指示認識テキスト埋め込みモデルの指示によるベクトル変位の分析

小川隼斗 [1,2]　福地成彦 [2]　李聖哲 [1,2]　河原大輔 [1]
[1] 早稲田大学理工学術院　[2]SB Intuitions
{cookie3120@ruri., shengzhe.li@asagi, dkw@}waseda.jp
{akihiko.fukuchi, shengzhe.li}@sbintuitions.co.jp

概要

テキスト埋め込みモデルはテキストの意味を埋め込みとして表現するためのモデルであり, 情報検索や質問応答システム, 文書分類など様々なタスクで活用されている. 近年では, 入力テキストにタスク固有の「指示」を付与することで, 各目的に最適化された埋め込みを獲得する手法が増えている. 本研究では, そのような指示認識埋め込みモデルにおいて, クエリに対する異なる指示の付与が埋め込み空間上の配置にどの程度変化をもたらすかを分析する. 実験の結果, タスクの種類によって埋め込み空間上の配置の変化の度合いは大きく異なることが明らかになった. また, パラメータサイズの増大は必ずしもタスクごとの埋め込み空間の分離を増大するわけではないことが明らかになった.

1　はじめに

情報検索や文書分類といったタスクにおいて, テキスト全体の意味を密なベクトル, すなわち密な埋め込みとして表現するテキスト埋め込み技術の研究が盛んに行われている. 近年では, 対照学習の導入によりテキスト全体を適切に表現することに最適化されたモデルの開発が進展し, 情報検索や類似文検索など様々なタスクにおける性能が飛躍的に向上した [1, 2].

しかし, 同一のテキストであっても, 対象とするタスクの性質に応じて最適な埋め込み表現は異なると考えられる. 例えば, 「織田信長が生まれた年はいつ?」という入力文に対し, 情報検索を行いたい場合は「織田信長の生まれ年は 1534 年です」のような事実に焦点を当てた文に近い埋め込みが最適である. 一方で, 意味的類似性の高いテキストを取得したい場合は「織田信長がいつ生まれたか知りたい」のような, 文の意図が共通する文に近い埋め込みが最適となる. こうした背景から, テキストに自然言語によるタスク指示を追加し, それを文脈として扱うことで, タスクの目的に沿って埋め込みを動的に変化させる手法 [3, 4, 5] が注目を集めている. 指示を認識できるモデルは, タスク指示を与えることで, こうしたベクトル空間上での動的な調整を可能にしている.

塚越ら [6] は指示認識埋め込みモデルにおける埋め込みの冗長性を調査し, 分類等では次元削減に頑健だが検索では性能低下が大きいことを示しており, 指示が埋め込みの性質を変える可能性が示唆される. その一方で, こうしたモデルが指示に応じて具体的にどの程度埋め込み空間を変容させているのか, その挙動については十分に解明されていない. 指示を与えることで, どのように空間的な変位するかを明らかにすることは, モデルの解釈性を高め, テキスト埋め込みモデルのさらなる性能向上を図る上で重要である. また, 一般に言語モデルの大規模化は表現能力の向上に直結すると考えられているが, テキスト埋め込みモデルにおける指示による空間的変位能力が, パラメータサイズに比例して向上するかについても明らかでない.

本研究では, 異なるサイズおよびアーキテクチャを持ち, かつ指示を認識できるテキスト埋め込みモデルを独自に構築し, タスク指示が埋め込み空間に与える影響を分析する. 異なる次元や空間を持つモデルを公平に比較するため, 共通空間への射影を用いて同一基準での分析を行う. その結果, 指示による埋め込みの変位は一様ではなく, 検索タスクと比較して分類や意味的類似度計算（STS）などのタスクにおいて, 空間上の移動がより顕著に現れることが明らかになった. 加えて, パラメータサイズと指示による空間分離性能の間には単純な相関が見られないことが示唆された.

2 関連研究

異なるモデルや異なる言語で学習された埋め込みは, それぞれの埋め込み空間が異なるため, そのままでは直接的な比較が困難である. これらを比較可能にするため、一方の埋め込み空間を他方の空間へ線形変換を用いて射影する手法が知られている [7]. しかし、Xing ら [8] は, その手法に対し, 埋め込みの学習における目的関数と, 埋め込み間の距離尺度と, 変換行列の学習における目的関数の間に不整合があると指摘している. そこで代替する手法として直交変換を用いることで, 単語間の相対的な位置関係を崩すことなく, 異なる空間上の埋め込み同士を整合性高くマッピングすることが可能であると述べている. 本研究では, この直交法に基づく射影を用いて異なるサイズや指示を持つモデル間の埋め込みを比較分析する.

3 モデルの学習

本研究では, アーキテクチャやパラメータサイズの違いが指示付き埋め込みの挙動に与える影響を分析するため, 共通の学習設定とデータセットを用いて指示を認識できるテキスト埋め込みモデルを構築する. 学習手法には, 近年のテキスト埋め込みモデル開発において主流となっている 2 段階の学習プロセス [5, 9, 10] を採用する. Stage 1 の学習は, 指示を付与していない弱教師ありデータによる基礎的な埋め込み能力の獲得を目的とし, Stage 2 の学習は, タスク指示を付与した高品質な教師ありデータによる指示追従能力とより強力な埋め込み能力の獲得を目的としている. なお, 損失関数やハイパーパラメータを含む詳細な学習設定については付録 A に記述する.

3.1 ベースモデル

異なるアーキテクチャおよびパラメータサイズにおける特性を比較するため, 以下の 2 種類のモデルを採用した.

- ModernBERT-ja [11]: エンコーダモデルであり, パラメータサイズは 30M, 70M, 130M, 310M, 1.4B の 5 種類を使用.
- Sarashina2.2[1]: 双方向化していないデコーダモデルであり, パラメータサイズは 0.5B, 1B, 3B の 3 種類を使用.

1) https://huggingface.co/collections/sbintuitions/sarashina22

表 1 各学習段階で使用したデータセットの内訳

学習段階	データセット名	サンプル数
Stage 1	Auto-Wiki QA[2]	2,377,491
	Auto-Wiki NLI Triplet[3]	198,895
	Auto-Wiki QA (Nemotron)[4]	156,088
	JSQuAD [12]	62,859
	総計	**2,795,333**
Stage 2	Dataset for Retrieval	407,632
	Dataset for STS	209,375
	Dataset for Classification	22,000
	Dataset for Clustering	13,884
	総計	**652,891**

3.2 学習データセット

各学習段階において使用したデータセットの内訳を表 1 に示す.

Stage 1: 弱教師あり学習 Stage 1 では, クエリと文書のペアからなる弱教師ありデータセットを用いた. これらは Stage 2 で用いるデータ数と比べ大規模であり, モデルに基礎的なテキスト表現能力を獲得させることを目的としている. 形式として (クエリ, ポジティブ文書) の 2 つ組を採用した. なお, ここでいうポジティブ文書とは, 情報検索タスクにおいてはクエリに対する回答や関連情報を含む文書を指し, NLI タスクにおいては前提文に対し, 含意のラベルが付けられた仮説文を指す.

Stage 2: 指示付き学習 Stage 2 では, より高品質な教師ありデータセットであり, 形式として (指示, クエリ, ポジティブ文書, ハードネガティブ文書) からなる 4 つ組を採用した. 検索タスク用のデータセットの指示およびクエリは LLM によって合成されたものを用いており, ポジティブ文書, ハードネガティブ文書の選定にも LLM を用いている. その他のタスクについては, 各データセットに適した指示を人手で作成し使用している. 本段階で用いたデータセットの構築手法および加工の詳細は付録 B に記述する. また, 構築したモデルの性能を確認するため, 日本語テキスト埋め込みベンチマーク JMTEB [13] で評価実験を行い, 構築した全てのモデルが実用的な性能水準に達していることを確認した.

2) https://huggingface.co/datasets/cl-nagoya/auto-wiki-qa
3) https://huggingface.co/datasets/hpprc/emb
4) https://huggingface.co/datasets/cl-nagoya/auto-wiki-qa-nemotoron

表 2　各タスクとして使用する指示

Base	-（指示を付与しない）
Retrieval	クエリを与えるのでクエリに答えることができる文章を検索してください。
Reranking	クエリを与えるのでクエリに関連する文章を検索してください。
STS	クエリを与えるので、もっともクエリに意味が似ている一節を探してください。
Classification	与えられたクエリを適切なカテゴリに分類してください。
Clustering	与えられたクエリのトピックを特定してください。

4　分析

本節では，構築したモデルに加え，3 つの公開されているテキスト埋め込みモデル Qwen3-Embedding [3], InstructOR [4], bge-v1.5 [5] を用いて，タスク指示が埋め込み空間に与える影響を分析する．これらのモデルは全て指示を認識可能であり，特に Qwen3-Embedding はテキスト埋め込みのベンチマークである MTEB [14] において非常に高い性能を記録しているため選定した．はじめに，分析に使用したクエリの構築方法と，異なるモデル間の比較を可能にするための分析手法について述べる．その後，分析結果について述べる．

4.1　分析用クエリの構築

指示の違いによる埋め込みの変位を公平に観測するためには，特定のタスクに過剰に適合したテキストではなく，検索や分類，STS といった多様な指示を受け入れても自然に成立する汎用的なテキストが必要である．本研究では，日本語質問応答データセットである JaQuAD [15] を基に，LLM を用いて分析用クエリを作成した．具体的には，事後学習済みの Sarashina2-70B[5] を使用し，JaQuAD の質問文を「その答えとなる実体や概念を間接的に説明する名詞句」に書き換える生成を行った．入出力の例を次に示す．

> **入力:** 戦後日本のストーリー漫画の第一人者で、医学博士の一面もある漫画家は誰?
> **出力:** 戦後日本のストーリー漫画の第一人者で、医学博士の一面もある漫画家

この形式のテキストは，検索タスクにおいては「クエリ」，STS タスクにおいては「比較対象の文」，分類タスクにおいては「分類対象の文」として扱っても文脈上の違和感が生じにくい．生成されたテキストの中から，人手による確認を経て，書き換え意図に沿った 200 件を選定した．各クエリに対し，表 2 に示すタスクごとの指示を付与してモデルに入力し，得られた埋め込みを分析対象とした．また，3 つの公開モデル [3, 4, 5] には同様に事後学習済みの Sarashina2-70B を用いて英語に翻訳した分析用クエリおよび指示を入力する．

5)　https://huggingface.co/sbintuitions/sarashina2-70b

4.2　分析手法

各モデルは異なる埋め込み次元を持つため，はじめに各モデルの埋め込みを共通の次元へ圧縮した後，直交変換を用いて基準となるモデルの空間へ射影する手法を採用する．

次元の統一　はじめに，全てのモデルの埋め込みベクトルを共通の次元数 $k = 200$ へ圧縮する．具体的には，分析用クエリの総数を N，各モデル固有の埋め込み次元数を d とし，実数空間上の埋め込み行列 $\mathbf{E} \in \mathbb{R}^{N\times d}$ に対し主成分分析（PCA）を適用することで，圧縮された行列 $\mathbf{S} \in \mathbb{R}^{N\times k}$ を得た．これにより，出力次元の異なるモデルを同一の次元で扱うことが可能となる．

共通空間への射映　次に，異なる系列のモデル間を比較可能にするために，直交変換 [8] を用いて異なるモデル間の空間的なズレを補正する．ある比較対象モデルの埋め込み $\mathbf{S}_{\text{tgt}}$ を，基準モデルの埋め込み $\mathbf{S}_{\text{ref}}$ に対応させることを考える．ここでは，Hamilton ら [16] のアプローチを参考に指示を含まない「Base」状態の埋め込みを用いて，以下の目的関数を最小化する直交行列 $\mathbf{Q}^* \in \mathbb{R}^{k\times k}$ を求める．

$$\mathbf{Q}^*_{\text{tgt}\to\text{ref}} = \underset{\mathbf{Q}}{\operatorname{argmin}} \left\| \mathbf{S}^{(\text{Base})}_{\text{tgt}}\mathbf{Q} - \mathbf{S}^{(\text{Base})}_{\text{ref}} \right\|_F \text{ s.t. } \mathbf{Q}^\top\mathbf{Q} = \mathbf{I} \quad (1)$$

この最適化問題の解 $\mathbf{Q}^*_{\text{tgt}\to\text{ref}}$ は，$\mathbf{M} = (\mathbf{S}^{(\text{Base})}_{\text{tgt}})^\top \mathbf{S}^{(\text{Base})}_{\text{ref}}$ の特異値分解 $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ を用いて，$\mathbf{Q}^*_{\text{tgt}\to\text{ref}} = \mathbf{U}\mathbf{V}^\top$ として一意に求められる．本実験において，基準モデル $\mathbf{S}_{\text{ref}}$ の選定が分析結果に特定のバイアスを与えないことを保証することは重要である．そのために，事前に全てのモデルペア $(\mathbf{S}_i, \mathbf{S}_j)$ について式 (1) の最適解 $\mathbf{Q}^*_{i\to j}$ を算出し，以下の正規化プロクラステス距離 D を用いて類似度行列を作成した．詳細は付録 C に記述する．

$$D(\mathbf{S}_i, \mathbf{S}_j) = \frac{\left\| \mathbf{S}_i\mathbf{Q}^*_{i\to j} - \mathbf{S}_j \right\|^2_F}{\left\| \mathbf{S}_j \right\|^2_F} \quad (2)$$

この距離行列に基づき，他の全てのモデルに対する距離の平均値が最も小さくなる，すなわちモデル空間群の幾何学的な中心に位置するとみなせる

Sarashina2.2-3B を基準モデルとして採用した. なお, 付録 C にて議論する通り, 基準モデルの選定が分析の結論に大きな影響を与えないことを確認している.

可視化手法　共通空間への射映後の位置関係を俯瞰すべく, 計 18 モデル ×6 タスクの埋め込みベクトル間のコサイン距離行列をもとに, 多次元尺度構成法（MDS）を適用して 2 次元平面へ射影した.

4.3　分析結果

共通空間へ射影された各モデルの埋め込み配置と, タスク指示による変位を可視化した結果を図 1 に示す. 図中の各点はモデルとタスクの組み合わせを表し, 同一モデルにおける「Base」状態から各タスクへの変化を直線で結合している. この可視化結果から得られた主な知見を以下に述べる.

モデル系列による空間の局在性　第一に, 各モデルは系列ごとにそれぞれ独立した領域にクラスタを形成しており, 埋め込み空間の配置はタスクの種類よりも, モデルの系列によって強く支配されていることが確認できる. これは, タスク指示による空間的な調整を行っても, 事前学習やアーキテクチャに由来するモデル固有の幾何学的特性は維持され続けることを示唆している.

タスクごとの変位特性　第二に, タスク種別ごとに変位の大きさが異なる点が挙げられる. 多くのモデルにおいて, Retrieval, Reranking タスクよりも, STS や Classification, Clustering タスクにおいて, より大きな変位が生じている. 本実験のモデルは, Stage 1 において検索系のデータセットを多く含む弱教師あり学習を行っているため, 指示を与えない「Base」状態であっても, 本質的に検索タスクに近い特性を帯びていると推測される. そのため, 検索系の指示を与えても埋め込みの変位は微小に留まる一方, 分類や STS といった検索とは性質の異なるタスク指示が与えられた場合には, 検索志向の空間から大きく離れる必要が生じ, 結果として大きな変位が観測されたと考えられる.

パラメータサイズと変位の関係　第三に, モデルのパラメータサイズと指示による移動量の間には, 単純な比例関係が見られないことが確認できる. 本実験で構築した ModernBERT-ja および Sarashina2.2 系列では, パラメータサイズが増大しても軌跡は短く, Base 周辺に留まる傾向が見られた. また, Qwen3-Embedding 系列では, 4B サイズのモデルにおいて軌跡が 8B サイズのモデルよりも長く伸びている. このことから, モデルの大規模化が必ずしも指示への感度を一様に高めるわけではなく, モデルの設計思想や学習手法に依存する側面が強いと言える.

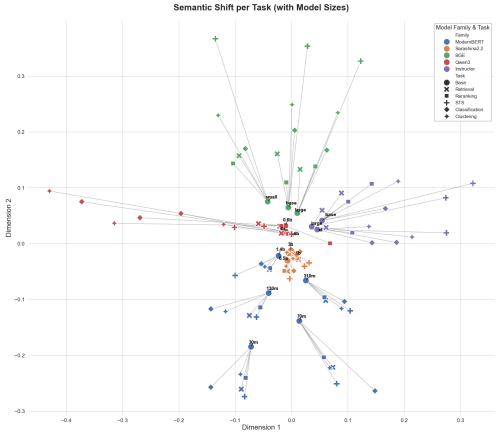

図 1　タスク指示による意味空間上の変位の可視化

5　おわりに

本研究では, 指示認識テキスト埋め込みモデルにおけるタスク指示が埋め込み空間に与える影響について, 分析した。分析にあたっては, パラメータサイズや構造の異なる複数のモデルを対象とし, 直交変換を用いることで, 公平な比較を実現した.

実験の結果, タスク種別による分析からは, 検索タスクよりも STS や Classification, Clusterin タスクにおいて指示による埋め込みの変位が大きい傾向が確認された. また, パラメータサイズの増大は必ずしも指示による埋め込みの変位を促進するわけではないことが示唆された. 特に, Qwen3-Embedding において 4B モデル最も大きな変位を示した点は, パラメータ数が指示への感度を決定づける支配的な要因ではないことを示している.

これらの知見は, 指示追従能力の高い埋め込みモデルの構築において, 単純なモデルパラメータの拡大だけでは不十分であることを示している.

一方で, 本実験における学習量は既存の大規模モデルと比較して少なく, 分析も 200 件のクエリに基づく限定的なものである点には留意が必要である. 加えて, 観測された空間上の変位量と実際のタスク性能との直接的な関連性については未検証である.

今後は, これらの課題を踏まえ, 意図した通りに埋め込み空間をより柔軟に操作できるような学習手法の検討に加え, 指示による変位量とタスク性能との関連性を解明することが必要となる.

謝辞

本研究は SB Intuitions 株式会社と早稲田大学の共同研究により実施した.

参考文献

[1] Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. Semantic re-tuning with contrastive tension. In **International Conference on Learning Representations**, 2021.

[2] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[3] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. **arXiv preprint arXiv:2506.05176**, 2025.

[4] Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 1102–1121, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

[6] 塚越駿, 笹野遼平. プロンプトに基づくテキスト埋め込みのタスクによる冗長性の違い. 言語処理学会 第 31 回年次大会発表論文集, 3 2025.

[7] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In **1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings**, 2013.

[8] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, editors, **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1006–1011, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[9] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2024.

[10] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning, 2023.

[11] 塚越駿. 日本語 ModernBERT の開発: 開発と評価編, 2025. https://www.sbintuitions.co.jp/blog/entry/2025/05/26/115815.

[12] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[13] Shengzhe Li, Masaya Ohagi, Ryokan Ri, Akihiko Fukuchi, Tomohide Shibata, and Daisuke Kawahara. JMTEB and JMTEB-lite: Japanese Massive Text Embedding Benchmark and Its Lightweight Version. **Vol.2025-NL-265,No.3,1-15**, Sep 2025.

[14] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein, editors, **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[15] ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. JaQuAD: Japanese Question Answering Dataset for Machine Reading Comprehension, 2022.

[16] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In Katrin Erk and Noah A. Smith, editors, **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1489–1501, Berlin, Germany, August 2016. Association for Computational Linguistics.

[17] Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. Gecko: Versatile text embeddings distilled from large language models, 2024.

[18] 吉越卓見, 河原大輔, 黒橋禎夫ほか. 機械翻訳を用いた自然言語推論データセットの多言語化. 研究報告自然言語処理 (NL), Vol. 2020, No. 6, pp. 1–8, 2020.

[19] Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. Kalm-embedding-v2: Superior training techniques and data inspire a versatile embedding model, 2025.

[20] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models, 2025.

表 3　ハイパーパラメータの設定

ハイパーパラメータ	Stage 1	Stage 2
Global Batch Size	8,192	1,024
Learning Rate	1×10^{-5}	1×10^{-5}
Warmup Ratio	10%	10%

A　学習設定の詳細

対照学習の損失関数には sentence-transformers の CachedMultipleNegativesRankingLoss [6] を用いる．また，本実験の学習で使用したハイパーパラメータの詳細を表 3 に示す．Stage 1 では大規模なバッチサイズを採用して基礎的な表現を獲得させ，Stage 2 ではバッチサイズを縮小し，高品質なデータによる微調整を行った．

B　Stage 2 で用いたデータセットの構築手法および加工の詳細

各タスクのデータセットの構築方法を記述する．また，全てのデータセットにおいてハードネガティブ文書は 1 事例のみを採用している．

B.1　Dataset for Retreival

検索タスク用のデータセット．Mistral-Small-3.2-24B-Instruct-2506[7] を用いて，文書に対する指示およびクエリを生成した．またポジティブ文書，ハードネガティブ文書の選定においては，Gecko [17] の手法を参考に，Sarashina2.2-3B[8] を用いたハードネガティブマイニングを行うことで，難易度の高いネガティブ例を収集し，ハードネガティブ文書を選定した．具体的にはモデルに対し，クエリと文書を入力し，関連度合い（1〜5）を出力する指示をした．そして，その関連度合いの出力確率を元に関連度順に文書を並び替えた．そのうち，関連度が最も高い文書をポジティブ文書，関連度が 20 番目のものをハードネガティブ文書に選定した．

B.2　Dataset for STS

STS タスク用のデータセット．JSNLI [18] および NU-MNLI[9] を使用し，これらには STS タスク向けの指示を人手で作成し付与した．

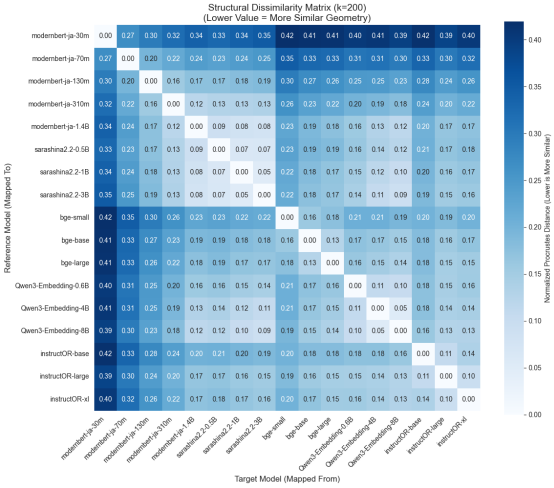

図 2　全モデルペア間の正規化プロクラステス距離

B.3　Dataset for Classification / Clustering

分類およびクラスタリング用のデータセット．これらデータセットには，それぞれのタスクに対応する指示を人手で作成し付与した．またこれらのデータセット構築には，example-based multi-class labeling を採用した．具体的には，クエリに用いる文書と同一のクラスまたはクラスタからランダムにサンプリングした事例をポジティブな文書とし，異なるクラスまたはクラスタからサンプリングした事例をネガティブな文書として扱う手法である．この手法は，KaLM-Embedding-V2 [19] や NV-Embed [20] において事例に対して対応するラベルをポジティブとしてデータセットを構築する手法と比較して，その有効性が確認されている．

C　基準モデル選定について

本研究で用いた全モデルペア間の正規化プロクラステス距離のヒートマップを図 2 に示す．同系列のモデル間での正規化プロクラステス距離は，他系列のモデル間との正規化プロクラステス距離より小さくなる傾向が見られる．また，基準のモデルのサイズを大きいものにするにつれ正規化プロクラステス距離が小さくなる傾向が見られる．この距離行列に基づき，他モデルとの平均距離が最も小さい Sarashina2.2-3B を基準モデルとして選定した．なお，Sarashina2.2-3B 以外のいくつかのモデルを基準モデルとして 4 節と同様の実験を行った結果，本実験における主要な観測結果は一貫することを確認した．

6）https://sbert.net/docs/package_reference/sentence_transformer/losses.html
7）https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506
8）https://huggingface.co/sbintuitions/sarashina2.2-3b
9）https://huggingface.co/datasets/cl-nagoya/nu-mnli