文脈依存性　JMTEB は，ニュース記事（livedoor_news），論文（nlp_journal），対話（wrime_classification）等の多様なドメインを含む [8]．Phase4_OrthOnly は，livedoor_news で大きな改善を示した一方，sib200（少数言語短文）では性能が低下した．これは，長文・定型文が多いドメインでは意味的構造が安定しており，低ランク変換が効果的である一方，短文や口語的表現では文脈依存性が高く [22]，固定的な変換が適合しにくいためと推察される．

検索タスクへの適応限界　日本語検索タスクでは，クエリと文書の語彙ギャップが大きい．教師ありコントラスト損失は同一ラベル内の類似度を高めるが，クエリ-文書間の語彙的・意味的ギャップを橋渡しする学習は行わない [4]．このため，検索タスクでは InfoNCE 等の専用損失が必要となると考えられる。