N-015

大規模言語モデルの埋め込み順序回帰による解釈しやすい難易度推定

江原 遥[1)]

1　はじめに

AI に教材の難易度を推定させる事は，難易度の基準と観点の解釈・理解が必要であり適切な教材生成のためにも重要となる，教育情報システムの基礎的なタスクである．既存研究では教育 AI の難易度推定の目的でも自然言語処理の手法をそのまま適用する研究が多数であり，「少なくとも簡単な問題ではない」といった人間が直感的に行っている知見をうまく活用できていない．

本研究では，BERT(Bidirectional Encoder Representations from Transformers) [1] や大規模言語モデルの In-Context Learning（文脈内学習）などの現代的な手法を用いた難易度推定を，埋め込み空間上の順序回帰問題として整理し定式化する枠組みを提案する．この定式化の方法は複数あり定式化ごとに異なるモデルができるため，既存のものも新規なものも含まれる．この定式化により，様々な難易度尺度のデータセットの難しさを AI がどのように捉えているのか埋め込み空間上で幾何学的・直感的に解釈する事が可能になる．実験では，実際の難易度データセット上で網羅的に性能を比較する．さらに学習履歴データと組み合わせた個別学習支援への応用も論じる．

図 1 に本研究の貢献の概要を示す．第 1 の貢献はテキストからの難易度推定の近年のニューラルネットワーク手法をサーベイし数理的にまとめ，その構造が図 1 のように分解できることを示した事である．ここでのテキストからの難易度推定は，語学学習だけでなく一般の科学等の質問も含む．前者は，あらかじめ大規模なコーパスで事前学習を行った Transformer に基づくニューラルネットワーク (NN) である BERT[1] 等の手法である．この部分は，意味の高次元ベクトル空間上の数値表現である**埋め込みベクトル**を出力する．直感的には，埋め込みベクトルは意味的に近いテキストが高次元空間上で近くなるように作られる．この BERT の埋め込みベクトルをそのまま用いる場合もあるが，多くの手法は難易度推定を行うニューラルネットワークに接続し，モデル全体を訓練データを用いて追加で訓練（ファインチューニング，微調整）する．訓練データとして，テキストに正解となる難易度を人手等で付与したデータをあらかじめ用意する．難易度推定は ChatGPT 等の対話型生成 AI の出現後である 2024 年の時点でも，ファインチューニングを行う手法が高性能で優勢である [2]．これは，難易度が一次元的な尺度であり，基準を正確に掴む多くの訓練事例に接する事が必要であるためであると思われる．

さて，この「難易度推定を行うニューラルネットワーク」には多くの種類が提案されており多くの既存研究はその手前の「BERT 等」の部分を含む手法を提案手法として提案されていた．一方，近年では，この「BERT 等」の部分にも技術革新があり，さらに精密に意味を捉えられるように発展してきている（例：ModernBERT[3]）．このため，**難易度推定値だけを解釈しても，「*BERT* 等」の部分と「難易度推定を行うニューラルネットワーク」の相性がたまたまよかったのか，それとも「*BERT* 等」の部分が出力する埋め込みベクトルそのものの性質が良いのかが分からない**など，解釈性に問題があった（図 1 下部）．機械学習モデルの解釈手法としては LIME[4] や SHAP[5] 等が有名であるが，いずれも**高次元かつ各次元が特定の特徴量に対応しているわけではない埋め込みベクトル空間には適用し難い**．そこで，解釈しやすい指標を導入することで部分ごとの性能を計測できるようにし性能改善につなげることが本研究の主な動機である．

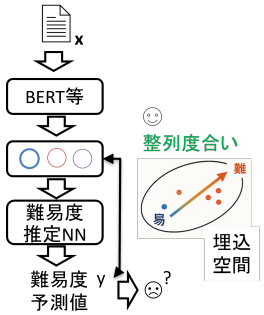

図 1　貢献の概要．NN はニューラルネットワーク．

この目的のため，2 点目の大きな貢献として，本研究では「BERT 等」部分の埋め込みベクトルが目的とする難易度のデータセットとどれだけ沿っているかを表す指標である「整列度合い」を提案し，解釈を容易にする（図 1 上部）．この指標は埋め込みベクトルが埋め込みベクトル空間上で幾何的にどれだけうまく整列しているかを表す指標である．**「*BERT* 等」部分が本当に意味を上手くとらえて，それを峻別しやすいような高次元ベクトルを出力しているのであれば，線形性など簡単な仮定に基づく「整列度合い」指標でも難易度を上手く識別できるはずである，という考え方が根底にある**．実際に，微調整時には埋め込みベクトルの空間がその先の「難易度推定を行うニューラルネットワーク」以上に更新されることはよくある．例えば，全体的な予測性能は悪かった場合でも「整列度合い」は高い場合は，「BERT 等」ではなく，その先の「難易度推定を行うニューラルネットワーク」部分の選択やパラメタ設定に問題がある可能性が考えられる．また，「整列度合い」は**次元を固定すればモデルに非依存な量である**ため，全く異なる構造の埋め込み間で比較を行う事ができる．さらに，「BERT 等」とは原理は異なるもののいわゆる対話型生成 AI も回答時には埋め込みベクトルを通じて回答を生成するため，モデル非依存性を利用して対話型生成 AI に対しても「整列度合い」を算出し所与の難易度データセットの回答に対話型生成 AI が適しているかの指標になる．大規模言語モデルの解析を行うプロービング手法の大半はモデルに依存した手法であり，順序回帰問題に特化しているもののモデル非依存な「整列度合い」は新規性・有用性が高いと期待される．

2　関連研究・背景

難易度推定問題は BERT[1] の登場以降，従来の人手の特徴量をベースにした手法を大幅に上回る性能を示し，事前学習済モデルを用いるアプローチが主流になった [6]．難易度推定問題は多クラス分類を用いて解かれるこ

1）　東京学芸大学

とも多い．多クラス分類損失では，クラス間の順序を考えないため例えば簡単なレベルではない時に「難しい方のレベルのうちどれかは分からないが，このレベルよりは難しい」といった情報が活用できない．この情報を活用しクラス間の順序を考慮して，例えば簡単なレベルではない事が分かった時に「難しい方のレベルのうちのどれか」の確率を上げるようにするのが順序回帰損失である．ただし，もともとのデータセットに段階が 2 つしかなければ，簡単な方のレベルではないときの難しい方のレベルは 1 種類であるために，順序回帰による予測性能向上は期待できない．

順序回帰自体は BERT より前から存在するが，ニューラルネットワークの最終層に順序回帰のための層を導入して分類する手法が提案されてきた．この手法には大別して 2 通りある．1 つは，多クラス分類をそのまま拡張し，損失関数の中でクラスの順序を考慮する手法が該当する．例えば，画像分野では**正解ラベルとの距離に応じて確率的なソフトラベルを生成し，クロスエントロピーで学習することで誤りの程度を滑らかに評価する** 順序回帰損失である Soft Ordinal Regression （SORD）がこれに該当する [7]．SORD は順序回帰損失として多用され，可読性推定にも導入されている [8] .

ニューラルネットを順序回帰に対応させる，もう 1 つの代表的な手法はペアワイズ判別である．これは，2 つの事例（事例のペア）が与えられたときに，どちらの方が順序が先かを判定する 2 値判別器を構成する手法である．結局行っていることは 2 値判別であるので適用しやすく，また，事例のペアを入力とするため，訓練データを増やしやすい利点もある．しかし，手法よりもレベル分けされた訓練データをどのようにペアのデータに変換するのかや，予測時にも与えられた事例が「他の事例と比較して」難しいか易しいかの判別しか行われないため，レベルの形式に変換するために，この比較対象の事例をどのように選ぶかによって性能が変わってくる問題がある．教育における難易度推定では，可読性判定で BERT にペアワイズ判別を組み合わせた研究がある [9].

2.1 多クラス分類（クロスエントロピー）

ここから，前節までで説明した「埋め込みベクトル」と「難易度推定を行うニューラルネットワーク」を具体的に数式で詳述する．まず，BERT 等の最終層が出力する埋め込みを $\mathbf{x}_i$ とし，その難易度ラベルを $\mathbf{y}_i$ とする．難易度ラベルは K 段階あるとし k は各段階を走る添え字とする．また，データは N 件あるとし i はデータを走る添え字とする．順序を考慮しない多クラス分類では，$p_{i,k} = \mathrm{softmax}\big(\mathbf{W}\mathbf{x}_i + \mathbf{b}\big)_k$ と定義する．ここで $\mathbf{W}$ と $\mathbf{b}$ は学習可能パラメータである．真のラベル y_i を one–hot ベクトル $\boldsymbol{y}_i$ で表すと，クロスエントロピー損失は次のようになる．ここで，単純にクラスに該当するかだけが考慮されており，クラスの順序については考慮されていない事に注意されたい．

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\,\log p_{i,k},$$

2.2 SORD（Soft Ordinal Regression Distillation）

SORD は **段階間の順序関係** を考慮した損失関数 L_{SORD} を最小化する．$\boldsymbol{s}_i = (s_{i,1}, \ldots, s_{i,K})$ とし $s_{i,k} = \max\!\Big(0,\ 1 - \frac{|k-y_i|}{K-1}\Big)$，　$\tilde{s}_{i,k} = \frac{s_{i,k}}{\sum_{j=1}^{K} s_{i,j}}$ とする．$\boldsymbol{p}_i$ との間で平均二乗誤差を取り，段階ラベル y_i から離れるほど損失が大きくなるよう損失関数を次のように定義する．

$$\mathcal{L}_{\mathrm{SORD}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\big(\tilde{s}_{i,k} - p_{i,k}\big)^2.$$

2.3 回帰（Regression）

最後に，段階ラベルを実数値として直接回帰する設定も比較用に導入する．二つの典型的な実装方法があるが，ここでは **確率の期待値型** を採用する[1)]．

$$\hat{y}_i = \sum_{k=1}^{K} k\, p_{i,k}, \qquad \mathcal{L}_{\mathrm{Reg}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2.$$

3 定式化

3.1 記法

N 個の埋め込みベクトル $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ と表記し，各々の $\mathbf{x}_i$ は D 次元ベクトルとする．すべての埋め込みベクトルが正規化されている，すなわち，$||\mathbf{x}|| = 1$ であると仮定する．ここで，$||\mathbf{x}||$ はベクトルのユークリッドノルムを表す．本手法は，これらの定義を満たす限り，単語埋め込みおよび文埋め込みの両方に適用できる．このため，以下では単語や文を単純化のため単に「項目」と呼ぶ．これらの埋め込みベクトル空間中の方向を示す D 次元のベクトルを $\mathbf{w} \in R^D$ とする． 我々の目標は，$\mathbf{w}$ の座標を求めることである．

3.2 最も簡単な事例

埋め込み空間上で「最も簡単な項目」というものの存在を仮定しよう．すると，埋め込みベクトル空間上では，この最も簡単な項目は最も広がりがない領域，すなわち，埋め込みベクトル空間上の点 $\mathbf{e}$ で表現できると仮定している事になる．また，埋め込みベクトル空間は，通常，意味的に類似しているものほど近くなる性質を持っている．この「意味的に類似」に難易度も含まれると考えると，「最も簡単な項目」を表す $\mathbf{e}$ との距離が小さい項目は易しく距離が遠い項目は難しいと考えられる．

ある $\mathbf{x}_i$ の方が，$\mathbf{x}_j$ より簡単であるとする．先の議論から，「最も簡単な項目」$\mathbf{e}$ との距離が，$\mathbf{x}_i$ の方が近いことになる．ここで，距離は単純にユークリッド距離で測ることにし，全ての $\mathbf{x}$ が $||\mathbf{x}|| = 1$ に正規化されているものとする．すると，次のように式変形することができる．

$$\begin{aligned}
||\mathbf{x}_i - \mathbf{e}|| &< ||\mathbf{x}_j - \mathbf{e}|| \qquad (1)\\
\Leftrightarrow ||\mathbf{x}_i - \mathbf{e}||^2 &< ||\mathbf{x}_j - \mathbf{e}||^2\\
\Leftrightarrow ||\mathbf{x}_i||^2 - 2\mathbf{x}_i^\top\mathbf{e} + ||\mathbf{e}||^2 &< ||\mathbf{x}_j||^2 - 2\mathbf{x}_j^\top\mathbf{e}^\top + ||\mathbf{e}||^2\\
\Leftrightarrow 1 - 2\mathbf{x}_i^\top\mathbf{e} + ||\mathbf{e}||^2 &< 1 - 2\mathbf{x}_j^\top\mathbf{e} + ||\mathbf{e}||^2\\
\Leftrightarrow \mathbf{e}^\top\left(\mathbf{x}_j - \mathbf{x}_i\right) &< 0\\
\Leftrightarrow \mathbf{w}^\top\mathbf{x}_i &< \mathbf{w}^\top\mathbf{x}_j \qquad (2)
\end{aligned}$$

ここで，式 2 では，$\mathbf{w}$ を $-\mathbf{e}$ と定義した．式 2 の解釈について考えよう．まず，$\mathbf{w}$ は埋め込みベクトル空間上の方向である．そして，$\mathbf{w}^\top\mathbf{x}_i$ は，この方向に見た順に $\mathbf{x}_i$ などの各項目を並べる事を意味する．すなわち，$\mathbf{w}$ は，埋め込みベクトル空間上で，この方向に進むと項目が難しくなることを示す．

1) 隠れ状態 $\mathbf{h}_i$ をそのまま線形変換し $\hat{y}_i = \mathbf{w}^\top\mathbf{h}_i + b$ とする純粋回帰型でも良い．実装しやすい方を選択されたい．

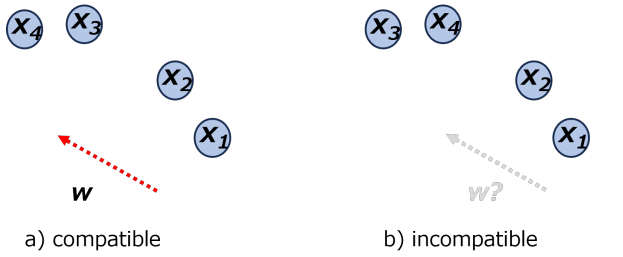

図 2　難易度方向の概要： a) $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ を 2 次元の埋め込みベクトルとして $\mathbf{x}_1 \prec \mathbf{x}_2 \prec \mathbf{x}_3 \prec \mathbf{x}_4$ と表記する．すなわち，$\mathbf{x}_1$ は $\mathbf{x}_2$ よりも簡単であり，$\mathbf{x}_2$ は $\mathbf{x}_3$ よりも簡単である，というようにアノテートされているとする．すると，この 2 次元空間において，$\mathbf{w}$ の方向にしたがって点を列挙すると，アノテーションと同じ順序になる．この埋め込みベクトルの集合は，アノテーションと「整列可能である」と定義する．b) この場合，$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ を，この 2 次元空間内で注釈されたのと同じ方法で順序付ける方向は存在しない．したがって，この埋め込みはアノテーションに対して「整列可能でない」と言う．

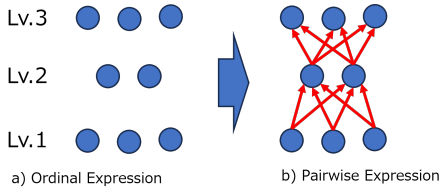

図 3　難易度アノテーションをペアワイズ制約に書き換える例．

3.3　難易度方向探索問題

ここでは，難易度方向 $\mathbf{w}$ の直感的な意味付けについて説明する．単純に言えば，難易度方向 $\mathbf{w}$ はデータセット（アノテーション集合）中の全ての項目が難易度順に並んで見えるような方向を表す．このような方向が見つかればよいが，大抵はそこまで都合の良い方向は存在しない．そこで，少し問題を緩和して，難易度順から少し外れた並びも許す事にする．

本節では埋め込みベクトル空間とアノテーション集合に対するあてはまりの良さとして解釈でき，大域的最適解が計算可能で近似などによらない「整列可能性」という量を定義する．そして，埋め込みベクトル空間と順序アノテーションが与えられたときに，両者の整列可能性の高さを，その埋め込みベクトル空間で計測した順序アノテーションの整列度合いとして定量化するものである．そこまで，まずは「整列可能性」の概念について図 2 で説明する．

これまでに埋め込みベクトル空間上で点が難易度順に並んで見える難易度方向 $\mathbf{w}$ の存在を仮定する問題について考えてきた．実際には，難易度のアノテーションは 2 項間関係というよりは段階 (レベル) のアノテーション (注釈) になっている事が多い．例えば，この質問の難しさは高等学校レベル，この質問の難しさは大学レベル，といったようなアノテーションである．こうした，段階順序付き注釈は容易にペアワイズ制約に変換できることを説明する．図 3 に例示する．図 3 a) では，8 つの項目があり，それぞれが Level1，Level2，Level3 の 3 つの尺度のいずれかで注釈されている．Level は「Lv」と略記することもある．レベルが大きいほど難易度が高くなるとする．図中，レベル 1 には 3 つの項目，レベル 2 には 2 つの項目，レベル 3 には 3 つの項目がある．アノテーション全体を，数学的に等価な変換により，図 3 b) に示される有向グラフに変換できる．

図 3 b) において，有向辺は順序対関係を表する．すなわち，辺の始点は終点よりも易しいことを意味する．一般性を損なうことなく，順序付き注釈の有限集合は等価な有向グラフに変換することができる．例えば，Level1 にある項目は，Level2 にあるすべての項目よりも簡単であることを意味する．レベル 2 には 2 つの項目があるため，レベル 1 の項目からそれら 2 つの項目に向かう 2 つの辺が存在する．辺はより易しいものからより難しいものへと向かうので，隣接するレベルの項目のみを考慮すればよいことに注意しよう．この例では，Level3 の項目は Level1 の項目よりも難しいという事実がグラフ構造に反映されており，Level1 の項目から Level3 の項目へと向かう有向パスをたどることができる．一方，その逆方向への移動は有向パスをたどっても不可能である．

同様に，まずアノテーションのデータセットを一対制約の集合に変換する．各レベルに I 個の項目があり，データセットに L 個の項目がある場合，変換後の一対辺の数は I^L 個となる．順序付きアノテーションが，図 3 に示すように，ペアワイズ関係に分解できることを以上のように直感的に説明した．

3.4　制約

図 3 に示したように，難易度アノテーションはペアワイズ制約の集合に変換できる．一般性を考慮して，ここでは一対比較制約の集合を用いる．各制約は，埋め込みベクトルの一方が他方よりも容易であると注釈付けされていることを示す．

以下のように，ペアワイズ順序制約の集合を定義する．$C = \{(i_1, j_1), \ldots, (i_M, j_M)\}$ とする．ここで，M は制約の数である．m 番目のペア，すなわち (i_m, j_m) は 1 つの制約を表す．ここで，$i_m \in \{1, \ldots, N\}$，$j_m \in \{1, \ldots, N\}$ は，いずれも埋め込みベクトルの添え字である．すなわち，(i_m, j_m) は，$\mathbf{x}_{i_m}$ が $\mathbf{x}_{j_m}$ よりも簡単であるとアノテートされている事をあらわす．ここでは，簡単のため，m の添え字を省略し，単に「より簡単」と評価されたベクトルを $\mathbf{x}_i$ と表記する．「より難しい」と評価されたベクトルは，$\mathbf{x}_j$ として表記する．

次に，ベクトル w によって作られた順序を扱うために，ベクトル x_i と x_j をベクトル w の方向に「射影」することを考える．ベクトル x_i と w のなす角を θ_i，ベクトル x_j と w のなす角を θ_j と表記する．目標は，x_i と x_j によって形成される制約を満たすように $\mathbf{w}$ を調整することである．一般性を損なうことなく，x_i が x_j よりも簡単であるという制約は，以下のように定式化できる．

$$
\begin{aligned}
||\mathbf{x}_i|| \cos\theta_i &< ||\mathbf{x}_j|| \cos\theta_j \\
\Leftrightarrow ||\mathbf{w}||||\mathbf{x}_i|| \cos\theta_i &< ||\mathbf{w}||||\mathbf{x}_j|| \cos\theta_j \\
\Leftrightarrow \mathbf{w}^\top \mathbf{x}_i &< \mathbf{w}^\top \mathbf{x}_j \\
\Leftrightarrow \mathbf{w}^\top (\mathbf{x}_i - \mathbf{x}_j) &< 0
\end{aligned} \tag{3}
$$

式 3 は，簡単に言えば，易しい問題と難しい問題のペアの 2 つの埋め込みの間で維持されるべき制約を述べている．ペアワイズ制約が M 個あり，単純化のため m を

省略したことを思い出そう．制約の添え字である m を導入すると次に説明する制約が満たされなければならない．

3.5 等式制約への書き換え

ここで，スラック変数である ξ_m を導入する．スラック変数は，不等式制約を等式制約に書き換えるために使用される．ξ_m は直観的には式 3 の制約を保つ度合いと理解できる．ξ_m の値が大きいということは，$\mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j)$ と 0 の間に大きな余裕（マージン）があることを意味する．なぜなら，大きな正の値の ξ_m を加えなければ，$\mathbf{w}^\top(\mathbf{x}_i - \mathbf{x}_j)$ を 0 にすることができないからである．つまり，これは m 番目の制約が大きなマージンで維持されることを意味する．一方，小さな ξ_m の値は，m 番目の制約を維持するマージンがほとんどないことを意味する．

$$\mathbf{w}^\top(\mathbf{x}_{i_m} - \mathbf{x}_{j_m}) + \xi_m = 0 \tag{4}$$

3.6 解の存在についての定理

式 3 の不等式を厳密に式 4 に合わせようとすると，非負の制約条件である $\xi_m \geq 0$ も必要となる．しかし，この制約条件を設定すると，ほとんどのデータセットにおける典型的な制約条件の数に対して，すべての制約条件を満たすような $\mathbf{w}$ を見つけることが事実上不可能になる．

線形代数の M 変数の 1 次連立方程式の解の存在可能性の議論を用いて，制約の数 M と埋め込みベクトルの次元 D，そして解の存在について，実用的な場面では通常満たされる条件の下で次の定理が成り立つ．

$M > D$ のとき 式 3 の実行可能解がない場合がある．
$M \leq D$ のとき 式 3 の実行可能解となる方向が存在する．

実用的なデータセットでは，$M > D$ であり，式 3 の実行可能解がない場合が多い．ただし，広く使われているデータセットの中にも M が小さいものがあり，その場合は逆に $M < D$ で適切な解を探し出せているかが問題になる．

3.7 緩和問題

実用性を考慮すると，むしろ，閾値 $T \leq 0$ を用いて負の ξ_m を許容し，これらの ξ_m が 0 以上であるという制約を除外し，すべての ξ_m の値の合計，すなわち制約を満たす度合いの合計を最大化することで，以下の最適化問題が得られる．

$$\begin{aligned} \text{maximize}_{\mathbf{w},\xi} \quad & \sum_{m=1}^{M} \xi_m \\ \text{s.t.} \quad & \forall m \in \{1,\ldots,M\}; \\ & \mathbf{w}^\top(\mathbf{x}_{i_m} - \mathbf{x}_{j_m}) + \xi_m = 0, \xi_m \geq T \end{aligned} \tag{5}$$

3.8 T は厳密さを制御する

この節では，式 5 を直感的に説明する．まず，図 2 では，2 次元の埋め込み空間において，2 つの埋め込みを整列可能性のないものにするのは，2 つのペアワイズ制約だけであることがわかった．それに対して，1 つのペアワイズ制約だけからなる制約に注目しよう．この場合，2 つの埋め込みの間の直線が方向すなわちベクトル $\mathbf{w}$ であるため，この集合は任意の 2 次元埋め込み空間と両立することは明らかである．同様に，制約の集合における制約の数が埋め込みベクトルの次元よりも少ない場合，常に整列可能である．

しかし実際には，アノテーションデータセットにおける制約の数は，埋め込み空間の次元より多い．典型的な単語や文の次元は 1 万未満である．しかし，ペアワイズ制約を考慮すると，それぞれ 100 件の事例からなる 2 つのレベルのみで構成されるデータセットでも 1 万を超える．したがって，実際には，式 5 において $T = 0$ とすると，式 5 は実行可能な解がなくなる．実際には，単に $T = -\infty$ と設定する，すなわち，制約条件 $\xi \geq T$ を削除すれば ξ_m への条件が緩和され式 5 は実行可能な問題となる．この場合，調整すべき変数として，すなわち $\{\xi_m | m = 1,\ldots,M\}$ と D 次元ベクトル $\mathbf{w}$ の座標として，M 個の等式制約条件がある一方で，$M + D$ 個の変数があるため，式 5 は実行可能である．これは，$T = -\infty$ の場合，必ず式 5 の実行可能解が存在することを意味しており，実用上優れた特性である．

式 5 の目的関数は，ξ_m の総和であり，制約条件を満たすためのマージンの総額と解釈できる．つまり，$T = -\infty$ とした場合でも，目的関数を最大化するように可能な限り負の ξ_m が回避される．これは，$T = -\infty$ とした場合の解における負の ξ_m は，与えられた埋め込み空間において，m 番目の制約が他の制約に従わないために，必然的に負に設定されたことを意味する．したがって，式 5 の解において，負の値を持つ ξ_m を持つ制約は，複雑性注釈の再確認の候補としてふさわしいものとなる．

なお，機械学習では，同様の緩和策がオリジナルのハードマージン SVM に適用され，ソフトマージン SVM が導出されている．ハードマージン SVM は，学習データが線形分離可能でない場合，実行可能な解を見つけることができないが，ソフトマージン SVM は可能である．自然言語処理で使用される SVM のほとんどは，実際には実用上の理由からソフトマージン SVM であることを考えると，$T = -\infty$ を設定するか，または ξ_m の正定値制約を外すことは実用的に妥当である．

3.9 $\mathbf{w}$ のノルム制約

式 6 に提案する問題の定式化を示す．これは式 5 で $T = -\infty$ とおいて ξ に対する条件を取り除き，$||w||$ のノルム制約を足した問題になっている．

$$\begin{aligned} \text{maximize}_{\mathbf{w},\xi} \quad & \sum_{m=1}^{M} \xi_m \\ \text{s.t.} \quad & \forall m \in \{1,\ldots,M\}; \\ & \mathbf{w}^\top(\mathbf{x}_{i_m} - \mathbf{x}_{j_m}) + \xi_m = 0 \\ & ||\mathbf{w}||^2 = 1 \end{aligned} \tag{6}$$

方向としては，ノルム制約がなくとも整列可能な方向 $\mathbf{w}$ は見つかる．しかし，この場合は $\mathbf{w}$ のノルムが問題ごとに変わってくるので，**問題間で目的関数値を比較することができない**．式 6 の目的関数は，アノテーションと埋め込み空間の当てはまりの良さを表していると解釈できる．**ノルム制約を課す理由は，異なる問題でも目的関数値を比較可能にし，よりアノテーションとの整列可能性（目的関数値）が高いアノテーションがより高い整列度合いを示す，解釈しやすい指標を作るためである．**

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
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

3.10　差分ベクトルの平均の最適性

本節では，この最適化問題の解が，差分ベクトルの単純な平均として解釈できることを示す．この差分ベクトルの平均は，平均ベクトルの差分とみなすこともできる．具体的には，図 2 に示すように，矢印の先端に位置する文または単語の平均埋め込みベクトルから，矢印の尾部に位置する文または単語の平均埋め込みベクトルを差し引いて得られるベクトルが，前節で議論した数学的最適化問題の解に対応する．

ここでは，上記の問題を解くために以下の操作を行う．$\xi_m = -\mathbf{w}^\top(\mathbf{x}_{i_m} - \mathbf{x}_{j_m})$ であることに留意して，この式を目的関数に単純に代入する．加えて，ラグランジュ乗数 λ を導入し，等式制約 $1 - ||\mathbf{w}||^2$ に付加することで，目的関数に $\lambda(||\mathbf{w}||^2 - 1)$ を加え，ノルム制約を除去する．最終的に，以下の式が得られる．

$$\begin{aligned} \text{maximize}_{\mathbf{w}} \quad & \sum_{m=1}^{M} -\mathbf{w}^\top(\mathbf{x}_{i_m} - \mathbf{x}_{j_m}) \\ + \quad & \lambda(1 - ||\mathbf{w}||^2) \end{aligned} \tag{7}$$

これは連続な目的関数を持つ制約のない最適化問題であるため，$\mathbf{w}$ に関して微分することで式 7 の最大値を求めることができる．式 7 を $\mathbf{w}$ で微分すると，次の式が得られる．

$$\begin{aligned} \sum_{m=1}^{M} \quad - \quad & (\mathbf{x}_{i_m} - \mathbf{x}_{j_m}) - 2\lambda\mathbf{w} = \mathbf{0} \\ \mathbf{w} \quad \propto \quad & \sum_{m=1}^{M} (\mathbf{x}_{j_m} - \mathbf{x}_{i_m}) \end{aligned} \tag{8}$$

式 8 は，目的関数の最大値が単に差分ベクトルの平均で表されることを示している．つまり，これらの差分ベクトルの平均を計算し，そのノルムを 1 に正規化することで，序論で述べた適合方向を得ることができる．この性質により，毎回最適化問題を解くことなく最適なベクトルを選択することが可能となる．

3.11　提案: 整列度合い指標

以上より，提案手法は以下のようにまとめられる：

1. 差分ベクトルの平均の計算：差分ベクトルの平均を計算し，そのノルムを 1 に正規化する．これにより，$\mathbf{w}$ の最適解が得られる．
2. 目的関数値の計算：最適な $\mathbf{w}$ が得られたら，目的関数値を計算する．この値は，与えられたデータセットと埋め込み空間との適合度を反映する．

具体的には次のようにする．式 8 の解は $\mathbf{w} \propto -\sum_{m=1}^{M}(\mathbf{x}_{i_m} - \mathbf{x}_{j_m})$ 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 6 の最適化問題の目的関数値は $-N_iN_j\hat{\mathbf{w}}^\top(\bar{\mathbf{x}_i} - \bar{\mathbf{x}_j}) = N_iN_j||\bar{\mathbf{x}_i} - \bar{\mathbf{x}_j}||$ と表せる．そこで 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ユークリッドノルム $||\bar{\mathbf{x}_i} - \bar{\mathbf{x}_j}||$ で定義する．

このように**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 2 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．**

提案する「整列度合い」の計算量は，**各段階の平均ベクトルの計算に必要な計算量が主である．段階間の制約について列挙したり最適化問題を解かずに求められる大きな利点がある．**

3.12　ペアワイズ線形判別分析との関係

提案する「整列度合い」は，ペアワイズ線形判別分析 [10] の考え方を用いても解釈できる．線形判別分析は，各次元をスケーリングしながら，2 つのクラス間の分離を最大化する方向を識別する．図 2 における矢印の底部をクラス 1，先端をクラス 2 とみなすと，「整列度合い」の最適化問題は各次元のスケーリングを行わない線形判別分析の一形態と見なすことができる．スケーリングを省略するのは，埋め込みベクトル空間は前段の「BERT 等」に相当するニューラルネットワークで特定の次元のみ大きなスケールを持つようなことがないように調整されているはずであるからである．

4　静的埋め込み実験

4.1　静的単語埋め込みでの実験

静的埋め込み（static embedding）とは，微調整しない埋め込みをそのまま特徴量に用いる手法の総称である．難易度推定の場合は，微調整しない埋め込みを特徴量に用いて古典的な機械学習手法で分類する場合がこれに当たる．単語埋め込みを用いた整列度合いの実験を行った．データセットとしては，英単語の難易度について CEFR（ヨーロッパ言語共通参照枠）の基準に基づいて人手でアノテーションされたデータセットである CEFR-J Vocabulary Profile (`https://github.com/openlanguageprofiles/olp-en-cefrj?tab=readme-ov-file`) を用いた．単語埋め込みとしては，FastText[11] を用いた．比較のための手法としては，すなわちサポートベクターマシン（SVM）を用いた．検証データセット上で調整するパラメータ C があるため，0.1，1.0，10.0 から検証データを用いて選んだ．

次の手順で，整列度合いを確認した．CEFR-J は 4 段階に難しさがアノテーションされている．このうち，2 番目に簡単な単語集合から 100 語をランダムに抜き出した．そして，各語について，前述のペアワイズ制約への書き換え手順を用いて，最も簡単な単語集合と各語のペアワイズデータを訓練データとして，各語から 3 番目に

簡単な単語集合とのペアワイズデータをテストデータとした．提案手法も，SVM も精度は 100%であった．

この方法では各語について 100 通りの部分データセットがあるとみなすことができる．英単語の難しさについては，SVL というより細かい 12 段階のアノテーションデータセットが提案されている (https://eow.alc.co.jp/svl_level12.html)．そこで，100 個のデータセットについて，アノテーションの整列度合いを目的関数値を通じて計測した．最も簡単な単語とのアノテーションの整列度合いであるので，より細かい 12 段階の SVL の方で難しい単語の方が，整列度合いが高くなっているはずである．実際に，順位相関係数で Spearman の R を用いて検定を行ったところ，$p < 0.01$ で統計的に有意な相関が示された．従って，提案手法で単語では整列度合いが検知できることが示された．

4.2 静的文埋め込み実験

提案手法は，埋め込みベクトルであれば単語埋め込みでも文埋め込みでも実行できる．次に，提案手法を用いて文埋め込みのアノテータ間の不一致が予測できるかを確認した．

データセットには CEFR-SP データセット [12] を用いた．このデータセットでは，前述の CEFR という基準に基づいて，英文の難しさが 2 名のアノテータによって 4 段階にアノテートされている．そこで，まず，1 名のアノテータのデータを用いて前述の単語の場合と同様に，各文についての「整列度合い」を計算した．文埋め込みには “multilingual-e5-small” を用いた (https://huggingface.co/intfloat/multilingual-e5-small)．「整列度合い」では文ごとの最適解の目的関数値を比較できる．具体体には，ξ の値はマージンと見なすことができるので，この値が大きい方が，整列度合いが高いと考えられる．

次に，もう 1 名のアノテータの結果と照らし合わせ，アノテーションの一致不一致と「整列度合い」のスピアマンの順位相関係数を計算した．その結果，相関係数は 0.262 であり，「やや相関」している事が確認された（https://www.study-channel.com/2015/08/spearmans-rank-correlation-coefficient.html）ものの，統計的有意性は見られなかった．

4.3 整列度合いを用いた分析

次に，前節で示した CEFR-SP データセットを用いて，複数の埋め込みベクトルの中から最も適合する埋め込みベクトルを選定する実験を行う．ここで用いた CEFR-SP データセットは，各レベルの組み合わせを単一のデータセットとして扱ったものである．

表 1 は，上記の文埋め込み実験の設定に基づいて提案手法により計算された整列度合いを示している．各列は CEFR の難易度注釈のペアに対応する．なお，*sentence-transformers/all-MiniLM-L6-v2*，*BAAI/bge-m3*，*intfloat/multilingual-e5-small* ，*intfloat/multilingual-e5-large* の出典は，https://huggingface.co/sentence-transformers/ の後にそれぞれの手法名を付けた URL から得られる．

これらの整列度合いは，前述の手法を用いて CEFR-SP レベルのペアからなる注釈集合に対する方向ベクトルを算出し，さらに目的関数値を計算することで求められる．スコアが高いほど，各難易度の注釈と埋め込み空間との適合度が高いことを示す．

すべての CEFR 難易度ペアにおいて，最初の埋め込みが最も良いスコアを示していることから，“sentence-transformers/all-MiniLM-L6-v2” が本 CEFR-SP 注釈データセットに最も適していると結論付けられる．

全体として最も高い整列可能性を示す all-MiniLM-L6-v2 モデルと，multilingual-e5-small モデルは共に埋め込み次元が 384 である．しかし，後者は全体として低い整列可能性を示しており，埋め込みの次元数が整列可能性に直接影響しないことを示唆している．同様に，BAAI/bge-m3 および intfloat/multilingual-e5-large モデルはそれぞれ 1024 次元の埋め込みを有するが，低い整列度合いを示しており，次元数の増加が必ずしも整列可能性を向上させるわけではないことが明らかとなる．

また，興味深い点として，全ての手法において (A1, B2) ペアの整列度合いが最も高いことが挙げられる．本データセットでは，A1 が最も容易なレベル，B2 が最も難易度の高いレベルを表す．これは，最も単純なレベルと最も複雑なレベルの文間の方向ベクトルが，埋め込み空間内でより一貫して整列していることを示唆しており，直感的な期待と一致する結果である．

対照的に，(A1, A2) ペアは全ての埋め込み空間において一貫して最も低いスコアを示す．A1 が最も容易なレベルであり，A2 がやや難易度の高いレベルであることから，単純なレベル間の区別が埋め込み空間内でより曖昧であることが示唆され，これも直感的な理解と合致する．

次に，先の埋め込みについて，SVM による予測実験の結果を示す．(A1,B1) を訓練データに，(B1,B2) をテストデータに用いた．実験の結果，表 1 の上から下の順に 0.63, 0.37, 0.26, 0.23 となった．すなわち，**同じ次元では，整列度合いの高い指標の方が高精度であった**．

4.4 静的埋め込みによる精度予測

整列度合いがテストデータでの難易度推定値を予測できるかを別のデータセットでも確認したい．そこで，「整列度合い」をさらにデータセット非依存にした FPNI 指標を次に定義する．

4.4.1 FPNI 指標

整列度合い指標は，段階の異なる 2 つの段階が埋め込みベクトル空間上でどれだけうまく並んで見えるかを表すもので式 6 の目的関数値である．ただ，データセットにより，どの段階間で埋め込みベクトルが並んで見えるのかが異なる．全ての段階間での整列度合い指標の平均を取る事も考えられるが，そのようにするとデータセットによって異なる段階の数や，データセット中の途中の段階もうまく難易度順に並べられているか，といったことに依存してしまう．例えば，最も簡単なレベルや最も難しいレベルは適切にデータセットがアノテーションされているが，途中の段階はアノテーターによってかなりレベル付けに差があるといった現象はよく見られる．そこで，なるべくデータセットの作り方や途中の段階のアノテーションのうまさに依存しない指標とするため，単純に所与のデータセットの最も簡単なレベルと最も難しいレベルの間で取った整列度合い指標を以後，FPNI (Farthest Pair Neatness Index) 指標と定義する．

第二言語学習者向けの可読性推定のデータセットとして代表的なものであり，全文が語学教師に

表 2　OneStopEnglish データセットでの静的埋め込み実験における検証 FPNI とテスト Macro-F1

モデル	検証 FPNI	テスト Macro-F1
all-MiniLM-L6-v2	0.6096	0.3653
multilingual-e5-small	0.1298	0.2022
multilingual-e5-large	0.1327	0.5169
sentence-bert-base	0.4200	0.3100
distilbert-base	0.3800	0.2900

表 3　学習中のエポック別の FPNI の動的な変化．F1 は Macro-F1 の略．検証は検証データ（開発データ）を用いて計測した値であることを示す．

エポック	BERT		ModernBERT	
	検証 F1	検証 FPNI	検証 F1	検証 FPNI
0	0.1667	0.1201	0.1795	0.0820
1	0.5556	0.0316	0.7541	0.1298
2	0.5556	0.0456	0.6588	0.1154
3	0.5746	0.0669	0.8665	0.5386
4	0.5556	0.0450	0.9470	0.6835
5	0.9381	0.6463	0.9386	0.8328

よってチェックされた質の高いデータセットとして有名な OneStopEnglish データセット [13] を用いて，検証 FPNI からテストデータの Macro-F1 を予測できるか確認した所，統計的有意な相関を得た（表 2, 厳密置換検定，$p < 0.05$）．ここで Macro-F1 は Macro-$F_1 = \frac{1}{K}\sum_{k=1}^{K} F_1^{(k)}$，$F_1^{(k)} = \frac{2P^{(k)}R^{(k)}}{P^{(k)}+R^{(k)}}$，$P^{(k)} = \frac{\mathrm{TP}^{(k)}}{\mathrm{TP}^{(k)}+\mathrm{FP}^{(k)}}$，$R^{(k)} = \frac{\mathrm{TP}^{(k)}}{\mathrm{TP}^{(k)}+\mathrm{FN}^{(k)}}$ と定義される．ただし，K はクラス数（段階数），TP,FP,FN は，それぞれ段階 k における真陽性，偽陽性，偽陰性である．OneStopEnglish データセットは初級（Elementary），中級（Intermediate），上級（Advanced）の 3 種類に分かれている．

5　動的埋め込み実験

本節では BERT などの学習中に前述の「整列度合い指標」を出力させる．BERT や ModernBERT のファインチューニングの学習量は，大まかにはデータを何回全件見たかという回数に相当する「エポック」で通常，測られる．学習が進む（エポックが大きくなる）につれて，BERT 等が出力する埋め込みも実際に訓練中のデータセットに沿ったものになると考えられるので，データセット中の順序に反するような事例，すなわち制約に違反するような事例の量は少なくなってくると考えられるため，式 6 の目的関数値である「整列度合い指標」は大きくなっていくと予想される．本節では，実際にこれが BERT と ModernBERT で起こっていることを確認する．

動的埋め込みモデル（BERT 系）では，訓練の進行に伴い FPNI 指標が変化する．OneStopEnglish データセットを用いて，BERT と ModernBERT の 2 つのモデルでエポック別の FPNI の変化を追跡した．

5.1　実験設定

次に，主な実験設定を列挙する．モデルは BERT (bert-base-multilingual-cased) と ModernBERT (answerdotai/ModernBERT-base) を比較した．データセットは前述の OneStopEnglish で，手法は多クラス分類とし，バッチサイズ 8, 学習率 2e-5, エポック数 5 とした．

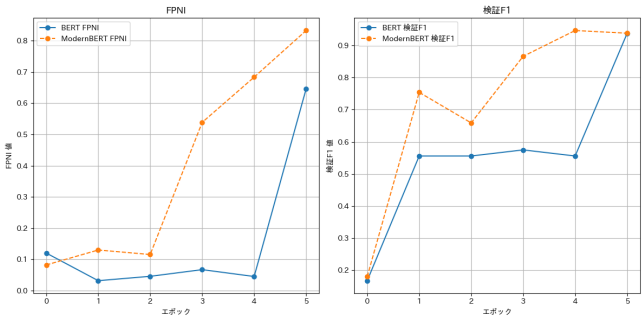

図 4　学習中のモデルのエポック別検証 FPNI 及び検証 Macro-F1 の変化．左図は FPNI 指標の変化を，右図は検証データでの Macro-F1 スコアの変化を示す．両モデルとも訓練が進むにつれて FPNI 値が適切に増加し，表現空間の分離度が改善している．

表 4　文脈内学習実験結果

モデル	データセット	検証データ		テストデータ
		Macro-F1	FPNI	Macro-F1
GPT-2	OneStopEnglish	0.3577	0.0261	0.3714
GPT-2	BEA2024	0.2491	0.0203	0.2183
ModernBERT-base	OneStopEnglish	0.3230	0.0113	0.2884
ModernBERT-base	BEA2024	0.2452	0.0407	0.3626

5.2　結果

表 3 に各エポックでの検証 F1 と FPNI 値を示す．図 4 は学習中の FPNI 指標を表している．参考のため，検証データでの Macro-F1 指標も示した．

まず，表 3，図 4 より，エポック数が増えるにしたがって，FPNI 値も上昇している事が分かる．これは直感的に**学習が進んでいない所では検証データが整列可能でなかった埋め込み空間が，学習が進むにつれて検証データが埋め込み空間中で整列可能になっていく**事を表す．すなわち，学習が進むにつれて埋め込み空間の形が変わり，検証データを整列できるような形にしていくのである．FPNI は検証データで測定しているため，例えば損失関数の値のように訓練データから計算されている値ではない事に注意されたい．また，埋め込み空間の次元は一定であるため，埋め込み空間の次元によって決まる値でもない．このように，この結果は，難易度推定の既存研究では著者の知る限り新規で有用な結果である．なお，この実験では，最終的に BERT は Macro-F1 値が 0.5556, ModernBERT が 0.9472 で，ModernBERT の方が予測性能が圧倒的に高かった．

6　文脈内学習実験

提案する「整列度合い指標」の素晴らしい点は，純粋に幾何的な指標であるため，**モデルに非依存である**点である．このため，文脈内学習 (In-Context Learning, ICL) に対しても，モデルが回答時する埋め込みベクトルが取得できれば，同様に計算可能である．文脈内学習は，大規模言語モデルに対してモデルのパラメータを更新することなく（すなわち，ファインチューニングをすることなく），入力プロンプトに少数の例示（few-shot examples）を含めることで，予測時に学習しているのと同様の効果を持たせる手法である．従って，本来は文脈内学習では検証データを用意して最適なハイパーパラメタを選択するようなことは必要ない．本研究では，純粋に性能を比較する目的で，ファインチューニングの必要なモデルで用いた検証データを，文脈内学習の場面でも用いて回

答の埋め込みベクトルを出力させ，検証データについている難易度ラベルから前述の「整列度合い指標」を計測している．なお，ChatGPT をはじめとする多くの生成 AI モデルでは，実際に使われている最新のモデルでは回答の埋め込みベクトルは取得できない（回答の出力確率は取得できる API も多い）．今回は，単純に文脈内学習でも提案手法が利用できることを示せれば十分であるため，GPT-2 を用いて実験を行った．

6.1 医学難易度予測タスク BEA2024

データセットとしては，前述の [13] に加えて，医学の難易度予測タスクのデータセットを用いた [2]．これは，自然言語処理分野の教育に関するワークショップである BEA 2024 の shared task として開催されたもので，大人数受験者の正答率から計算された信頼できる医学の設問の難易度の値を予測するものである．語学学習と比較して，科学等の質問データセット自体は大規模言語モデルのベンチマーク用のデータが多くあるものの，ベンチマーク目的であるため，人間に回答させたときの難易度が付与しにくい．また，科学などの質問の難易度は，その性質上，教員に難易度を付与させる場合でも相当に前提条件を共通化しないと教員間で難易度を一致させることが難しい．BEA2024 は現在入手可能な中で，設問文から難易度を予測する事を主眼とするタスクのデータセットで最も信頼できるものの 1 つである．

表 4 に結果を示す．FPNI は検証データでの FPNI であるので，テストデータはもちろん用いていない．この実験結果からは，同じ手法の中の FPNI の大小と，**テストデータでの** Macro-F1 性能の大小が一致している事が分かる．これは，検証データでテストデータなしで計算された FPNI 指標が，大まかにテストデータでの Macro-F1 を予測できている事を表す．特筆すべきなのは，ModernBERT-base では，検証データでの Macro-F1 の大小と検証データでの FPNI が逆転しており，FPNI の方が Macro-F1 値を上手く予測できている事である．

7 結論

本研究ではテキストからの難易度推定のための様々な順序回帰手法を数理的にまとめ，埋め込みベクトルの出力部分と埋め込みベクトルから難易度推定を行う部分の 2 つの部分に分かれることをまず示した．次に，埋め込みベクトル部分がどの程度難易度データセットに沿っているのかを評価する「整列度合い指標」を提案した．従来手法と異なり，提案する指標は純粋に幾何的な指標であるためモデルに依存せず，静的埋め込み，BERT 等の学習中，文脈内学習の全ての画面で使えることを示した．さらに，検証データ（開発データ）中の「整列度合い指標」を用いてテストデータの Macro-F1 が予測できることを示した．さらに，語学学習のみならず医学の質問難易度推定データセットでも提案手法の性能を確認し，提案手法が幅広く使える有望な尺度であることを示した．

今後の課題としては，提案する尺度は次元が変わると比較できないため同じ次元の中での大小を比較するしかないため，次元が変わっても比較できる尺度を提案する事が挙げられる．また，今回提案した手法が単に埋め込み空間をつぶして難易度の予測値を出力するだけの従来法と大きく異なるのは，本来ベクトルの方向の多様性によって多様な意味を扱える埋め込みの性質を利用し，多様な「難易度」（難しくする方向）扱える点にある．個別最適な学習支援のためには，絶対的な難易度を頭ごなしに出力しても仕方なく，学習者ひとりひとりにとっての難易度を計算できる枠組みが望ましい．提案した「整列度合い」はモデルに依存しない幾何的な指標である．今後，学習支援システムの技術が進展し，例えば「個々の学習者が理解する意味の空間」をモデル化するような埋め込み空間が提案された場合に，本研究で提案した指標は，「一般的な難易度尺度とは異なる理解の仕方をしている学習者」を見つけ出す事や，「理解の仕方が異なる学習者を見つけ出す」ことにもつながり，有用な応用が期待される．

謝辞

本研究は，科学技術振興機構さきがけ研究費 (JPMJPR2363)，JSPS 科研費 22K12287 の支援を受けた．

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 *Proc. of NAACL–HLT*, pp. 4171–4186, 2019.

[2] Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proc. of BEA*, pp. 470–482, June 2024.

[3] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, December 2024.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[5] Andreas Messalas, Yiannis Kanellopoulos, and Christos Makris. Model-agnostic interpretability with shapley values. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–7. IEEE, 2019.

[6] Joseph Marvin Imperial. BERT embeddings for automatic readability assessment. In *Proc. of RANLP*, pp. 611–618, 2021.

[7] Raul Díaz and Amit Marathe. Soft labels for ordinal regression. In *Proc. of CVPR*, pp. 4732–4741, 2019.

[8] Ho Hung Lim and John S. Y. Lee. Improving readability assessment with ordinal log-loss. In *Proc. of BEA*, pp. 343–350, 2024.

[9] Justin Lee and Sowmya Vajjala. A neural pairwise ranking model for readability assessment. In *Findings of ACL*, pp. 3802–3813, 2022.

[10] Hidehiko Kamiya and Akimichi Takemura. On rankings generated by pairwise linear discriminant analysis ofmpopulations. *Journal of Multivariate Analysis*, Vol. 61, No. 1, pp. 1–28, 1997.

[11] Joydeep Bhattacharjee. *fastText Quick Start Guide: Get started with Facebook's library for text representation and classification*. Packt Publishing Ltd, 2018.

[12] Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. CEFR-based sentence difficulty annotation and assessment. In *Proc. of EMNLP*, pp. 6206–6219, December 2022.

[13] Sowmya Vajjala and Ivana Lučić. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proc. of BEA*, pp. 297–304, June 2018.