単語埋め込みの独立成分分析の軸が解釈できる粒度はどれくらいか？

佐藤祥太[1] 木山朔[2] 中島秀太[3] 小町守[3] 唐堂由其[1]
[1] 金沢大学 [2] 東京都立大学 [3] 一橋大学
sato11sho03@stu.kanazawa-u.ac.jp

概要

単語の意味や関連性を低次元のベクトルなどで表現することを単語埋め込みと呼ぶ．単語埋め込みの次元（軸）は，各次元の値が上位の単語から類推されるカテゴリを人手で判断することで解釈できる．本研究では，このカテゴリが解釈できる次元の割合を単語埋め込みの軸の解釈性と定義する．単語埋め込みの軸の解釈性は，独立成分分析による変換を行うと向上することが先行研究で示されている．しかし，その研究の中では単語埋め込みの次元数と同じ次元数で独立成分分析を実施しており，独立成分の粒度を変化させた場合にどの程度の次元が解釈できるかは明らかではない．そこで，本研究では独立成分分析の次元数を変えた際に解釈できる粒度がどの程度かを調査する．実験結果より，独立成分の粒度が大きいほど解釈性が低下することを示した．

1 はじめに

言語や画像，音声などをベクトルや行列，テンソルなどで低次元で表現する技術を表現学習と呼ぶ．この表現学習によって得られた数値列を埋め込み表現と呼ぶ．埋め込み表現の登場により，様々な自然言語処理のタスクを数値的な処理で扱えるようになった．本研究では単語の埋め込み表現（単語埋め込み）に着目して分析を行う．単語埋め込みはさまざまなタスクに有用だが，各次元が何を表現しているかを理解するのは困難である．

本研究では単語埋め込みの各次元の値が上位の単語からカテゴリを類推することを解釈と呼ぶ．単語埋め込みの**軸の解釈性**とは，単語埋め込みの中で解釈できる次元の割合と定義する．単語埋め込みの軸の解釈性を高めることは，埋め込みを用いた手法の信頼性を分析するのに有用である．

単語埋め込みの軸の解釈性を向上させるために，

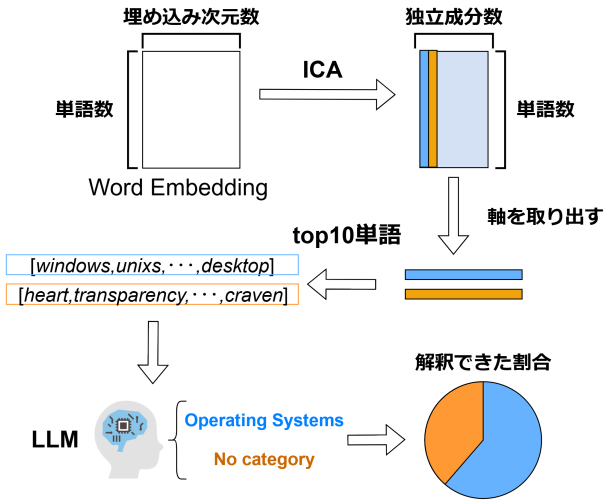

図 1 本研究の概要図．単語埋め込みを ICA で変換する際の独立成分数を変えて，軸の解釈性の評価を LLM を用いて実施する．

独立成分分析（Independent Component Analysis; ICA）を用いる研究が存在する [1, 2]．独立成分分析とは，**独立成分数**を指定し入力された行列が各列ごとに独立になるように線形変換する手法である．先行研究では ICA を用いた変換の際の独立成分数として単語埋め込みと同じ次元数を用いているものの，独立成分数を変えた場合の振る舞いは明らかではない．解釈性が最大となるような独立成分数が分かれば，単語埋め込みに ICA を適応する際の最適な粒度を用いることができる．

そこで，本研究では ICA で変換する独立成分数を変えた際に単語埋め込みの軸の解釈性がどのように変化するかを分析する．具体的には，ICA の独立成分数を変えた際に次元ごとに値が上位の単語から大規模言語モデル（LLM）によって軸が解釈できる独立成分数を調査する．さらに解釈できた/できなかった独立成分数の傾向についての分析をする．本研究の貢献は以下となる．

- LLM による軸の解釈の妥当性を人手評価に

表 1　LLM に入力した top-10 単語と生成されたカテゴリの例

top-10 単語	生成カテゴリ
microsoft, windows, unix, os, linux, desktop, macintosh, operating, kde, gui	Operating Systems
proteins, cell, cells, protein, membrane, cytoplasm, eukaryotic, dna, mrna, eukaryotes	Cell biology
heart, j, units, ec, cathar, transparency, televised, nist, priscilla, craven	No category

よって示した．特に，LLM が解釈できると判断した軸は人手評価とほぼ一致し，解釈できないと判断した軸の一部は人間によって解釈できる軸だと判定された．

- ICA を適応する際の独立成分の粒度が大きいほど軸の解釈性が低下することを示した．一方で，独立成分数が大きくなるほど軸の独立性が失われる傾向がわかった．また，解釈できる/できない軸の違いは単語の概念の類似度に関連することを示した．

2　関連研究

単語埋め込みに ICA を行い，埋め込み表現に共通する幾何構造を発見した研究について紹介する [1]．この研究では，単語埋め込みにおける次元圧縮手法である PCA（独立成分分析）と ICA を比較し，複数のモデルやモーダル，言語において ICA で変換した軸の解釈性が高いことを示した．ICA は入力された行列に対し，その成分が独立となるように線形で変換する手法である．単語埋め込みの行列を $X \in \mathbb{R}^{N \times D}$ とする．ここで N は単語数，D は単語埋め込みの次元数を示す．変換後の次元数（独立成分数）を d として，$N \times D$ 次元から $N \times d$ 次元に独立になるように線形変換する行列 $B \in \mathbb{R}^{D \times d}$ を求めることで，$S = XB$ となる独立成分行列 $S \in \mathbb{R}^{N \times d}$ が獲得できる．本研究では，独立成分行列 S における独立成分数 d を変え，独立成分の粒度を変化させた場合の単語埋め込みの軸の解釈性について調査する．最適な独立成分数が分かれば，ICA を用いて単語埋め込みやその類似度を分析する研究群 [3, 4, 5] のさらなる拡張が期待できる．

大規模言語モデルに評価をさせる研究群は LLM-as-a-judge と呼ばれる [6, 7]．LLM による評価は人手評価との相関が高くなるタスク [8, 9, 10, 11, 12] も存在しており，人手評価の代替として LLM による評価を用いることができる．本研究では，ICA の次元の値が上位の単語集合からカテゴリを生成させるタスク（軸の解釈タスク）を LLM に評価させる．LLM のカテゴリ出力を人手で確認し，LLM による評価が人間による評価と一致する結果を得た．

表 2　LLM カテゴリ生成の人手評価の混同行列．左がアノテータ 1，右がアノテータ 2．

LLM \ Annotator	解釈できない	解釈できる
解釈できない	(46,46)	(6,6)
解釈できる	(0,1)	(18,17)

3　軸の解釈性の評価方法

3.1　実験設定

実験では先行研究 [1] にならい，Skip-gram with Negative Sampling (SGNS) [13] で text8 コーパス[1] から学習した英語の 300 次元の単語埋め込みを使用する．低頻度語除去などの前処理を行い，トークン数 115,058，タイプ数 30,000 の埋め込みを得た．この埋め込みを変換前の埋め込みとし，scikit-learn (ver1.5.2) の FastICA を用いて埋め込みの変換を行う．[2] ICA によって変換する独立成分数は d' = [300, 200, 100, 80, 40, 20, 10] の計 7 つである．また，実験中で使用する LLM には Gemini 1.5 Flash を用いる．[3]

3.2　LLM による解釈

軸の解釈性を評価するため，ICA 変換後の埋め込みの各軸から歪度を算出し，必要に応じて符号を反転させ，すべての歪度を正にしたうえで，表 1 のように成分値の大きい上位 10 単語を取り出して LLM に与え，それらの単語群の意味から意味軸を表すカテゴリを生成できるかを確認する．LLM に与えるプロンプトには，単語の意味を考慮してカテゴリを生成するように明記し，カテゴリが作れない場合は “No category” と出力するように指示する．[4]

3.3　LLM による軸の解釈の妥当性

本研究で行う実験は，LLM による軸の解釈が人手と一致することが前提となる．よって，LLM による軸の解釈の妥当性を人手で評価する．

1）https://mattmahoney.net/dc/textdata.html
2）https://scikit-learn.org/stable/
3）https://deepmind.google/technologies/gemini/flash/
4）モデルに与えるプロンプトは付録図 4 に示す．

表 3 各独立成分数での解釈できた軸の割合

独立成分数	解釈できる割合
300	0.63
200	0.59
100	0.42
80	0.24
40	0.05
20	0.00
10	0.00

ICA によって次元を変換した軸の独立成分 d' からランダムに 10 軸を抽出し，top-10 単語と LLM が生成したカテゴリのペアを対象に，2 名のアノテータ（情報科学分野の大学院生と社会科学分野の大学院生）による計 70 件[5]の人手評価を行う．アノテーションでは，与えられた 10 単語のうち，5 単語以上がそのカテゴリにとって適しているかどうかで判定を行う．また，アノテータによる妥当性評価の際には，独立成分数がわからない状態にしたうえでアノテーションを行う．

表 2 に LLM カテゴリ生成の人手評価の結果を示す．アノテータ 1 によるアノテーションでのカッパ係数は 0.798 となり，アノテータ 2 によるアノテーションでのカッパ係数は 0.760 となった．これらの値は LLM とアノテータ間の一致度がかなりあることを示しており，LLM による軸の解釈を信頼できるものとして実験を進める．

4 分析 1: 解釈できる軸の調査

独立成分数を変えた際の軸の解釈性について評価する．ICA 変換後の各次元の埋め込みから，各軸の成分値の大きい top-10 の単語を LLM に与え，その単語群のカテゴリを生成させる．LLM の出力結果から，各次元での解釈できた軸の数を数え，解釈できた軸の割合の変化を分析する．

また，解釈できる軸と解釈できない軸の違いを分析するため，WordNet [14] を用いた分析を行う．解釈できる軸と解釈できない軸に対しそれぞれのカテゴリ生成に用いた top-10 単語の共通パスの長さを算出し，軸の解釈性と単語間の synset の類似度の関係性を分析する．加えて，埋め込みの各次元の歪度を算出し，歪度が正になるように必要に応じて符号を反転させたうえで，歪度の降順でソートした軸を用いて，歪度と軸の解釈の関係性を分析する．

5） ここで対象となるデータは，LLM によって解釈できると評価された軸 (17/70) と解釈できないと評価された軸 (53/70) の両方を含む．

表 4 top-10 単語の WordNet での共通パスの長さ

独立成分数	解釈できる軸	解釈できない軸
300	4.073	2.879
200	4.037	2.910
100	3.781	2.874
80	3.896	2.875
40	4.757	2.574
20	—	2.785
10	—	2.300

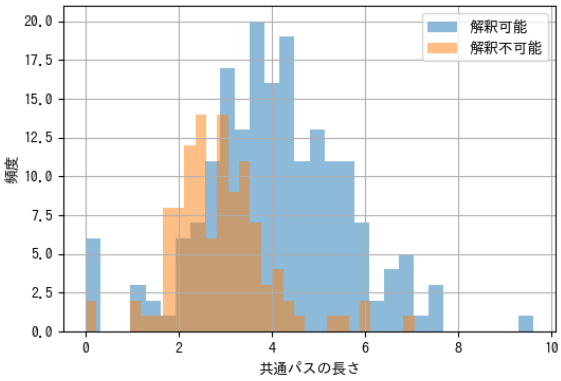

図 2 300 次元埋め込みでの各軸 top-10 単語の WordNet 共通パス長の分布

4.1 独立成分数ごとの軸の解釈性

表 3 に，各独立成分数に対し LLM が解釈できた軸の割合を示す．これらの結果から，ICA 変換後の埋め込みに関しては，独立成分数が小さいほど軸の解釈性が低下することが分かった．

4.2 解釈できる/できない軸の比較

軸の解釈性と top-10 単語の類似度の関係性 表 4 は，軸の解釈に用いた top-10 単語内の名詞から算出した WordNet での共通パスの長さを示す．[6]解釈できる軸と解釈できない軸の WordNet 上の共通パスの長さを比較すると，どの埋め込み次元においても 1 程度の共通パスの長さの差が確認できる．これより，解釈できる軸は解釈できない軸に比べて共通概念が多い単語が上位単語として出現する，つまり top-10 単語の概念の類似度が高いほど解釈しやすくなる傾向があるのがわかる．

図 2 は 300 次元の ICA 変換後の埋め込みにおいて，解釈できる場合と解釈できない場合のそれぞれで各軸の top-10 単語内の名詞の WordNet 上の共通パスの長さの分布を示したヒストグラムである．この

6） 共通パスの算出の際に参照する上位概念が品詞によって存在しないものがあり，簡単のため名詞に限定した．

表 5　ICA の軸上の top-n 個の単語集合のカバーする単語タイプ数 (top-10 および top-k)

		top-10			top-k		
独立成分数	k	解釈可能	解釈不可能	合計	解釈可能	解釈不可能	合計
300	100	1,837 / 1,900	1,066 / 1,100	2,832 / 3,000	14,785 / 19,000	9,215 / 11,000	20,044 / 30,000
200	150	1,148 / 1,170	815 / 830	1,940 / 2,000	13,978 / 17,550	10,393 / 12,450	20,130 / 30,000
100	300	417 / 420	575 / 580	988 / 1,000	10,787 / 12,600	13,870 / 17,400	20,399 / 30,000
80	375	190 / 190	610 / 610	799 / 800	6,589 / 7,125	16,995 / 22,875	20,492 / 30,000
40	750	20 / 20	379 / 380	399 / 400	1,480 / 1,500	20,266 / 28,500	20,934 / 30,000
20	1,500	—	200 / 200	200 / 200	—	21,345 / 30,000	21,345 / 30,000
10	3,000	—	100 / 100	100 / 100	—	21,528 / 30,000	21,528 / 30,000

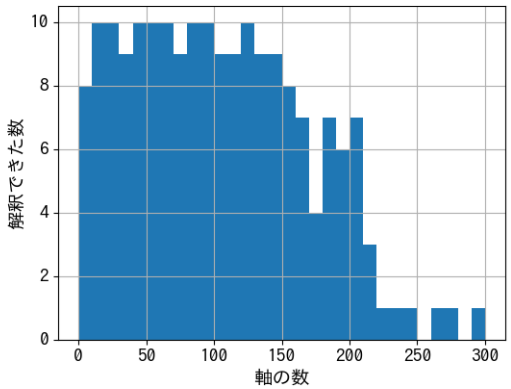

図 3　300 次元の単語埋め込みを歪度でソートした場合での 10 軸ごとの解釈できる軸の数

図より，解釈できる軸の中には多くの共通概念を持つ単語によって解釈される軸が存在していることが分かる．また，解釈できる軸は解釈できない軸に比べて，共通パスの長さが長い軸が多いことも確認できる．

軸の解釈性と歪度の関係性　図 3 は，ICA 変換後の 300 次元の埋め込みに対して正の歪度の降順にソートをし，10 軸ごとに解釈できる軸を集計したヒストグラムである．[7] この図より，軸の解釈性については，正の歪度の値が大きいほど解釈しやすく，値が小さくなるにつれて軸の解釈性も減少するという傾向が確認できる．

5　分析 2: 軸の表現能力の調査

ICA による単語埋め込みの変換によって表現される軸上に，どの程度の語彙が網羅されているかを確認し，解釈できた軸集合の表現力を調べる．各次元 d' の埋め込みの解釈できる軸と解釈できない軸のそれぞれで，1) top-10 単語，2) top-k 単語 ($k = 30,000/d'$) のタイプ数を数える．[8]

top-10 単語の場合　表 5 の左側は，各次元の埋め込みに対して，解釈できる軸と解釈できない軸のそれぞれで top-10 単語のタイプ数を数えた結果である．top-10 の単語に関しては，ほぼ重複なく単語が分散していることが確認できる．

top-k 単語の場合　表 5 の右側は，各次元の埋め込みに対して，解釈できる軸と解釈できない軸のそれぞれで top-k 単語のタイプ数を数えた結果である．top-k の結果を見ると，$\frac{2}{3}$ 程度の語彙しか出現しておらず，語彙をカバーしきれていないことが確認できる．解釈できる軸の中でも出現する語彙数が少ないだけでなく，解釈可能・不可能の軸で分けた場合と合計した場合で出現する語彙数が減少していることから，2 つの軸の間にも語彙の重複があることが分かる．この傾向は独立成分数が大きいときに顕著であり，独立成分の粒度を細かくすると，必ずしも独立性の高くない軸が抽出されていると考えられる．

6　おわりに

本研究では，LLM による解釈の妥当性を人手評価で示し，独立成分数を変えた場合の単語埋め込みの軸の解釈性を定量的に評価する実験を行った．解釈できる粒度の分析の結果，独立成分数が小さくなるほど軸の解釈性が低くなる一方，独立成分数が大きくなるほど軸の独立性が失われる傾向がわかった．また，解釈できる軸と解釈できない軸の違いとして単語の概念の類似度に関連することを示した．

今後の展望としては，BERT などの文脈化埋め込みに対する分析と多言語での分析を行い，本分析の妥当性を検証していきたい．また，軸の解釈において上位 10 単語を用いて解釈を行なっているが，この単語数が最適かどうかを調査していきたい．

7)　先行研究 [1] では可視化した図をわかりやすくするために歪度の絶対値によってソートしているため，本研究でも先行研究に従った．

8)　top 単語数の選定理由として，軸のカテゴリ生成に用いた単語数である top-10，理想的には $k = 30,000/d'$ で元の埋め込みの全語彙が網羅できるとして top-k 単語を対象とした．

参考文献

[1] Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. Discovering universal geometry in embeddings with ICA. In **EMNLP 2023**, pp. 4647–4675, Singapore, December 2023. Association for Computational Linguistics.

[2] Tomáš Musil and David Mareček. Exploring interpretability of independent components of word embeddings with automated word intruder test. In **LREC-COLING 2024**, pp. 6922–6928, Torino, Italia, May 2024. ELRA and ICCL.

[3] Rongzhi Li, Takeru Matsuda, and Hitomi Yanaka. Exploring intra and inter-language consistency in embeddings with ICA. In **EMNLP 2024**, pp. 19104–19111, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[4] Momose Oyama, Hiroaki Yamagiwa, and Hidetoshi Shimodaira. Understanding higher-order correlations among semantic components in embeddings. In **EMNLP 2024**, pp. 2883–2899, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[5] Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. Revisiting cosine similarity via normalized ICA-transformed embeddings. In **COLING 2025**. International Committee on Computational Linguistics, 2025.

[6] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on LLM-as-a-judge. In **arXiv**, 2024.

[7] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods. In **arXiv**, 2024.

[8] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In **NeurIPS 2023 Track Datasets and Benchmarks**, 2023.

[9] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In **ACL 2023**, pp. 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics.

[10] Masamune Kobayashi, Masato Mita, and Mamoru Komachi. Large language models are state-of-the-art evaluator for grammatical error correction. In **BEA 2024**, pp. 68–77, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[11] Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. TMU-HIT at MLSP 2024: How well can GPT-4 tackle multilingual lexical simplification? In **BEA 2024**, pp. 590–598, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[12] Ayako Sato, Kyotaro Nakajima, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. TMU-HIT‘s submission for the WMT24 quality estimation shared task: Is GPT-4 a good evaluator for machine translation? In **WMT 2024**, pp. 529–534, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In **NeurIPS 2013**, Vol. 26. Curran Associates, Inc., 2013.

[14] George A. Miller. WordNet: a lexical database for English. **Commun. ACM**, Vol. 38, No. 11, p. 39–41, November 1995.

A　実験に用いたプロンプト

From the meanings of the next 10 words, please generate a category with one word for the word group. However, if the category cannot be identified from the meanings of the 10 words, please output "No category".
[*'word-1'*, *'word-2'*, ... , *'word-10'*]

図 4　LLM に与えたカテゴリ生成のプロンプト

B　LLM とアノテーター間での軸解釈の不一致に関わる分析

LLM では No category となったがアノテータにはカテゴリが類推できたものには，人種や宗教に関する単語が与えられた場合や，記号を表すような単語や機関の省略名が与えられた場合，10 単語すべてを考慮に入れるとカテゴリが作りづらい場合などがみられた．具体例は表 6 に示す通りである．このことから，軸が解釈できない要因には LLM の知識や安全性に関する制約が起因している可能性がある．また，表 2 では LLM とアノテータ間の不一致のほとんどが LLM が解釈できずアノテータが解釈でき，LLM が解釈できてアノテータが解釈できないものが概ねないことから，LLM は人間よりも軸の解釈に関して厳しい評価をする傾向が確認できる．

表 6　LLM と人手で評価が異なったカテゴリ生成の例

top-10 単語	LLM	人手評価
black, white, racial, panther, blacks, races, klansmen, sabbath, morbid, holes	No category	人種
x, y, n, rangle, langle, p, qquad, cdots, frac, phi	No category	数学記号
field, magnetic, fields, interest, electric, clone, carotene, generated, musicology, subfields	No category	サイエンス

C　WordNet 上での共通パスの長さと単語の具体例

表 1 に示した単語群の WordNet 上での共通パス長を表 7 に示す．WordNet 上に名詞として存在する単語は下線によって示されている．解釈できた単語群では，実際に同じ分野の単語が存在していることが確認できる．また，解釈できなかった単語群に着目すると，単語間の意味的類似度が低いことが確認できる．この表 7 から，例えば “j” のような単語が名詞されていることが分かる．“j” の場合は，WordNet 上で熱力学の単位である “joule” と紐づいて，名詞として認識された．また，“os” という単語は，“子宮口” の意味合いの名詞として認識された．このように，本研究における WordNet を用いた分析では本来の意図とは異なった単語の処理が一部含まれている．

表 7　LLM に入力した top-10 単語と top-10 単語の名詞の WordNet 上の共通パスの長さの例

top-10 単語	共通パスの長さ
microsoft, windows, unix, os, linux, desktop, macintosh, operating, kde, gui	6.333
proteins, cell, cells, protein, membrane, cytoplasm, eukaryotic, dna, mrna, eukaryotes	3.139
heart, j, units, ec, cathar, transparency, televised, nist, priscilla, craven	2.389