逆埋め込み難易度制御のための整列度合い計算法と難易度段階間の観点差分析法

江原 遥
Yo EHARA
東京学芸大学　教育学部
Email: ehara@u-gakugei.ac.jp

あらまし：テキストの意味を数値的に表現する埋め込みベクトル空間は生成 AI の基盤技術である．難易度は意味の重要な側面だが，空間での表現との対応を簡便に求めたい．本研究では，埋め込み座標からテキストを復元する逆埋め込み技術を活用し，埋め込み座標から目的の難易度のテキストを生成する直感的な手法を提案する．埋め込み手法に依らず空間と難易度の対応の良さを測れる整列度合いを理論的に提案し，語学教育のための難易度段階間の観点の差を数値的に評価する分析法も提案する．
キーワード：逆埋め込み，生成 AI，難易度制御

1　はじめに

意味的に近接するテキストに対して，ベクトル空間上で近接する座標を与える「埋め込み」技術は，現代の自然言語処理における基盤技術となった [4]．一方，埋め込みベクトルからテキストを復元する「逆埋め込み」技術は新しい技術である [3]．逆埋め込みは，埋め込みから個人名等を復元できるかという情報セキュリティ上強い動機があるため，技術開発への投資が集まりやすく復元技術が精緻化していくと考えられる．

教育 AI 分野での難易度制御については，文献 [5] の研究が挙げられる．しかし，この研究では設問に対して回答した試験データが難易度推定に使われていたり，語学教育とは直接関係しない AI の質問応答性能評価のデータセット SQuAD で評価されている．この他にもテキスト生成の難易度制御手法はあるが，**基本的にはどの手法も難易度付与された一定規模以上のデータセットによるモデルの訓練を要する**．語学教育では語学教師等が人手で難易度を評価したデータを基準にしたい場面が多く，こうしたデータは作成コストが高いため少量のデータでも動作する手法が望ましい．さらに，段階間での難易度評価の「観点」の分析も行いたい．

本研究では，こうした語学教育難易度のための難易度制御のニーズを満たす難易度制御手法として逆埋め込みを用いた手法を提案する (図 1)．まず，容易な文例をさらに難しい方向に延ばすことで，与えられた 2 つの段階にはない新しい難しい文例を生成する「外挿」の応用が考えられる．図 1 では，外挿を使用してより難しい文例の座標を算出し，その座標を入力として逆埋め込みを適用することで座標に対応するより難しい文例を生成する．また内挿を使用して中間点の座標を算出し，中間の難易度を持つ文例を生成する応用も考えられる．

図 1 のように埋め込みの線形補完で難易度制御する場合，次のリサーチクエスチョン (RQ) が考えられる：*RQ1:* **図 *1* のように都合よくテキストが埋め込み空間上に並んでいるか？**，*RQ2:* **段階ごとにテキストを難しくする方向（観点）は異なるのではないか？**RQ1 のため複数の埋め込み手法を比較し，難易度制御に適した空間構造を持つ埋め込みを選択するための**整列度合い**という指標を提案する．また RQ2 のために，アノテータが外国語の文の難易度段階を評価する「観点」を分析する手法を提案し難易度段階ごとに観点が異なる事を示す．

2　整列度合い

2 段階の間で図 1 のように文埋め込みベクトルが適切に並んでいる「整列度合い」を簡単に数値化したい．2 つの段階の文数を N_i, N_j とする．簡単のため段階 i の埋め込みベクトル集合を $\{\mathbf{x}_1, \ldots, \mathbf{x}_{N_i}\}$ とし，その 1 つを $\mathbf{x}_i$ と書く．段階 j も同様とする．全ての $\mathbf{x}$ は D 次元で，ユークリッドノルムで $||\mathbf{x}|| = 1$ に正規化されているとする．この空間上で容易な段階 i から難しい段階 j の順に文例の埋め込みが整列する方向 $\mathbf{w}$ を探す問題を考える (図 1)．この整列の制約は $\mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j) < 0$ と書ける．2 段階間の全ての点の組に対してはこの制約は成り立ちづらいので，余裕を表すスラック変数 ξ を導入し全余裕を最大化する最適化問題を考える．

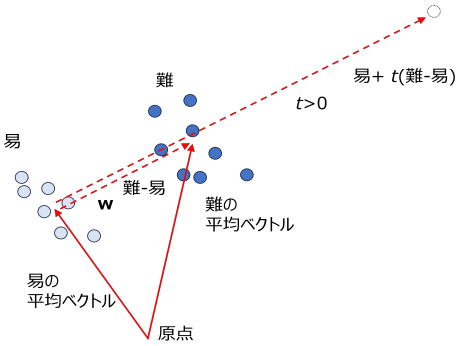

図 1: 概念図．各○は文．2 段階の文集合の平均ベクトルを用いて指定の難易度の文の座標（点線白○）を求め，逆埋め込みを用いて対応する文を生成することで，数値指定した難易度の文を直感的に作る難易度制御を行いたい．本図は t を制御しより難しい文を作る外挿例．

$$\begin{aligned} \text{maximize}_{\mathbf{w},\xi} \quad & \sum_{k=1}^{K} \xi_k \\ \text{s.t.} \quad & \forall k \in \{1, \ldots, N_i N_j\}; \\ \mathbf{w}^\top(\mathbf{x}_{i_k} \quad - \quad & \mathbf{x}_{j_k}) + \xi_k = 0, ||\mathbf{w}||^2 = 1 \end{aligned} \tag{1}$$

ここで k は 2 つの段階の文の全組み合わせ $K = N_i N_j$ の k 番目を表す．この最適化問題の解は，$\mathbf{w} \propto -\sum_{k=1}^{K}(\mathbf{x}_{i_k} - \mathbf{x}_{j_k})$ である．段階 i の平均ベクトルを $\bar{\mathbf{x}}_i$ とし段階 j も同様とすると $\mathbf{w} \propto -(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$ と表せる．そこで，$\hat{\mathbf{w}} = -\frac{\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j}{||\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j||}$ とすると，式 1 の最適化問題の目的関数値は $-N_i N_j \hat{\mathbf{w}}^\top(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j) = N_i N_j ||\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j||$ と表せる．N_i と N_j に依存しないよう，段階 i と段階 j の**整列度合い**を平均ベクトルの差のノルム $||\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j||$ で定義する．この**整列度合い**は閉じた式で簡単に計算でき，大きいほど，「2 つの段階の埋め込み点が図 1 のようにそろって見える最も良い方向 $\hat{\mathbf{w}}$ を探したとき，制約を満たさない度合いが小さい」という理論的な意味付けがある．**この整列度合い指標は，ノルムが *1* である事以外は埋め込み空間での分布の仮定を課していないため，任意の埋め込み空間に対して汎用的に適用可能である．**

3　実験

3.1　データセット

第二言語学習者向けに英語文を CEFR(Common European Framework of Reference) の段階で難易度付けしたデータセット *CEFR-SP* を用いた [1]．CEFR-SP は Wiki-Auto と SCoRE の 2 つに分かれるが，本研究では SCoRE のみを使用した．CEFR-SP は 2 名の専門家によって難易度付けされている．[1] とは異なり両者の判断が一致した文のみを用い，簡単な段階から難しい段階に次の 4 段階が残った：A1，A2，B1，B2.

表 1: 各埋め込みの CEFR の各段階間の整列度合い

埋め込みモデル名	次元数	A1,A2	A1,B1	A1,B2	A2,B1	A2,B2	B1,B2
text-embedding-ada-002	1,536	0.19	0.23	0.31	0.09	0.24	0.23
bge-m3	1,024	0.29	0.34	0.47	0.14	0.37	0.36
multilingual-e5-large	1,024	0.20	0.24	0.32	0.10	0.25	0.23
all-MiniLM-L6-v2	384	0.37	0.41	0.60	0.17	0.43	0.41
multilingual-e5-small	384	0.19	0.23	0.31	0.10	0.24	0.23

表 2: 難易度を表す方向ベクトル間のコサイン類似度（大きいほうが類似）

-	A2-A1	B1-A1	B1-A2	B2-A1	B2-A2	B2-B1
A2-A1	1.000	0.920	0.198	0.648	0.046	-0.031
B1-A1	0.920	1.000	0.566	0.685	0.157	-0.061
B1-A2	0.198	0.566	1.000	0.349	0.298	-0.088
B2-A1	0.648	0.685	0.349	1.000	0.790	0.685
B2-A2	0.046	0.157	0.298	0.790	1.000	0.925
B2-B1	-0.031	-0.061	-0.088	0.685	0.925	1.000

表 3: 差分 B2-A1 での難易度制御 (数値評価は本文記載)

v	She had a beautiful necklace around her neck.
v+0.4(B2-A1)	She had a beautiful necklace around her neck around her neck.
v+1.0(B2-A1)	She had a beautiful necklace around her neck adorned with a beautiful pearl necklace..
v+1.1(B2-A1)	Her necklace was very beautiful around her neck adorned with a small pearl necklace.
v+1.7(B2-A1)	Her necklace was admired greatly by a young female nurse from a very fine necropolis etch..
v+1.8(B2-A1)	Her necklace was admired greatly by a young female narwhal from the famous jewelry store.
v+2.0(B2-A1)	Her narwhal necklaces were greatly admired by the female researcher from a small history center.
v+5.0(B2-A1)	The history of forensics graduates from acorns can be greatly delved into by a modern forensic scientist.

逆埋め込み　本研究全体を通じて，[3] が提供するコード（`https://github.com/jxmorris12/vec2text`）と，付随する 2025 年 5 月現在唯一の公式提供の逆埋め込みモデルを使用する．これは OpenAI の 1,536 次元の埋め込みである *text-embedding-ada-002* に対する逆埋め込みモデルである．まず逆埋め込みが埋め込みベクトルの座標から元の文をどの程度復元できるかを簡単に検証したところ，比較的簡単なテキストでは元の文とほぼ完全に一致する復元が得られた．例えば “She had a beautiful necklace around her neck.” では，ピリオドの位置の違いを除き完全に復元された．

RQ1: 整列度合い　RQ1 への回答として，表 1 に節 2 で定義した整列度合いを示す．様々な意味処理の性能比較サイト MTEB(`https://huggingface.co/spaces/mteb/leaderboard`) で高性能な埋め込みから表 1 中の 2 行目以降の 4 種の埋め込みモデルを比較した．どの埋め込みでも最もレベル差のある A1 と B2 が最も高い整列度合いを示している事，埋め込みの種類によって整列度合いが大きく違う事，埋め込みの次元数と整列度合いには直接の関係はない事が分かる．ここで最も整列度合いの高い *all-MiniLM-L6-v2* に対して逆埋め込みモデルが提供されれば，逆埋め込みを用いた難易度制御実験を行う事が考えられる．しかし，今回は適切な比較のため，唯一逆埋め込みが公式提供されている *text-embedding-ada-002* を用いた．表 1 から，近年高性能を達成している *multilingual-e5-small* と同等の性能が期待される．

RQ2: 観点　図 1 のように，難しいレベルの平均埋め込みベクトルから簡単なレベルの平均埋め込みベクトルを引いて得られる差分ベクトルは，意味空間上で簡単な文を難しくなる方向，すなわち文の難易度を評価する「観点」を表していると解釈できる．さて，差分ベクトルを用いて難易度制御を行う事を考えよう．今回は 4 つのレベル A1,A2,B1,B2 を考えているので，差分ベクトルだけでも 6 通りある．観点の違いを測る事が RQ2 であった．観点は方向ベクトルであるので，観点の近さはコサイン類似度を用いて計算できる（表 2）．

表 2 を見ると，各段階間の差の観点はかなり多様である事が分かる．例えば，簡単な方の 2 段階 A2 と A1 を分ける観点は，難しい方の 2 段階 B2 と B1 を分ける観点とは類似していない．しかし，最も難しい B2 と最も簡単な A1 を分ける観点 B2-A1 はどの段階間の観点とも全体的に類似している．従って，B2-A1 をテキストを難しくする方向と考えることは妥当と考えられる．

実際に，原文の埋め込みベクトル v を B2-A1 の差分ベクトル（観点）を足していく事で難易度制御し，逆埋め込みを用いて生成した結果を表 3 に示す．差分ベクトルの係数が増加するにつれて，文が段階的により難しくなることが定性的にも分かる．興味深いことに，係数が約 2.0 まででは元の文の “necklace” という単語が残るが，係数がさらに増加すると，生成される文は元のテキストとは無関係な難しい単語を使用するようになった．例えば，係数 1.8 以上では “narwhal”（イッカク）という単語が現れ，これは TOEIC950 点以上レベルの高度な語彙である．

逆埋め込みによる生成された文の難易度を定量的にも評価した．各埋め込み座標に対して，最もユークリッド距離が近い 10 例に対する人手の評価値の平均値を計算した．A1，A2，B1，B2 をそれぞれ 1.0，2.0，3.0，4.0 とした．v+0.5(B2-A1): 1.9，v+1.0(B2-A1): 2.3，v+1.5(B2-A1): 3.0，v+2.8(B2-A1): 3.4 となり，難しい方向に移動させるほど，10 近傍の人手評価の難易度が増加することが確認された．

科学教育における MMLU 実験　提案法は一般の埋め込み空間を用いるため**内容面での難易度制御への応用も見込める**利点がある．高校水準で 150 の高校物理問題のデータを用いて [2]，100 の訓練データと 50 のテストデータに分割し，multilingual-e5-small 埋め込みを使用して実験した．結果は，v+0.5d: 1.4，v+0.6d: 1.6，v+0.7d: 1.7，v+0.8d: 1.9 となり（v は物理基礎の平均ベクトル，d は物理基礎と物理の差分ベクトル，値は 10 近傍の平均値），物理学的な内容も難易度制御が可能と示唆された．

4　おわりに

本研究では，上記の語学教育にとって望ましい性質を全て満たす有望な手法として，逆埋め込みを用いる手法を提案した．本研究では語学教育に用いる場合にどの埋め込みを使えばいいのかを選択する「整列度合い」の理論的な導出を行い，逆埋め込みによる難易度制御が可能なことを定性的・定量的に示し，逆埋め込みによる手法では難易度段階間の「観点」の分析も行える事も示した．提案手法は大規模データを要さないため，今後の研究として，本稿で例示した科学分野のほか，特定層の学習者に適応的に難易度制御など幅広い応用が期待される．さらに提案法は特定の埋め込み手法に依存しない汎用的な手法であるため，新しい逆埋め込み技術の登場時にも適用可能であることや，文脈内学習中や微調整中の動的な埋め込み空間に対しても有効であることが期待される．

謝辞　本研究は JSPS 科研費 22K12287 および JST PRESTO JPMJPR2363 の支援を受けた．

参考文献

(1) Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. CEFR-based sentence difficulty annotation and assessment. In *Proc. of EMNLP*, pp. 6206–6219, 2022.
(2) Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
(3) John X. Morris, Wenting Zhao, Justin T. Chiu, Vitaly Shmatikov, and Alexander M. Rush. Language model inversion. In *Proc. of ICLR*, 2024.
(4) N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
(5) 後藤照佳, 富川雄斗, 宇都雅輝. 問題と模範解答を同時に生成する難易度調整機能付き読解問題自動生成手法. 第 49 回教育システム情報学会全国大会 Web 講演論文集, 2024.