平均プーリングによる文埋め込みの再検討：平均は点群の要約として十分か？

原 知正 [1]　栗田 宙人 [1]　横井 祥 [1,2]　乾 健太郎 [3,1,2]
[1] 東北大学　[2] 理化学研究所　[3]MBZUAI
{hara.tomomasa.s8, hiroto.kurita.q4}@dc.tohoku.ac.jp
yokoi@tohoku.ac.jp　kentaro.inui@mbzuai.ac.ae

概要

文や文書のベクトル化は，検索拡張生成（RAG）をはじめとした広範な自然言語処理アプリケーションを実装するための基盤技術である．本稿では，文埋め込みの標準的な構成方法である平均プーリングが，構成要素となる単語埋め込み集合の持つ空間的な広がりの情報を潰し得る，という問題を指摘する．また実験により，上記の問題が実際のテキストと深層学習モデルで確かに生じていることと，一方でその割合は小さいことを示す．実験結果は平均プーリングの経験的な有用性を支持するものだが，同時に，文表現の構成方法を再検討する必要性を示唆するものである．

1　はじめに

文・段落・文書といった単語より大きな単位のテキストをベクトル化することで，情報検索や文書分類をはじめとする広範な自然言語処理タスクが，深層学習のフレームワークで統一的に解けるようになる [1, 2]．とりわけ最近では，大規模言語モデルに外部知識を効率的に取り込むための手法である検索拡張生成（RAG）においても，文埋め込みが主要な役割を担っている [3, 4]．

自然言語の表現学習における基本単位は単語（トークン）であり，文埋め込みも一般に単語の埋め込み表現から作られる．中でも最も標準的で人気のある方法は，文中の各単語埋め込みを平均することにより文埋め込みを構成する**平均プーリング**である．静的単語埋め込みに基づく手法群 [5] から動的単語埋め込みに基づく手法群 [6, 7] まで，これまで経験的な有用性が広く示されてきた．

人気の高い平均プーリングであるが，よく考えるといささか乱暴な計算方法に見える．問題の起きる

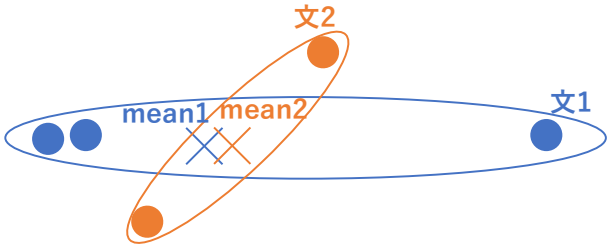

図 1　平均プーリングが点群の空間的な広がりを潰す例．**青色の点群**と**橙色の点群**はそれぞれ別の文の単語埋め込み集合である．これらの点群の配置は明らかに異なっているが，平均プーリングで点群を 1 点に集約することで文埋め込み同士はほとんど同じになってしまう．

例として，図 1 に，二つの単語埋め込みの集合と，これらを平均プーリングでそれぞれ 1 点に集約した様子を示す．図の**青色の点群**と**橙色の点群**（単語埋め込み集合）の配置は明らかに異なり，つまり文を構成する単語集合の表す意味は異なる状態を図示している．一方で平均プーリング（×印）で作られた文埋め込み同士はほとんど同じものになり，つまり意味の異なる二つの文にほとんど同じ表現が与えられてしまう．まとめると，**もともと点群として表現されている文を 1 点に要約することで，点群の空間的な広がりが持つ意味の異なりの情報を潰してしまう可能性がある**，ということだ[1)]．

本稿では，**(Q1)** 上記の仮説上の問題が実際のテキストとモデルで起きているのかどうか，**(Q2)** また起きているとしてそれはどの程度の割合なのかを，文間の意味類似度（STS）データセットを利用して経験的に検証する．実験の結果，**(A1)** 前述の問題が生じるような文ペアが確かに存在していること，つまり意味的に異なる文が平均プーリングで「同一視」されてしまうという問題が，実際のテキストとモデ

1)　単語埋め込みの集合を高次元空間の経験分布と見れば，平均プーリングでの要約は，1 次のモーメントのみを参照することを意味する．分散共分散以後（2 次のモーメント以後）の情報を無視することに注目すれば，統計的な意味でもその乱暴さは明らかであろう．

ルでも起きていること，**(A2)** 一方でその割合は小さいこと, がわかった．問題が生じる確率が低いことは，平均プーリングが経験的に有用であるという従来の知見を支持するものである．一方で，前述した問題は実際のテキストとモデルでも生じる明確なエラーであり，今後適切な措置を講じていく必要もあると考えられる．本稿が，文という自然言語の基本単位の表現を深層学習時代にどう構成すべきかという問題を再検討する契機となれば嬉しい．

2 平均プーリング

本節では文埋め込みを構成するための標準的な手法である平均プーリングを説明し，続いて平均プーリングにより生じ得る問題点について述べる．

2.1 平均プーリングを用いた文埋め込み

文埋め込みを構成するための最も標準的な手法は，文中の各単語の埋め込み表現を何らかのプーリング処理を介して 1 点に集約するというものである[2)]．プーリング手法には最大プーリングなどあるが，中でも**平均プーリング**が標準的であり，これまで理論的・経験的な有用性が示されてきた．[9, 10, 5, 11]

平均プーリングは，文字通り文中の各単語埋め込みの平均 (重心) をもって文埋め込みとする計算方法である． 形式的には，文 $s = (w_1, \cdots, w_i, \cdots w_n)$ の各単語 w_i に対応する単語埋め込みを $\boldsymbol{w}_i \in \mathbb{R}^d$ とし，文 s の埋め込み表現 $\boldsymbol{s}$ を以下で計算する．

$$\boldsymbol{s} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i \in \mathbb{R}^d \tag{1}$$

2.2 平均プーリングの問題

しかし，平均プーリングは，「はじめに」でも述べたように，全く異なる単語ベクトル集合を似た文埋め込みに表現し得るという問題がある．図 1 を改めて確認するが，このような例では**青色の点群**と**橙色の点群** (単語埋め込み集合) の配置は異なる一方，平均プーリングで作られた文埋め込み同士はほとんど同じものになっている．「実験」では，ここで述べた「点群として異なる」「埋め込みとして近い」をもう一歩形式的に扱い，定量的な検証を進める．

2) 文埋め込み自体を何らかのネットワークを介して直接出力させるモデルも過去にはあったが [8]，現状もっとも人気があり標準的なアプローチは，単語の埋め込み表現が得られる基盤モデルからまず単語の埋め込み表現を取り出し，これを何らかの方法で要約するというものだ．

異なる意味を持つ文が同じ点に埋め込まれてしまうと，実用上のさまざまなシーンで問題が生じる．たとえば文埋め込みを入力として何らかの予測問題を解く際には，予測モデルの学習ないし構成が不可能になる可能性がある．感情分析の問題を解く際に，“この映画は最高” と “こんな映画は二度と見たくない” が同じベクトル表現となる状況では，精度の高い分類器は構成できないだろう．文埋め込み同士を比較する際にも同様の問題が起きる．クエリベクトルと文書ベクトルを比較しながら関連文書を検索する際に，全く異なる内容を扱った文書が同じ埋め込み表現を持っていては困る．あるいは文埋め込み同士の比較を通じて文同士の意味類似度を推定する際も，異なる意味を持つ文が埋め込み空間で同一の点に埋め込まれている状況は避けたい．

3 分析の方針

以後では，「点群としては明らかに異なるのに，平均プーリングをすると同じ位置に埋め込まれてしまう」(図 1) という理論上の問題が，実際のテキストと実際のモデルで生じるのかどうかを経験的に検証したい．このためには，(i) 「点群同士がどの程度異なるか」と (ii)「これを要約した平均プーリング同士がどの程度異なるか」をそれぞれ定量的に評価する必要がある．本研究では，(i) として Word Mover's Distance (WMD) [12] と呼ばれる最適輸送に基づく点群距離を採用する．つまり，図 1 における「青色の点群と橙色の点群の配置の違い」を，「青色の点群の位置に置かれた荷物を橙色の点群の位置に距離空間で移し替えるために必要なコスト」で定量化する．(ii) としては，平均プーリング後の文埋め込み同士の L2 距離 を用いる．L2 距離は，単語埋め込み空間においてもっとも自然に考えられる距離構造であり，かつ，WMD が想定している距離構造とも一貫する．まとめると，「点群としては明らかに異なるのに，平均プーリングをすると同じ位置に埋め込まれてしまう」という現象を，「点群間の WMD は大きいが平均プーリング後の L2 距離は小さい」という条件で探していく．

4 実験

本節では，「単語埋め込み集合の配置が異なっているにもかかわらず平均プーリングで作られた文埋め込み同士がほとんど同じものになる」問題が実際のテキストやモデルで起きているのか，起きている

としてどの程度起こっているのか検証する．本稿では静的単語埋め込みを用いて，単語埋め込み集合としての距離より平均プーリングで作られた文埋め込みの距離が小さい文ペアを STS データセットの中から探し，上記の問題が生じているのか確認する．

4.1 実験設定

データセット データセットには STS Benchmark [13] の訓練データを使用した．このデータセットには全部で 5509 件の文ペアが存在し，各文ペアに対して人手で評価された文間の意味類似度が 0.0 ～ 5.0 の範囲で付けられている．

モデル 静的な単語埋め込みとして word2vec-google-news-300 [3] を使用した．

実験手順 まず，それぞれの文から文埋め込みを構成する．文を単語に分割して[4]，各単語に対応する単語埋め込みを取得し，平均プーリングで文埋め込みを構成する．次に，文埋め込みの違いと単語埋め込み集合の違いを比較する．各文ペアに対して文埋め込みの L2 距離と WMD を求め，データセット中の L2 距離と WMD それぞれで相対順位に基づき文ペアの意味類似度を 0.0〜1.0 の範囲で算出した[5]．そして，「文埋め込みの L2 距離に基づいた意味類似度」＞「WMD に基づいた意味類似度」となる文ペアに対して，単語埋め込み集合は異なっているが平均プーリングで作られた文埋め込み同士は近くなっているか確認する．主成分分析で二次元に次元削減して各埋め込みを可視化し，散布図から前述した問題が起こっているのか人手で確認する．

4.2 実験結果

4.2.1 定量分析

図 2 の赤線から右側が，「点群としては明らかに異なるのに，平均プーリングをすると近い位置に埋め込まれてしまう」例の割合を表している．x 軸は「平均プーリングで測った相対的な文類似度 – WMD で測った相対的な文類似度」を表しており，これが 0.3 を超えることは，5 段階（しかない）STS スコアが 3 ずれること，つまり平均プーリングと WMD で類似度の推定値に大きな乖離があることを意味する．

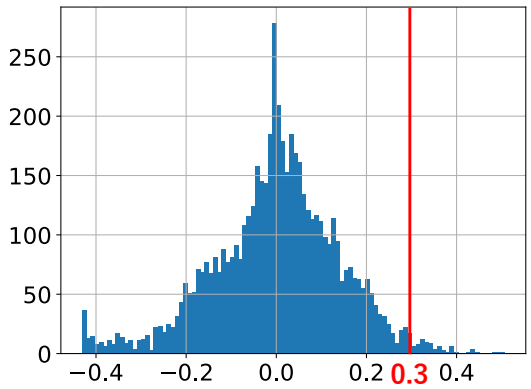

図 2 「平均プーリングの L2 距離の相対順位から算出した類似度」 – 「WMD の相対順位から算出した類似度」のヒストグラム．赤線よりも右にある文ペアに対して実際に人手で「単語埋め込み集合の配置は異なるのに平均プーリングで構成された単語埋め込みが近くなる」問題が起きているか判断する．横軸が大きくなるほど上記の問題が起こりやすくなると考えられる．

る．このような文ペアを機械的に抽出すると，データセットの中に 75 件存在し，これはデータセット全体の 1.3%に該当する．このうち，実際に「単語埋め込み集合は異なっているが平均プーリングで作られた文埋め込みが近くなっている」と人手で判断できた文ペアは 71 件存在していた．

割合の小ささ（1.3%）は，平均プーリングが経験的に有用であるという従来の知見を支持するものである．一方で前述した問題が実際のテキストとモデルで確かに生じていることも今回明らかになった．文表現の手法として平均プーリングは常に最良とは言えず，ケースに応じて適切な措置を講じる必要性を示唆している．

4.2.2 定性分析

「単語埋め込み集合の配置は異なっているが平均プーリングで作られた文埋め込みは近くなっている」と判断された 71 件の文ペアの中で，元々の意味類似度が小さい一例を図 3 に示した．この例ではデータセットに存在している二つの文は “a woman holds a **baby** while a **man looks** at it as another **man holding** a **child watches**.”, “a woman **stands** with her **arms** out in a **store** while another woman holds a **camera**.” となっている．この例では文埋め込みの L2 距離に基づいた意味類似度は 0.70， WMD に基づいた意味類似度は 0.37 であり，その差は 0.30 となっており，文埋め込みに基づく意味類似度が高く

3) https://github.com/piskvorky/gensim-data
4) WMD の設定に習い [12]，文を単語に分割する際にストップワードを除去した．
5) 意味類似度の値が大きいほど，文の意味的に近いように定義する．

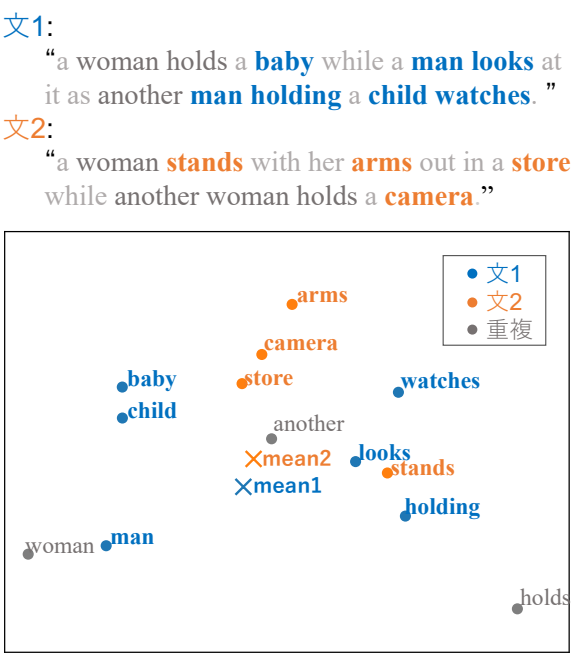

図 3 データセットに存在する単語埋め込みの点群の配置が異なるが平均プーリングで作られた文埋め込みはほとんど同じになる例．**青色の点**は “a woman holds a **baby** while a **man looks** at it as another **man holding** a **child watches**.”，**橙色の点**は “a woman **stands** with her **arms** out in a **store** while another woman holds a **camera**.” のそれぞれの色を割り当てた単語，灰色の点は文 1 と文 2 で重複している単語である．図の単語埋め込みの点群の配置は異なるものにもかかわらず，平均プーリングで作られた文埋め込み同士はほとんど同じものになっている．

なっていた．図を見てみると，単語埋め込み点群の配置は異なっているが平均プーリンで作られた文埋め込みは埋め込み空間で近くなっていることがわかる．またこの文ペアは人手で評価された意味類似度の相対順位は 0.09 であり，平均プーリングで作られた文埋め込みが文間の意味類似度を不当に高く計算しており，実用上の問題として本稿で指摘した問題点が起こっていることが確認できた．

5 関連研究

平均プーリングによる文表現 平均プーリングは文埋め込みを作る手法として人気のある手法で，静的単語埋め込みに基づく手法群 [5] から動的単語埋め込みに基づく手法群 [6, 7] まで，理論的・経験的な有用性が広く示されてきた [9, 10, 5, 11]．本稿では取り扱わなかったが，今後の展開として本稿で指摘した問題が動的単語埋め込みでも生じるのか検証したい．

点群の広がりを保持する文表現 単語埋め込み集合を 1 点に集約しない文表現も，これまで様々に提案されてきた．最適輸送に基づく文類似度尺度 [12, 14] は，点群をそのまま用い，距離空間上で点群を「動かす」コストを計算することで文の意味類似度を計算している．生成文の自動評価に幅広く用いられている BERTScore [15] もやはり単語埋め込み集合を集合のまま利用し，単語埋め込み間のコサイン類似度を集めた上で F 値に類する文間類似度を計算している．点群の持つ 2 次のモーメントまでを考慮する文表現として，文をガウス埋め込み [16] として表現する方法も提案されている [17]．本稿では，平均プーリングによって点群としての情報が大きく損なわれるケースが，実用上も少数だが確かに存在することを確認した．このことは，本節で述べた「広がりの情報を維持する表現」を使うべき状況が少数ではあるが確かに存在することを意味する．今後，上述したようなリッチな文表現方法を，いつ，どのように，いかなる修正を加えて利用することが肝要なのかを理解することは，分野の重要な将来研究であろう．

6 結論

本稿では，単語埋め込みの点群集合は異なるが平均プーリングで作った文埋め込みがほとんど同じものになる問題を指摘した．また，実際のテキストとモデルでは上記の問題が確かに生じているが，その割合は小さいことを実験で示した．この結果は平均プーリングの経験的な有用性を支持するが，文表現の構成方法を再検討する必要性を示唆するものである．

今後の展開として動的単語埋め込みや平均プーリング以外のプーリング手法，さらに L2 距離ではなくコサイン類似度で点の違いを定量化した場合についても分析を行うなどがある．また本稿では実際に単語埋め込みの配置は異なっているのに平均が近くなることを最終的には人手で判断したが，この人手による判断を何らかの量で記述することでより分析を精緻で一般的なものにしたい．さらに本稿では指摘した問題が起こる確率が低いという結果を理論的に解析することも考えられる．

謝辞

本研究は JSPS 科研費 JP22H05106，JP22H03654 の助成を受けたものです．また，本研究は JSPS 科研費 JP22H00524 の助成を受けたものです．さらに本研究は，JST，CREST，JPMJCR20D2 の支援を受けたものである．本研究の遂行にあたり多大なご助言，ご協力を賜りました TohokuNLP グループの皆様に感謝申し上げます．

参考文献

[1] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. 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. 6769–6781, Online, November 2020. Association for Computational Linguistics.

[2] 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.

[3] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[4] Harrison Chase. LangChain, October 2022.

[5] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. **CoRR**, Vol. abs/1511.08198, , 2015.

[6] 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.

[7] 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.

[8] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In **5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings**. OpenReview.net, 2017.

[9] Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. **Cogn. Sci.**, Vol. 34, No. 8, pp. 1388–1429, 2010.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, **Advances in Neural Information Processing Systems**, Vol. 26. Curran Associates, Inc., 2013.

[11] Carl Allen and Timothy Hospedales. Analogies explained: Towards understanding word embeddings. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, **Proceedings of the 36th International Conference on Machine Learning**, Vol. 97 of **Proceedings of Machine Learning Research**, pp. 223–231. PMLR, 09–15 Jun 2019.

[12] 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, 07–09 Jul 2015. PMLR.

[13] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors, **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)**, pp. 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[14] 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.

[15] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.

[16] Luke Vilnis and Andrew McCallum. Word Representations via Gaussian Embedding. In **3th International Conference on Learning Representations (ICLR)**, 2015.

[17] Shohei Yoda, Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. Sentence representations via gaussian embedding, 2023.