# なぜ平均プーリングはうまく動くのか？<br>テキスト埋め込みの二次統計量の崩壊の定量化

原知正 [1] 栗田宙人 [1] 今泉允聡 [2,3,4] 乾健太郎 [5,1,4] 横井祥 [6,1,4]
[1] 東北大学 [2] 東京大学 [3] 京都大学 [4] 理化学研究所 [5]MBZUAI [6] 国立国語研究所
{hara.tomomasa.s8,hiroto.kurita.q4}@dc.tohoku.ac.jp
imaizumi@g.ecc.u-tokyo.ac.jp kentaro.inui@mbzuai.ac.ae yokoi@ninjal.ac.jp

## 概要

テキスト埋め込みは，標準的には単語埋め込みの平均プーリングで得られる．本稿では，平均プーリングが元の単語埋め込み集合の一次の統計量（平均）のみを使っており，その空間的な広がりを表す二次以後の統計量が失われる点に着目する．実モデルの状況を調べるため，まず平均プーリングによる二次統計量の崩壊の度合いを定量化する指標を提案する．次にこの指標を実テキストと実モデルに適用し，事前学習済みモデルを対照学習で微調整した近年のモデルでは，ベースモデルより二次統計量の崩壊が生じづらいことを経験的に確かめた．粗い集約手法に見える平均プーリングの再検討から，近年のモデルの有用性に新しい視点を与える．

## 1 はじめに

文や文書を単一のベクトルで表現するテキスト埋め込みにより，情報検索 [3] や自動評価 [4] など幅広い自然言語処理タスクを，統一的な枠組みで扱えるようになった．一方で，訓練データ [5] や次元 [6] の観点から検討されてきたように，性能向上と性質の理解は今なお重要な課題である．. 本稿では，テキスト埋め込みを構築する際に，単語レベルの表現をどのように集約するかという集約手法に着目する．

集約手法としては，テキスト中の単語埋め込みを平均する**平均プーリング**が標準的である．この単純な手法は，古典的な静的単語埋め込み [7] から 最先端の Transformer [8] エンコーダによる文脈化埋め込み [5, 9] に至るまで，様々な埋め込み表現で経験的な有用性が確認されてきた．

平均プーリングは標準的な手法である一方で，粗い集約手法にも見える．本稿では，平均プーリングは単語埋め込み集合を一次の統計量（平均）のみで

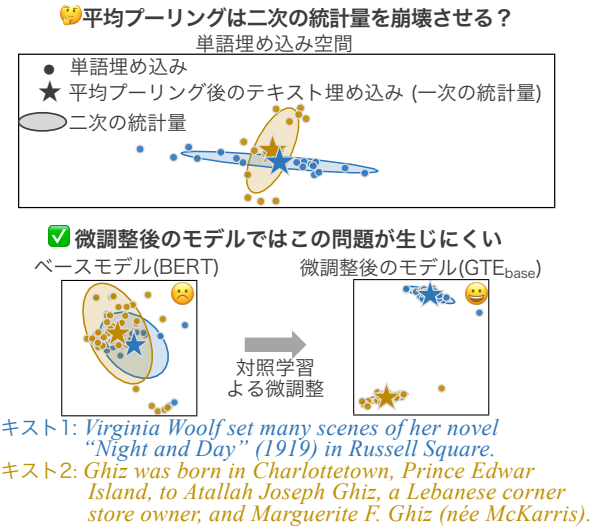


**図 1** 本研究の概要．**上部：**平均プーリングは，異なる単語埋め込み集合であっても，ほとんど同じテキスト埋め込みに集約しうる．**下部：**微調整後のモデルでは平均プーリングによる崩壊が，そのベースモデルよりも生じにくいことを経験的に確認した（§ 5）．各図は BERT [1] と GTE$_{base}$ [2] の埋め込みを主成分分析で可視化している．この例は § 4 の指標によって発見した．

要約するため，空間的な広がりを捉える二次以降の統計量が失われる点に着目する図 1 上部にその例を示す．この例では，元の単語埋め込み集合は異なる（$[\bullet,\cdots,\bullet] \neq [\bullet,\cdots,\bullet]$）にもかかわらず，平均プーリング後のテキスト埋め込みはほとんど同じものになる（$\bigstar \fallingdotseq \bigstar$）．こうした問題を回避するため，テキストを単語埋め込み集合として表現する手法も提案されてきた [10, 11, 12, 13]．しかし，テキスト埋め込みはその計算コストの軽さ [14] や経験的な有効性 [13] から，依然として広く利用されている．

では，実際のテキストとモデルにおいて平均プーリングはどの程度情報を捨てているのか？これを調べるために，まず平均プーリングによる二次統計量の崩壊度を定量化する指標を提案する（§ 4）．次に，この指標を実テキストと実モデルに適用し，事前学

習済みモデルを対照学習で微調整した近年のモデルでは，ベースモデルよりも二次統計量の崩壊が生じにくいことを経験的に確かめた（§ 5，図 1 下部）．粗い集約手法に見える平均プーリングの再検討から，近年のモデルの有用性に新たな視点を与える．

## 2 関連研究

本稿は平均プーリングの再検討から近年のテキスト埋め込みの有用性に新たな視点を与える．本節では，平均プーリングによるテキストの埋め込み表現と，元々の単語埋め込み集合としてテキストを表現する手法それぞれを述べ，その両者を比較する．

**平均プーリングによる埋め込み表現**　平均プーリングは，静的単語埋め込み [7] から Transformer による文脈化単語埋め込み [15] まで広く用いられてきた．近年は，事前学習済み Transformer エンコーダを対照学習 [16] で微調整したテキスト埋め込みモデルが高い経験的な性能を達成しており，最先端のモデルもこの枠組みが採用されている [17, 2, 5, 9]．

**単語埋め込み集合表現**　一方で，テキストをプーリング前の単語埋め込みの集合として表現する手法も提案されてきた．例えば，単語埋め込み集合間の最適輸送コストで意味類似度を計算する手法が提案されている [10, 11, 12, 13]．また，ColBERT は，各クエリ単語を最も類似した文書単語に対応付け，単語レベル類似度を集約して関連度を算出する [18, 14]．

**埋め込み表現の優位性**　しかし依然として，テキスト埋め込みが計算コストの軽さ [14, 19] やその経験的な有用さ [13] から広く用いられている．本稿は，平均プーリングの再検討を通じて，この近年のテキスト埋め込みの有用性に新たな視点を与える．

## 3 準備

本稿では，単語埋め込み集合の一次統計量のみを使う平均プーリングが引き起こす二次統計量の崩壊の度合いを定量化する．このために，単語埋め込み集合とそれぞれの統計量を形式的に定義する．

**単語埋め込み集合**　以降では，モデル $\boldsymbol{f}$ によって二つのテキスト $t_1$ と $t_2$ のテキスト埋め込みを構築することを考える．$\boldsymbol{f}$ は $t_i$ を入力として単語埋め込み集合 $\boldsymbol{X}_i$ を出力する：

$$\boldsymbol{X}_i := [\boldsymbol{x}_{i,1}, \cdots, \boldsymbol{x}_{i,n_i}] = \boldsymbol{f}(t_i) \in \mathbb{R}^{d \times n_i} \tag{1}$$

$\boldsymbol{x}_{i,j} \in \mathbb{R}^d$ は $t_i$ における $j$ 番目の単語埋め込み，$d$ は埋め込みの次元，$n_i$ は $t_i$ の単語数である．

**一次の統計量（平均プーリング）**　平均プーリングはこれらの単語埋め込み集合の一次の統計量（平均）$\boldsymbol{\mu}(\boldsymbol{X}_i) \in \mathbb{R}^d$ をテキスト埋め込みとする：

$$\boldsymbol{\mu}(\boldsymbol{X}_i) := \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{x}_{i,j}. \tag{2}$$

**二次の統計量**　しかし，$\boldsymbol{X}_i$ の空間的な広がりを捉える二次の統計量 (共分散行列) $\boldsymbol{\Sigma}(\boldsymbol{X}_i) \in \mathbb{R}^{d \times d}$ も存在する[1)]：

$$\boldsymbol{\Sigma}(\boldsymbol{X}_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} (\boldsymbol{x}_{i,j} - \boldsymbol{\mu}(\boldsymbol{X}_i))(\boldsymbol{x}_{i,j} - \boldsymbol{\mu}(\boldsymbol{X}_i))^\top. \tag{3}$$

これらの定義に基づいて，次節では平均プーリングによる二次統計量の崩壊を定量化する．

## 4 二次の統計量の崩壊

本節では，図 1 上部が示すような，二つの単語埋め込み集合平均プーリングしたときに生じる二次統計量の崩壊を，直感的・形式的に議論する．

### 4.1 直感的説明：崩壊が生じる条件

図 2 の四隅に，平均プーリングによる二次統計量の崩壊が生じる場合と生じない場合の直感的な例を示した．本稿では，一次および二次の統計量の類似度によってこの崩壊を特徴付ける．

**崩壊が生じる場合**　崩壊は，一次統計量が似ているが二次統計量が異なる，つまり $\boldsymbol{\mu}(\boldsymbol{X}_1) \fallingdotseq \boldsymbol{\mu}(\boldsymbol{X}_2) \wedge \boldsymbol{\Sigma}(\boldsymbol{X}_1) \neq \boldsymbol{\Sigma}(\boldsymbol{X}_2)$ のときに生じる．このとき，二次統計量まで見れば元々の単語埋め込み集合は異なるが，その平均はほぼ同じになる．

**崩壊が生じない場合**　崩壊が生じない状況も同様に整理できる．一つは，一次と二次の統計量の両方が異なる，つまり $\boldsymbol{\mu}(\boldsymbol{X}_1) \neq \boldsymbol{\mu}(\boldsymbol{X}_2) \wedge \boldsymbol{\Sigma}(\boldsymbol{X}_1) \neq \boldsymbol{\Sigma}(\boldsymbol{X}_2)$ のときである（図 2 右上）．このとき，元の分布が異なることは一次統計量の違いから捉えられる．もう一つは，二次統計量が類似する，つまり $\boldsymbol{\Sigma}(\boldsymbol{X}_1) \fallingdotseq \boldsymbol{\Sigma}(\boldsymbol{X}_2)$ のときである（図 2 左下・右下）．このとき，一次統計量も類似すれば元の分布も類似する（左下）．一次統計量が類似しない場合も，単語埋め込み集合の分散が極端に大きくないと仮定すれば，平均によって分布は分離される（右下）．

### 4.2 形式的説明：二次統計量の崩壊度

本節では，平均プーリングによる二次統計量の崩壊度（**S**econd-**O**rder **C**ollapse by **M**ean pooling，以下

1) 簡単のため，平均プーリングによって保持されない最も低次の統計量である二次の統計量に着目する．

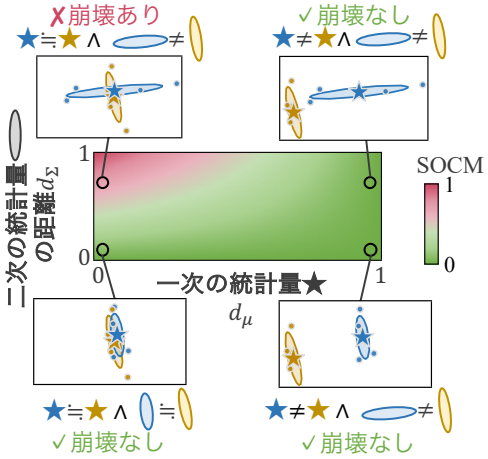


**図 2**　平均プーリングによる二次統計量の崩壊の直感とその定量化．四隅の図は崩壊が生じる場合と生じない場合を一次の統計量と二次の統計量の類似度で分類している．ヒートマップは $(d_\mu, d_\Sigma)$ の各組み合わせに対する SOCM の値．崩壊が生じる場合に SOCM の値は大きく，生じない場合には小さい値となる．

**SOCM**）を形式的に導入する．§ 4.1 で述べたように，本稿では崩壊は一次および二次の統計量の類似度によって特徴付けた．そのため一次と二次それぞれの統計量同士の距離によって SOCM を構成する．また，SOCM に望ましい性質を整理し，それらがすべて満たされるように尺度を構築する．

**SOCM の定義**　二つの単語埋め込み集合 $\boldsymbol{X}_1$ と $\boldsymbol{X}_2$ が与えられたときの平均プーリングによる二次統計量の崩壊度 SOCM とを次式で定義する：

$$\mathrm{SOCM}(d_\mu, d_\Sigma) := (1 - d_\mu) d_\Sigma. \tag{4}$$

ここで，$d_\mu$ は一次統計量 $\boldsymbol{\mu}(\boldsymbol{X}_1)$ と $\boldsymbol{\mu}(\boldsymbol{X}_2)$ の距離であり，ユークリッド距離に基づいて次式で定義する：

$$d_\mu := \|\boldsymbol{\mu}(\boldsymbol{X}_1) - \boldsymbol{\mu}(\boldsymbol{X}_2)\|_2^2 / 4 \tag{5}$$

テキスト埋め込みの慣例として一般的なノルム正規化 $\|\boldsymbol{\mu}(\boldsymbol{X}_i)\| = 1$ を仮定すると，$d_\mu \in [0, 1]$[2] となり，また $d_\mu$ はテキスト埋め込みの標準的な距離尺度に対応する．また，$d_\Sigma$ は二次統計量 $\boldsymbol{\Sigma}(\boldsymbol{X}_1)$ と $\boldsymbol{\Sigma}(\boldsymbol{X}_2)$ の距離であり，ブレスワッサースタイン距離 [20] に基づいて次式で定義する：

$$d_\Sigma := \mathrm{tr}\big(\boldsymbol{\Sigma}(\boldsymbol{X}_1) + \boldsymbol{\Sigma}(\boldsymbol{X}_2) - 2\big(\boldsymbol{\Sigma}(\boldsymbol{X}_1)^{1/2}\boldsymbol{\Sigma}(\boldsymbol{X}_2)\boldsymbol{\Sigma}(\boldsymbol{X}_1)^{1/2}\big)^{1/2}\big)/4. \tag{6}$$

ここで，比較を簡単にするために，$d_\Sigma$ を $d_\mu$ と同じ範囲にするために $\mathrm{tr}(\boldsymbol{\Sigma}(\boldsymbol{X}_i)) \le 2$ を仮定する．この仮定は，単語埋め込み集合の分散が極端に大きくないという § 4.1 の仮定にも対応する．このとき，$d_\Sigma \in [0, 1]$[3] となる．$d_\mu, d_\Sigma \in [0, 1]$ と式 (4) から $\mathrm{SOCM} \in [0, 1]$ であり，値が大きいほど崩壊の度合いも大きいことを表す．§ 4.1 で述べたように，崩壊は一次統計量が似ていて（$d_\mu$ が小さい）かつ二次統計量が異なる（$d_\Sigma$ が大きい）ときに生じる．SOCM は $(1 - d_\mu)d_\Sigma$ という形により，この直感を反映する．以降では，この SOCM の設計の妥当性を議論する．

**統計量同士の距離**　一次と二次の統計量の同士距離として $d_\mu$ と $d_\Sigma$ を採用した．これは単語埋め込み集合の違いが一次の統計量の違い $d_\mu$ と二次の統計量の違い $d_\Sigma$ に分解できることに基づく．形式的には，$d_\mu$ と $d_\Sigma$ は $L_2$-ワッサースタイン距離 $W_2^2$ を分解する [20]：

$$W_2^2\big(\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{X}_1), \boldsymbol{\Sigma}(\boldsymbol{X}_1)), \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{X}_2), \boldsymbol{\Sigma}(\boldsymbol{X}_2))\big)/4 = d_\mu + d_\Sigma. \tag{7}$$

ここで簡単のため，単語埋め込み集合 $\boldsymbol{X}_i$ を，一次および二次の統計量で定まる正規分布 $\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{X}_i), \boldsymbol{\Sigma}(\boldsymbol{X}_i))$ で特徴づける．ワッサースタイン距離は，正規分布などの確率分布間の差異を，高次の統計量まで考慮して測る標準的な距離尺度である [21]．また，§ 2 で述べた，テキスト間の意味類似度計算するための単語埋め込み集合表現同士の最適輸送コストもこのワッサースタイン距離に基づく [22]．以上より，SOCM を構成する一次および二次の統計量間の距離として $d_\mu$ と $d_\Sigma$ を選択した．

**SOCM の形式の妥当性**　採用した SOCM の妥当性を確かめるため，平均プーリングによる二次統計量の崩壊を定量化する指標に望ましい性質を整理し，式 (4) がそれらを満たすことを示す．まず，次の (a)–(e) の 5 つを望ましい性質として考える：

(a) **崩壊条件：** $d_\mu = 0 \wedge d_\Sigma = 1 \Leftrightarrow \mathrm{SOCM} = 1$.
§ 4.1 で述べたように，一次統計量が一致し（$d_\mu = 0$），二次統計量が最大限異なる（$d_\Sigma = 1$）ときに，崩壊の度合いは最大となる．

(b) **非崩壊条件：** $d_\mu = 1 \vee d_\Sigma = 0 \Leftrightarrow \mathrm{SOCM} = 0$.
§ 4.1 で述べたように，一次統計量が最大限異なる（$d_\mu = 1$）か，二次統計量が一致する（$d_\Sigma = 0$）とき，崩壊は生じない．

(c) $d_\mu$ **に関する単調性：** $\dfrac{\partial\,\mathrm{SOCM}}{\partial d_\mu} \le 0$.
一次統計量が異なる（$d_\mu$ が大きくなる）と崩壊の度合いは小さくなる．

(d) $d_\Sigma$ **に関する単調性：** $\dfrac{\partial\,\mathrm{SOCM}}{\partial d_\Sigma} \ge 0$.

2)　1/4 は $d_\mu$ を $[0, 1]$ に収めるためのスケーリングである．

3)　1/4 はトレース上界の下で $d_\Sigma$ を $[0, 1]$ に収めるためのスケーリングである．

二次統計量が異なる（$d_\Sigma$ が大きくなる）と崩壊の度合いは大きくなる．

(e) $d_\mu$ **と** $d_\Sigma$ **の相互作用：** $\dfrac{\partial^2 \mathrm{SOCM}}{\partial d_\mu \, \partial d_\Sigma} \le 0.$
一次の統計量が異なるときは二次の統計量が異なっても崩壊の度合いは大きくなりにくい

式 (4) で定義される SOCM は，性質 (a)–(e) のすべてを満たす．証明は付録 A に示す．図 2 は $d_\mu$ と $d_\Sigma$ に対する SOCM の値を示しており，性質 (a)–(e) が満たされいていることが確認できる．また，§ 4.1 で述べた崩壊が生じる状況では SOCM の値は大きく，それ以外では小さいことも確認できる．

# 5 実験

この SOCM を用いて，実テキストと実モデルで平均プーリングによる二次統計量の崩壊がどの程度生じているかを確認する．

## 5.1 実験手法

**実験手順** 実験では，テキストペア集合 $D = \{(t_1, t_2)\}$ とモデル $\boldsymbol{f}$ を用いる．テキスト $t_i$ を $\boldsymbol{f}$ に入力し，単語埋め込み集合 $\boldsymbol{X}_i = \boldsymbol{f}(t_i)$ を得る．テキストペア $(t_1, t_2)$ について，対応する単語埋め込み集合 $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ から SOCM を計算する．

**単語埋め込み集合の正規化** SOCM を計算する前に単語埋め込み集合 $\boldsymbol{X}_i$ を正規化する．SOCM の定義では $\|\boldsymbol{\mu}(\boldsymbol{X}_i)\| = 1$ が仮定され，これはノルム正規化されたテキスト埋め込み，つまり $\dfrac{\boldsymbol{\mu}(\boldsymbol{X}_i)}{\|\boldsymbol{\mu}(\boldsymbol{X}_i)\|}$ を用いる慣例に対応する．この仮定を満たすため，$\boldsymbol{X}_i$ に対し，次の正規化後の集合 $\boldsymbol{X}_i^{\mathrm{norm}}$ を用いる：

$$\boldsymbol{X}_i^{\mathrm{norm}} = \left[ \frac{\boldsymbol{x}_{i,1}}{\|\boldsymbol{\mu}(\boldsymbol{X}_i)\|}, \ldots, \frac{\boldsymbol{x}_{i,n_i}}{\|\boldsymbol{\mu}(\boldsymbol{X}_i)\|} \right] \in \mathbb{R}^{d \times n_i}. \quad (8)$$

このとき $\boldsymbol{\mu}(\boldsymbol{X}_i^{\mathrm{norm}}) = \dfrac{\boldsymbol{\mu}(\boldsymbol{X}_i)}{\|\boldsymbol{\mu}(\boldsymbol{X}_i)\|}$ が成り立つ [4]．この $(\boldsymbol{X}_1^{\mathrm{norm}}, \boldsymbol{X}_2^{\mathrm{norm}})$ に対して SOCM を計算する．

**データセット** 100 万個のテキストが含まれる Wikipedia [16] からテキストペア集合を構築した．1,000 個のテキストを乱択し，それらをペアワイズに比較して 499,500 個のテキストペアを生成した．

**モデル** 近年のテキスト埋め込みモデルとして標準的な事前学習済みモデルを対照学習によって微調整したテキスト埋め込みモデルを選択した．具体的には，Unsupervised-SimCSE [16]，E5 [17]，GTE [2] を使用した[5]．モデルの詳細は付録に示した．また，これらのベースモデルである BERT [1] を使用した．

4） 証明は付録 B に示した
5） すべてのモデルはベースサイズである．

**表 1** 各モデルにおける SOCM の平均．ベースのモデルから派生したテキスト埋め込みモデルについて，括弧内の値は SOCM の変化を示す．太字の値は減少を示す．

| モデル | SOCM の平均↓ |
|---|---|
| BERT | 0.402 |
| → Unsup-SimCSE-mean | **0.197** (−0.205) |
| → E5 | **0.036** (−0.366) |
| → GTE | **0.025** (−0.377) |

## 5.2 実験結果

**定量分析** 表 1 に各モデルの SOCM 平均を示す．微調整後のモデルでは，ベースモデルに比べて SOCM が小さい傾向が見られた．これは，微調整後のモデルでは平均プーリングによる二次統計量の崩壊が生じにくいことを示唆する．

**定性分析** 実テキストペアに対して，BERT と GTE の単語埋め込みを可視化した一例を図 1 下部に示す．この例では，二つのテキスト間の意味的な関連は低い．BERT では SOCM = 0.618 と大きく[6]，可視化上でも単語埋め込み集合の広がり方は異なるがその平均は近くなっていた．一方，$\mathrm{GTE}_{\mathrm{base}}$ では SOCM = 0.024 と小さく，平均によって二つの分布が分離されている．この可視化例からも，微調整後のモデルでは平均プーリングによる二次統計量の崩壊が生じにくいことが確認できる．

これらの結果は，近年のモデルにおける平均プーリングの有用性を示唆するものである．平均プーリングは一見粗い集約手法に思えるが，近年のテキスト埋め込みモデルは実際には高い性能を示している．本稿の分析は，近年のモデルでは，このような粗い手法のように思える平均プーリングでも，情報が失われにくくなることを示唆している．

# 6 おわりに

本稿では，平均プーリングによって単語埋め込み集合の二次以降の統計量が失われる点に着目した．まず，この平均プーリングによる二次統計量の崩壊を定量化した．実験により，事前学習済みモデルを対照学習で微調整したモデルでは，ベースモデルよりもこの崩壊が生じづらいことを経験的に確認した．今回の経験的な結果がなぜ生じたのかを理論付けることは今後の興味深い方向の一つである．

6） SOCM は次元削減前の 768 次元で計算している．

## 謝辞

本研究は，AMED の課題番号 JP25wm0625405，JSPS 科研費 22H05106，および JST 創発 JPMJFR2331 の助成を受けたものです．

## 参考文献

[1] 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, **Proceedings of the 2019 Conference of the North**, pp. 4171–4186, Stroudsburg, PA, USA, June 2019. Association for Computational Linguistics.

[2] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, and et al. Towards general text embeddings with multi-stage contrastive learning, 2023.

[3] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In **Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)**, 2021.

[4] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

[5] Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder. **Transactions on Machine Learning Research**, February 2025.

[6] Sotaro Takeshita, Yurina Takeshita, Daniel Ruffinelli, and Simone Paolo Ponzetto. Randomly removing 50% of dimensions in text embeddings has minimal impact on retrieval and classification tasks. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 27693–27714, Stroudsburg, PA, USA, 2025. Association for Computational Linguistics.

[7] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, and et al. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, Stroudsburg, PA, USA, 2018. Association for Computational Linguistics.

[8] Ashish Vaswani, Noam M Shazeer, Niki Parmar, Jakob Uszkoreit, and et al. Attention is all you need. **Neural Inf Process Syst**, Vol. 30, pp. 5998–6008, June 2017.

[9] Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, and et al. Gemini embedding: Generalizable embeddings from gemini, March 2025.

[10] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In Francis Bach and David Blei, editors, **Proceedings of the 32nd International Conference on Machine Learning**, Vol. 37 of **Proceedings of Machine Learning Research**, pp. 957–966, Lille, France, 2015. PMLR.

[11] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, and et al. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **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. 563–578, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics.

[12] Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. Word rotator's distance. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2944–2960, Online, November 2020. Association for Computational Linguistics.

[13] Seonghyeon Lee, Dongha Lee, Seongbo Jang, and Hwanjo Yu. Toward interpretable semantic textual similarity via optimal transport-based contrastive sentence learning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5969–5979, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics.

[14] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3715–3734, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics.

[15] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **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. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[16] 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.

[17] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, and et al. Text embeddings by weakly-supervised contrastive pre-training, 2024.

[18] Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In **Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval**, New York, NY, USA, July 2020. ACM.

[19] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, and et al. The faiss library, 2025.

[20] D C Dowson and B V Landau. The fréchet distance between multivariate normal distributions. **J. Multivar. Anal.**, Vol. 12, No. 3, pp. 450–455, September 1982.

[21] Cedric Villani. **Optimal Transport: Old and New**. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, Germany, December 2009.

[22] Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020.

[23] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, and et al. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In **Advances in Neural Information Processing Systems**, February 2020.

# A　SOCM の証明

本節では，§ 4.2 で導入した SOCM

$$\mathrm{SOCM}(d_\mu, d_\Sigma) = (1 - d_\mu)d_\Sigma \tag{9}$$

が望ましい性質 (a)–(e) を満たすことを証明する：

(a) $d_\mu = 0 \wedge d_\Sigma = 1 \Leftrightarrow \mathrm{SOCM} = 1$
(b) $d_\mu = 1 \vee d_\Sigma = 0 \Leftrightarrow \mathrm{SOCM} = 0$
(c) $\dfrac{\partial\,\mathrm{SOCM}}{\partial d_\mu} \le 0$
(d) $\dfrac{\partial\,\mathrm{SOCM}}{\partial d_\Sigma} \ge 0$
(e) $\dfrac{\partial^2\,\mathrm{SOCM}}{\partial d_\mu\,\partial d_\Sigma} \le 0$

**性質 (a) の証明**　性質 (a)：$d_\mu = 0 \wedge d_\Sigma = 1 \Leftrightarrow \mathrm{SOCM} = 1$ を証明する．$d_\mu = 0$ かつ $d_\Sigma = 1$ のとき，$\mathrm{SOCM}(0, 1) = 1 \cdot 1 = 1$．逆に $\mathrm{SOCM} = 1$ とすると，$(1 - d_\mu)d_\Sigma = 1$．$d_\mu, d_\Sigma \in [0, 1]$ より $(1 - d_\mu), d_\Sigma \in [0, 1]$ なので，この等式が成り立つのは $(1 - d_\mu) = 1$ かつ $d_\Sigma = 1$ のときのみ．したがって $d_\mu = 0$ かつ $d_\Sigma = 1$．

**性質 (b) の証明**　性質 (b)：$d_\mu = 1 \vee d_\Sigma = 0 \Leftrightarrow \mathrm{SOCM} = 0$ を証明する．$d_\mu = 1$ のとき，$\mathrm{SOCM}(1, d_\Sigma) = 0 \cdot d_\Sigma = 0$．$d_\Sigma = 0$ のとき，$\mathrm{SOCM}(d_\mu, 0) = (1 - d_\mu) \cdot 0 = 0$．逆に $\mathrm{SOCM} = 0$ とすると，$(1 - d_\mu)d_\Sigma = 0$ なので，$(1 - d_\mu) = 0$ または $d_\Sigma = 0$，すなわち $d_\mu = 1$ または $d_\Sigma = 0$．

**性質 (c) の証明**　性質 (c)：$\frac{\partial\,\mathrm{SOCM}}{\partial d_\mu} \le 0$ を証明する．

$$\frac{\partial\,\mathrm{SOCM}}{\partial d_\mu} = \frac{\partial}{\partial d_\mu}[(1 - d_\mu)d_\Sigma] = -d_\Sigma \le 0 \tag{10}$$

$d_\Sigma \in [0, 1]$ より $-d_\Sigma \le 0$ が成り立つ．

**性質 (d) の証明**　性質 (d)：$\frac{\partial\,\mathrm{SOCM}}{\partial d_\Sigma} \ge 0$ を証明する．

$$\frac{\partial\,\mathrm{SOCM}}{\partial d_\Sigma} = \frac{\partial}{\partial d_\Sigma}[(1 - d_\mu)d_\Sigma] = (1 - d_\mu) \ge 0 \tag{11}$$

$d_\mu \in [0, 1]$ より $(1 - d_\mu) \ge 0$ が成り立つ．

**性質 (e) の証明**　性質 (e)：$\frac{\partial^2\,\mathrm{SOCM}}{\partial d_\mu\,\partial d_\Sigma} \le 0$ を証明する．性質 (c) より $\frac{\partial\,\mathrm{SOCM}}{\partial d_\mu} = -d_\Sigma$ なので，

$$\frac{\partial^2\,\mathrm{SOCM}}{\partial d_\mu\,\partial d_\Sigma} = \frac{\partial}{\partial d_\Sigma}[-d_\Sigma] = -1 \le 0 \tag{12}$$

が成り立つ．

# B　正規化の性質の証明

§ 5 で定義された正規化のもとで，$\boldsymbol{\mu}(\boldsymbol{X}_i^{\mathrm{norm}}) = \boldsymbol{\mu}(\boldsymbol{X}_i)/\|\boldsymbol{\mu}(\boldsymbol{X}_i)\|$ という関係が成り立つことを証明する．単語埋め込みの集合 $\boldsymbol{X}_i = [\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,n_i}] \in \mathbb{R}^{d \times n_i}$ とその正規化されたバージョン

$$\boldsymbol{X}_i^{\mathrm{norm}} = \left[\frac{\boldsymbol{x}_{i,1}}{\|\boldsymbol{\mu}(\boldsymbol{X}_i)\|}, \ldots, \frac{\boldsymbol{x}_{i,n_i}}{\|\boldsymbol{\mu}(\boldsymbol{X}_i)\|}\right] \in \mathbb{R}^{d \times n_i} \tag{13}$$

について，平均プーリングの定義より

$$\boldsymbol{\mu}(\boldsymbol{X}_i^{\mathrm{norm}}) = \frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\boldsymbol{x}_{i,j}}{\|\boldsymbol{\mu}(\boldsymbol{X}_i)\|} \tag{14}$$

$$= \frac{1}{\|\boldsymbol{\mu}(\boldsymbol{X}_i)\|} \cdot \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_{i,j} = \frac{\boldsymbol{\mu}(\boldsymbol{X}_i)}{\|\boldsymbol{\mu}(\boldsymbol{X}_i)\|} \tag{15}$$

**表 2**　Wikipedia のテキストペアに対する各モデルの平均 SOCM 値．ベースのモデルから派生したテキスト埋め込みモデルについて，括弧内の値は SOCM の変化を示す．太字の値は減少を示す．

| モデル | SOCM の平均 ↓ |
|---|---|
| MiniLM | 0.286 |
| → all-MiniLM-L12-v2 | 0.313 (+0.027) |
| → $\mathrm{E5}_{\mathrm{small}}$ | **0.099** (−0.187) |
| → $\mathrm{GTE}_{\mathrm{small}}$ | **0.055** (−0.231) |
| nomic-bert-2048 | 0.139 |
| → nomic-embed-text-v1.5 | **0.122** (−0.017) |

が成り立つ．さらに，これは $\|\boldsymbol{\mu}(\boldsymbol{X}_i^{\mathrm{norm}})\| = 1$ を意味し，SOCM の定義に使用した一次の統計量のノルムが 1 であることを満たしている．

# C　追加モデルでの実験結果

本節では，§ 5 で使用したモデルに加えて，§ 5 では，モデルとして，BERT をベースに対照学習で微調整したテキスト埋め込みモデルを対象とした．本節では，異なる事前学習済みモデルをベースとして対照学習で微調整したテキスト埋め込みモデルに対して実験を行う．その結果，§ 5 と似た結果が得られた．

**モデル**　平均プーリングを用いる近年の代表的なテキスト埋め込みモデルとして，all-MiniLM-L12-v2 [15]，$\mathrm{E5}_{\mathrm{small}}$ [17]，$\mathrm{GTE}_{\mathrm{small}}$ [2]，nomic-embed-text-v1.5 [5] を選択した．これらのモデルはいずれも，事前学習済み言語モデルを対照学習によって微調整されている．対応するバックボーンモデルとして，MiniLM [23]，nomic-bert-2048 [5] を使用した．

**実験手順**　実験手順は § 5 と同一である．Wikipedia から構築した 499,500 のテキストペアに対して各モデルの SOCM 値を計算した．

**結果**　Table 2 に各モデルの平均 SOCM 値を示す．§ 5 の結果と同様に，対照学習によって微調整されたテキスト埋め込みモデルではそのベースのモデルよりも SOCM の値が減少する傾向にあった．これは微調整されたテキスト埋め込みモデルでは，そのベースのモデルよりも平均プーリングによる二次統計量の崩壊が生じにくいことを示唆している．一方で，細かく結果を分析すると，微調整前後のテキスト埋め込みモデルでは SOCM の値の変化はモデルペアによって異なっていた．例えば，MiniLM→$\mathrm{GTE}_{\mathrm{small}}$ では − 0.231 の減少が見られた．一方，MPNet→all-mpnet-base-v2 では − 0.017 と比較的変化幅が小さいケースも存在した．これは，MPNet が既にベースモデルの段階で比較的小さい SOCM 値（0.117）を示しており，微調整による改善の余地が限られていたためと考えられる．