文埋め込みにおける意味とスタイルの相違性

山内悠輔 [1]　相澤彰子 [2,1]
[1] 東京大学 大学院情報理工学系研究科　[2] 国立情報学研究所
{yamauchi_y, aizawa}@nii.ac.jp

概要

Semantic Textual Similarity (STS) タスクは文ベクトルの評価に用いられる自然言語処理の基盤的なタスクである。”文章の類似性”の定義は本質的に曖昧であり、データセットによって異なるが、多様な種類のデータセット間の包括的な分析を行った研究は限られており、言語モデルが人間と同様に多様な種類の意味やスタイルの違いを捉えているかどうかは不明である。本研究では、凍結した Encoder に軽量な pooler を取り付ける統一的フレームワークを導入し、STS、Paraphrase identification (PI)、Triplet データセット間で統一的な分析を行う。21 種類のデータセットを用いた実験の結果、意味の概念は STS/PI タスク間で高い相関を示す一方、スタイルは意味的類似性とは独立した概念であり、明示的な分離が必要であることが分かった。

1　はじめに

Semantic Textual Similarity (STS) [1, 2] は文章間の意味的な近さを表し、情報検索 [3] や文書クラスタリング [4]、重複検出 [5] など様々な分野に応用され、STS タスクは文埋め込みモデルの性能を測るベンチマークとして広く採用されてきた。「しかし、意味的な近さという概念は本質的に曖昧で、データセットごとに定義が異なる。また、言語モデルはコーパス中のトークンの共起頻度に強く影響され、人間とは異なる基準で単語の意味をとらえていると報告されている [6]。したがって、言語モデルが文章の意味をどのように捉えているかは明確ではない。

こうした背景から、近年は与えられた文対に対して言語モデルがなぜ高いあるいは低い類似度を予測したのかを解釈・説明する研究が注目されている。[7]。既存研究では、予測類似度に対するトークンペア間の寄与を算出する手法 [8, 9, 10] や、モデルの埋め込みの各次元に事前に意味を割り当てて学習を行う手法 [11, 12] などが提案されている。これらの研究は STS タスクの透明性を向上するものであるが、一方で扱うデータセットがイメージキャプションや表層的な変換に限られるなどスコープが限定的である。文章の違いには語彙や構造の変化だけでなく、数値表現や文体など多様な要素が含まれる。したがって、モデルがこれらの違いをどのように捉えているかを分析するには、様々なドメインやタスクから収集したデータセットによる包括的な検証が必要である。本研究では異なる様式を持つデータセット間での統一的かつ包括的な比較を行うフレームワークを提案する。本アプローチでは、凍結したエンコーダに軽量な pooler を取り付け、その出力をコサイン類似度による共通スケールで比較することで柔軟な分析を可能にする。

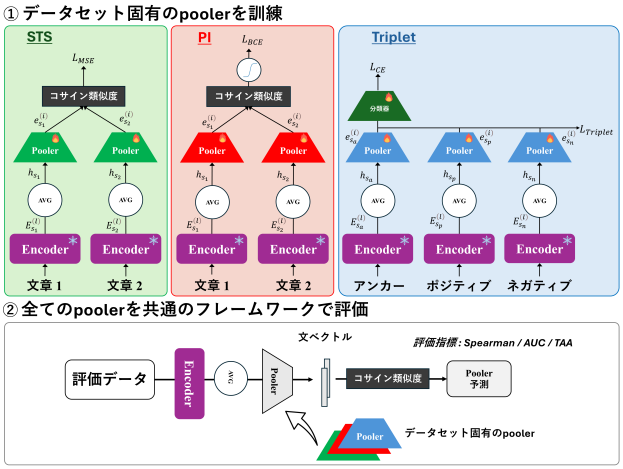

図 1　本研究の訓練・評価フレームワーク。(1) データセット固有の文章の違いを学習し、(2) 共通の推論設定で比較を行う。

本研究の貢献は次の三点である。1) 21 種類の文章類似度データセットを収集し、共通のフレームワーク上で各データセットが定義する意味やスタイルを比較可能にした。2) 意味の違いは STS データセットと PI データセットの間で概ね共通の概念として捉えられる一方、スタイルは意味とは独立した概念として捉えられることを発見した。3) 階層クラ

表 1　各データセットが定義する文章間の違いの例。

データセット	文章 1	文章 2	文章間の違い
PAWS-Wiki	It was chosen as the 19th best movie at the 7th Yokohama Film Festival .	It was chosen as the 7th best film at the 19th Yokohama Film Festival .	数の値
QQP	How do I fill in Address Line 1 and Address Line 2?	How do I register desired web address?	同じ質問であるか

データセット	アンカー	ポジティブ	ネガティブ	文章間の違い
APPDIA	So sad nobody boycotts this shit.	Taking it easy fuck that ! ! winning	So sad no body boycott this	意味 または スタイル (中立/毒性)

スタリング分析により、意味とスタイルの表現の違いを特定した。[1)]

2　手法

本論文では Text Style Transfer での定義 [13, 14] に倣い、”意味”を話し手が伝えたい意図や文章の内容、”スタイル”を同じ文章の内容をどのように伝えるか (例: 丁寧に、ユーモラスに、感情的に) として定義する。まず本研究で収集したデータセットの概要について説明し (節 2.1)、提案フレームワークの詳細を節 2.2 で説明する。

2.1　データセット

本研究で収集した 21 種類のデータセットは、STS、Paraphrase Identification (PI)、Triplet の 3 つのカテゴリに分類される。各データセットは文のペアまたは三つ組から構成され、それぞれ異なる観点から文章間の違いを定義している (表 1)。全ての STS と PI データセットは文の意味的同等性を測るが、各データセットで意味の定義基準は異なる。例えば、PAWS-Wiki は文中の数値を重要視し、QQP は二つの文が同じ質問かどうか（同じ答えが期待できるかどうか）を“意味”として定義している。Triplet データセットは各サンプルがアンカー文、ポジティブ文、ネガティブ文の三文から構成される。本研究で収集した Triplet データセットでは、アンカーとポジティブは同じスタイルだが意味が異なり、アンカーとネガティブは同じ意味だがスタイルが異なるように作成した（ELSA-Emotion は意味もスタイルも異なる）。この構成により、モデルがどの要素を類似性の基準と見なすかを検証できる。各データセットは訓練・テストセットを結合した後に 8:2 の割合で再分割し、訓練セットは最大 5,000 件となるようにダウンサンプリングした。各データセットの詳細は付録 A に記載する。

1)　本研究の内容は、国際学会 EACL 2026 Findings に採択されたものに基づく。

2.2　データセット固有の pooler

データセット間で包括的な検証を行うため、本研究では Bi-Encoder [15] 構造を用いて各文を独立に埋め込みへ変換する。各データセットはカテゴリごとに異なる目的関数で学習し、共通の評価指標で評価する (図 1)。まず共通の Encoder で文をトークンベクトルの系列 $E_s^{(l)} \in \mathbb{R}^{T\times d_{\text{emb}}}$ に変換する。形式的に、Encoder モデルの l 層目の文章 s の埋め込み表現を $E_s^{(l)} \in \mathbb{R}^{T\times d_{\text{emb}}}$ と表記する。(T は系列長、d_{emb} は次元数)

平均プーリングを $E_s^{(l)}$ に適用して固定長の文ベクトル $\mathbf{h}_s$ を得る。この文ベクトルは**データセット固有の pooler** により、共通の埋め込み空間へ投影される:

$$\mathbf{e}_s^{(i)} = W_i\,\mathbf{h}_s + \mathbf{b}_i, \quad W_i \in \mathbb{R}^{d_p\times d_{\text{emb}}},\ \mathbf{b}_i \in \mathbb{R}^{d_p} \tag{1}$$

ここで i はデータセットのインデックス、d_p は pooler の次元数を表し、各データセット D_i の pooler は $(W_i, \mathbf{b}_i)$ を学習可能なパラメータとして持つ。以下に STS、PI、Triplet の訓練・推論パラダイムについて説明する。

(1) STS: 各文章ペア $\left(s_1^{(k)}, s_2^{(k)}\right) \in D_{i,\text{sts}}$ は連続値の教師ラベル $y_k \in [0,1]$ を持ち、予測類似度は次式で与えられる:

$$\hat{y}_k = \cos\left(\mathbf{e}_{s_1}^{(k)}, \mathbf{e}_{s_2}^{(k)}\right) \tag{2}$$

STS では平均二乗誤差 (MSE) 損失を用いて pooler を訓練し、評価には Spearman の順位相関係数を用いる

(2) PI: 各文章ペア $\left(s_1^{(k)}, s_2^{(k)}\right) \in D_{i,\text{pi}}$ は二値の教師ラベル $y_k \in \{0,1\}$ を持ち、予測類似度はコサイン類似度の出力にシグモイド関数を適用することで得られる:

$$\hat{y}_k = \sigma\left(\cos\left(\mathbf{e}_{s_1}^{(k)}, \mathbf{e}_{s_2}^{(k)}\right)\right) \tag{3}$$

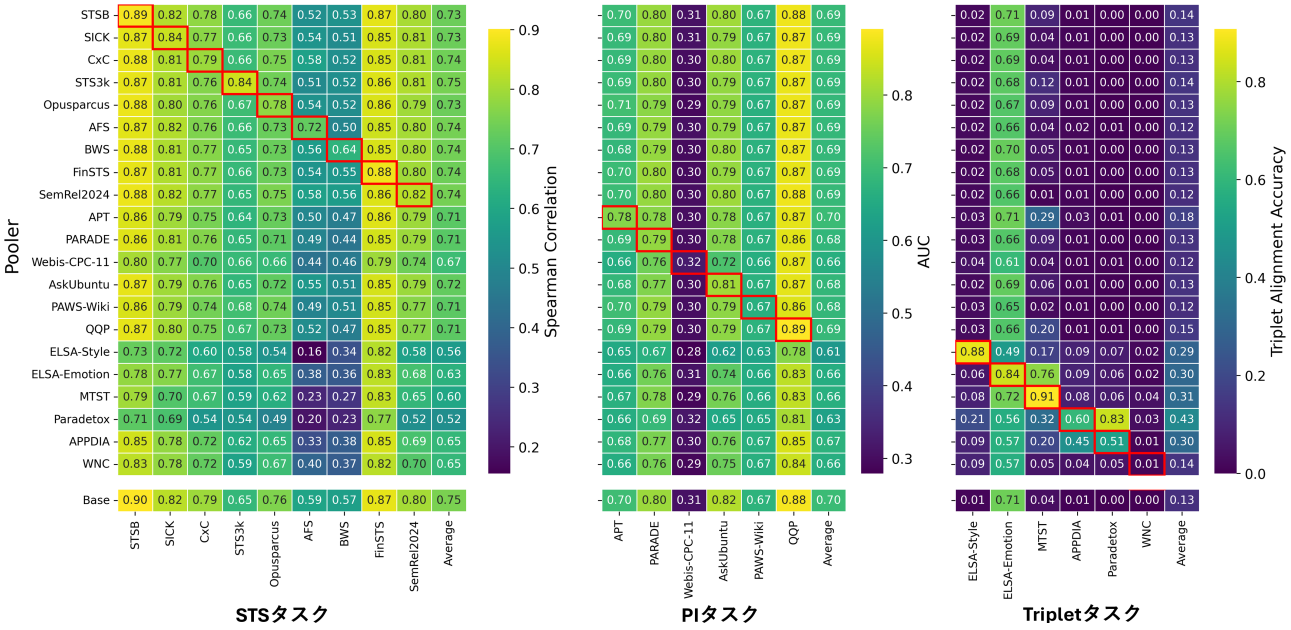

図 2　各 pooler の STS, PI, Triplet タスクに対するスコア。各タスクは異なる評価指標を適用し、セル内の数値が各 pooler のスコアを示す (高いほど良い)。***Base*** は凍結された Encoder モデルを表す。訓練データセットと評価タスクが同一の pooler のスコアは赤枠で囲まれている。

PI ではシグモイド関数の出力に対しバイナリクロスエントロピー損失 (BCE) で学習し、評価には ROC 曲線下面積 (AUC) を用いる。

(3) Triplet: 各サンプル $(s_a, s_p, s_n) \in D_{i,\text{trip}}$ は三文で構成され、訓練時の損失関数にはマージン付きトリプレット損失 loss [16] を適用する:

$$L_{\text{Triplet}} = \max\bigl(0,\ m + \cos(\mathbf{e}_{s_a}, \mathbf{e}_{s_n}) - \cos(\mathbf{e}_{s_a}, \mathbf{e}_{s_p})\bigr) \tag{4}$$

m はマージンハイパーパラメータである。加えて、アンカーベクトル $\mathbf{e}_{s_a}$ のスタイルラベルを softmax 分類器で予測し、クロスエントロピー損失 L_{CE} で学習する。最終的な損失は重み α を用いて $L_{\text{CE}} + \alpha\, L_{\text{Triplet}}$ で表される。評価指標は Triplet alignment accuracy (TAA) [17] を適用し、これはテストセット中の $\cos(\mathbf{e}_{s_a}, \mathbf{e}_{s_p}) > \cos(\mathbf{e}_{s_a}, \mathbf{e}_{s_n})$ となるサンプル数の割合を算出する。

3　実験

3.1　実験設定

Encoder モデルは BERT-large を基にした mxbai-embed-large-v1 [18] を使用する。pooler の出力次元は $d_p = 256$ マージン m は 0.2、損失重み α は 1.0 とした。その他の設定は付録 B に記す。

3.2　結果

まず、文が同じ意味を持つかどうかを判定するデータセット群に対して、意味的類似度の概念が共通かどうかを分析する。図 2 に示すように、Encoder 単体でも高い精度を示しており、意味的類似度に関する基本的な概念がデータセット間で共通していることがうかがえる。AFS・BWS・Webis-CPC-11 ではスコアが低く、モデルが意味的な違いを捉えにくい概念も存在した。一方、他の多くのデータセットでは未学習の状態でも一貫して高いスコアが得られた。より詳細にモデルの予測傾向を検証するために各文ペアに対する類似度予測について pooler 間の相関係数を計算したところ、すべての STS および PI データセットで学習した pooler 同士の相関が 0.9 以上となった (付録 C)。この結果は、pooler が普遍的な意味的類似性は捉えているものの、微細な意味の違いを十分に捉えきれていないことを示す。データセット固有の意味的差異を表現するには、より洗練された手法が必要である。

Triplet タスクに着目すると、STS や PI タスクとは結果が大きく異なり、対応するデータセットで訓練した pooler 以外は低いスコアにとどまった。さらに、損失関数の重み α を変化させて意味とスタイルの分離を強調したところ、スタイル分類精度が向上

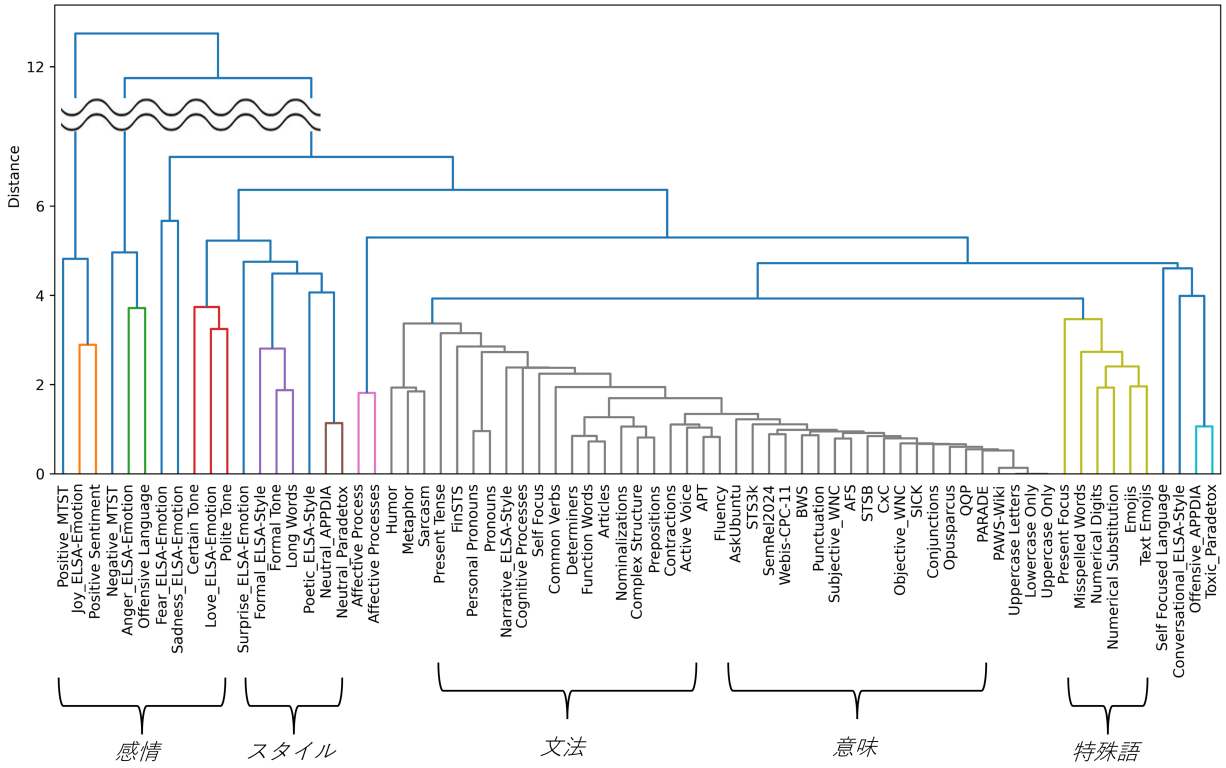

図 3　Mean Difference ベクトルを用いた階層クラスタリングの結果 (デンドログラム)

する傾向が観察された。これらの結果は、各スタイルが意味や他のスタイルとは独立した概念であり、モデルがこれらを区別して捉えるためには明示的な学習が必要であることを示唆している。

3.3　分析

概念間の関係性をより明示的に分析するために、デンドログラムによるモデルの概念構造の可視化を行う。より具体的には、各データセットの文章間の差異を代表するベクトルを算出し、階層クラスタリングを適用する。まず、各データセットから文章が異なるとラベル付けされた文章ペアのサンプル集合 D_i^- を構築する。STS データセットからは文章類似度 y_k が下位 100 件のサンプル、PI データセットからは $y_k = 0$(言い換えではない) 文章ペアをランダムに 100 件、Triplet データセットからはアンカーとネガティブのペアをランダムに 100 組ずつサンプリングする。このサンプル集合を基に、以下で定義される *Mean Difference*(MD) ベクトルを算出する。

$$h_{\mathrm{MD}_i} = \frac{1}{K} \sum_{n \in D_i^-} \left(h_{s_1}^{(n)} - h_{s_2}^{(n)} \right) \tag{5}$$

直感的に、このベクトルは各データセット内で一貫する意味やスタイルの違いを表すベクトルであると考えられる。このベクトルは文対の順序に依存するため、Triplet データセットではアンカーとネガティブを入れ替えた場合も別の MD ベクトルとして扱った (StyleDistance を除く)。最後に、全ての MD ベクトルに対して Ward 法による階層クラスタリングを適用し、結果のデンドログラムを可視化する。図 3 はその結果を示し、各クラスタが MD ベクトルに対応しベクトル間の距離が近いものから順にマージされる。デンドログラムでは意味的な概念同士が最初に結合し、次いで文法・特殊語・スタイル・感情の順に結合した。この結果は、意味的な違いが多様でありながら互いに密接な関係にある一方、スタイルや感情の違いは意味とは明確に区別されることを示している。

4　おわりに

本研究では 21 種類の文章類似度に関するデータセットを収集し、言語モデルがこれらのデータセットで定義される文章の違いをどのように捉えているかを分析した。意味とスタイルは大きく異なる概念であり、単一の埋め込み空間で双方を同時に扱うことは難しい。そのため、本研究の pooler のように異なるモジュールを用いて多角的に特徴を捉える設計が有効であると示唆される。

謝辞

本研究は、国立情報学研究所大規模言語モデル研究開発センターおよび JSPS 科研費 24K03231 の助成を受けたものである。

参考文献

[1] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret, editors, ***SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)**, pp. 385–393, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[2] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic textual similarity. In Mona Diab, Tim Baldwin, and Marco Baroni, editors, **Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity**, pp. 32–43, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[3] Hiroki Iida and Naoaki Okazaki. Incorporating semantic textual similarity and lexical matching for information retrieval. In Kaibao Hu, Jong-Bok Kim, Chengqing Zong, and Emmanuele Chersoni, editors, **Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation**, pp. 582–591, Shanghai, China, 11 2021. Association for Computational Linguistics.

[4] Muhammad Rafi and Mohammad Shahid Shaikh. An improved semantic similarity measure for document clustering based on topic maps, 2013.

[5] Preetam Prabhu Srikar Dammu and Omar Alonso. Near-duplicate question detection. In **Companion Proceedings of the ACM Web Conference 2024**, WWW '24, p. 493–496, New York, NY, USA, 2024. Association for Computing Machinery.

[6] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from bert for semantic textual similarity. **ArXiv**, Vol. abs/2011.05864, , 2020.

[7] Juri Opitz, Lucas Möller, Andrianos Michail, and Simon Clematide. Interpretable text embeddings and text similarity explanation: A primer, 2025.

[8] 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, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[9] Lucas Moeller, Dmitry Nikolaev, and Sebastian Padó. Approximate attributions for off-the-shelf Siamese transformers. In Yvette Graham and Matthew Purver, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2059–2071, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[10] Alexandros Vasileiou and Oliver Eberle. Explaining text similarity in transformer models. 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. 7859–7873, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[11] Juri Opitz and Anette Frank. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, **Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 625–638, Online only, November 2022. Association for Computational Linguistics.

[12] Yiqun Sun, Qiang Huang, Yixuan Tang, Anthony Kum Hoe Tung, and Jun Yu. A general framework for producing interpretable semantic text embeddings. **ArXiv**, Vol. abs/2410.03435, , 2024.

[13] Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. Text style transfer: A review and experimental evaluation, 2023.

[14] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep learning for text style transfer: A survey. **Computational Linguistics**, Vol. 48, No. 1, pp. 155–205, March 2022.

[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] Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. 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. 4892–4903, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[17] Yingchaojie Feng, Yiqun Sun, Yandong Sun, Minfeng Zhu, Qiang Huang, Anthony Kum Hoe Tung, and Wei Chen. Don't reinvent the wheel: Efficient instruction-following text embedding based on guided space transformation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 24511–24525, Vienna, Austria, July 2025. Association for Computational Linguistics.

[18] Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. Open source strikes bread - new fluffy embedding model, 2024.

表 2　各データセットの詳細な情報

データセット	カテゴリ	ソースドメイン	サンプル数	平均トークン数
STSB	*STS*	news headlines, video/image captions, NLI	8,577	10.2
SICK		video/image captions	9,840	9.6
CxC		video/image captions	88,052	10.4
STS3k		hand crafted by author	2,800	8.1
Opusparcus		OpenSubtitles database	2,900	4.7
AFS		Internet Argument Corpus	6,000	22.2
BWS		UKP's Argument Facets, IBM Debater data	3,400	26.7
FinSTS		Corporate Annual Report	3,988	26.9
SemRel2024		news, Wikipedia, SNS	8,350	12.2
APT	*PI*	MSRP, PPNMT	4,449	11.6
PARADE		Computer Science web platforms	10,182	17.0
Webis-CPC-11		Project Gutenberg	4,243	119.8
AskUbuntu		Stack Exchange technical forum	6,557	92.0
PAWS-Wiki		Wikipedia	65,345	21.4
QQP		Quora	404,308	11.1
ELSA	*Triplet*	GoEmotions	10,434	20.5
MTST		Yelp	2,000	8.9
Paradetox		Reddit, other online forums	11,927	10.3
APPDIA		Reddit	1,981	11.9
WNC		Wikipedia	6,990	21.3
StyleDistance		generated by GPT-4	4,000	14.7

A　データセットの詳細

表 2 に本研究で使用したデータセットのメタ情報の一覧を表示する。本研究で収集されたデータセットは異なる文章の違いを定義するように選別した。

B　訓練時の詳細情報

訓練には `mixedbread-ai/mxbai-embed-large-v1` を使用し、以下のハイパーパラメータを採用した。

- 精度: float16, エポック数: 10, 学習率: 1e-4
- バッチサイズ: 64, シード値: 42
- Pooler 次元数: $d_p = 256$

C　Pooler 間の予測類似度の相関

図 4 に本研究で収集したデータセットの全ペアに対して、各 pooler が予測した類似度の相関ヒートマップを可視化する。STS データセットと PI データセットで学習した pooler は学習前の Encoder モデルの予測類似度と非常に高い相関を示す一方で Triplet データセットで学習した pooler は相関が低くなる傾向にある。

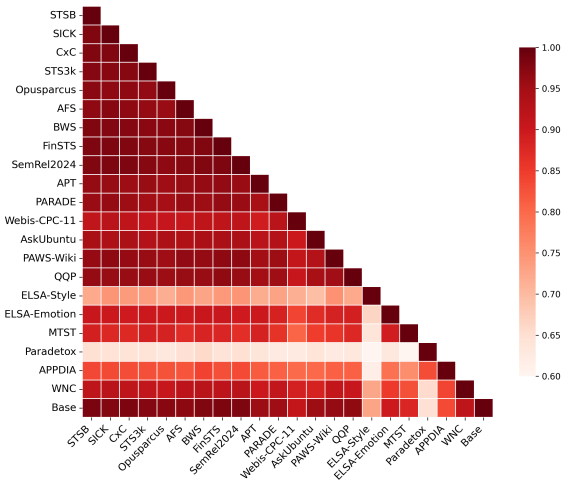

図 4　データセット全ペアにおける各 pooler 予測類似度の相関ヒートマップ