文章と数式の表現学習を用いた数学における類題検索

居樹　彩乃†　　湯本　高行†

† 兵庫県立大学大学院情報科学研究科　〒 651-2197 神戸市西区学園西町 8 丁目 2-1
E-mail: †ad24l031@guh.u-hyogo.ac.jp, ††yumoto@sis.u-hyogo.ac.jp

あらまし　ある問題の類題演習は特に数学学習において非常に重要となる。そこで本研究では高校数学における類題検索について扱う。数学の問題は文章部分と数式部分の 2 種類が存在するため、それぞれ別のアプローチで類似度を算出する。まず、文章部分と数式部分を各文、各式単位でベクトルに変換を行う。この時、文章部分は数式を除く問題の文章部分、数式部分は数学の公式を入力として教師なし SimCSE で作成したモデルを使用する。ここで求めた各文、各式のベクトルを使用して文章部分、数式部分それぞれにおける問題間の双方向の類似度を算出する。なお文章部分に関しては、ベクトルをもとに算出した類似度に対して数学用語の集合間類似度をかけ合わせてることで文章部分の類似度を算出する。以上の手順で求めた文章部分と数式部分の類似度にパラメータを加えたものを問題間の類似度とする。

キーワード　情報検索, 教育, 表現学習, 数学

1 はじめに

ある問題の解法を定着させるためには、その問題の類題を演習することが重要である。しかしながら、類題を問題集から探し出すことは非常に時間を要する上に、判断基準の個人差が生じる。特に新人の講師や学生は経験値が少ないため、類題を見つけることが困難である。中でも高校数学に着目すると出現する解法パターンが非常に多いため、特に手間がかかってしまう科目といえる。また、日本語を使用した数学の問題を解くための研究は多く存在するが、数学の問題の類似性に関する研究は少ない。そこで本研究では、高校数学における問題を対象とする類題検索方法について記す。なお、本研究における数学の問題とは、ある問題の問題文と解説文の両方を指す。

数学の問題は文章部分と数式部分から構成されているため、それぞれ別の方法で類似度を算出した上で全体の類似度を求める。文章部分は問題集に記載されている例題、数式部分は公式集を元に教師なし SimCSE でモデルの作成を行う。このモデルを利用し、文章部分と数式部分それぞれをベクトル化した上で、そのコサイン類似度を利用する。それに加えて、文章部分は数学の問題の特性を反映させるために数学用語の集合間類似度を用いる。このように文章と数式で別のアプローチを行うことで、数学の問題の特性を上手く反映させた類似度を求めることができると考える。そして、この提案手法が数学の類似性を正しく表すことができているのかの評価を行う。この際のテストデータには、実際の問題とそれに対する類題を使用する。

本論文の構成は以下の通りである。まず、第 2 節では関連研究の紹介を行う。続く第 3 節では具体的な数学の問題間の類似度算出方法を提案する。第 4 節では、その提案手法を用いた実験を行い、評価ならびに考察を述べる。最後の第 5 節では、本論文のまとめを記す。

2 関連研究

2.1 数学問題の類似性評価

上江洲らの研究 [1] においても、文章部分と数式部分それぞれの類似度を定義した上で数学の問題の類似度を算出している。しかしながら 1 文程度で完結する単純な問題しか扱っておらず、テストや入試で出題されるような問題に対応するのは難しい。そのため、本研究では入試問題のように複数の要素が含まれるような問題における類似度を算出方法について記す。

2.2 MathBERT

MathBERT [2] とは、数式に特化した BERT の事前学習済みモデルである。事前学習の際には数学だけでなく、物理や化学といった幅広い学問で登場する数式を使用しており、数学関連タスクにおいて高い精度を誇る。しかしながら、多方面の数式を対象としているがゆえに高校数学という狭い領域においては、どの数式も似た表現になってしまうことが課題点としてあげられる。加えて MathBERT は学習時の入力として数式とともに、英語で記載されたコンテクストを与えている。よって、本研究で扱う日本語の類題を含めた数学の問題には対応していないため、使用するには工夫が必要となる。

2.3 SimCSE

対照学習の一種である SimCSE [3] は、ある同一の入力に対して毎回異なるドロップアウトを用いて自身の推測する学習である。これは教師あり学習と同等の精度を発揮すると論文内で述べられている。この SimCSE は生物医学分野 [4] や医療分野 [5] などの広い分野で応用されている。また、日本語を対象としても有効性が示されている [6]。以上より、本研究では数式と日本語で書かれた文章に対して SimCSE を用いる。

例題 I.三角関数の合成

$\sin\theta + \cos\theta$ の最大値を求めよ。
ただし、$0 \leq \theta \leq 2\pi$ とする。

指針

$\sin\theta + \cos\theta$ の合成を行う

解答

$$\sin\theta + \cos\theta$$
$$= \sqrt{2}\left(\frac{1}{\sqrt{2}}\sin\theta + \frac{1}{\sqrt{2}}\cos\theta\right)$$
$$= \sqrt{2}\sin\left(\theta + \frac{\pi}{4}\right)$$

ただし、$0 \leq \theta \leq 2\pi$ より $\frac{\pi}{4} \leq \theta + \frac{\pi}{4} \leq \frac{9}{4}\pi$
このとき、$-1 \leq \sin\left(\theta + \frac{\pi}{4}\right) \leq 1$
よって、$-\sqrt{2} \leq \sqrt{2}\sin\left(\theta + \frac{\pi}{4}\right) \leq \sqrt{2}$
したがって、$\sin\theta + \cos\theta$ の最大値は$\sqrt{2}$

図 1　数学の問題例

3　提 案 手 法

本研究では、文章部分と数式部分でそれぞれの類似度を用いて問題の類似度を算出する。問題 p に対する問題 q の問題間類似度 $sim(p,q)$ について、文章部分の類似度を $sim_{text}(p,q)$、数式部分の類似度を $sim_{math}(p,q)$ とし、パラメータ α を用いると式 1 と表すことができる。

$$sim(p,q) = (1-\alpha)sim_{text}(p,q) + \alpha \cdot sim_{math}(p,q) \quad (1)$$

3.1　数学の問題について

本研究では、図 1 のような緑色で示される問題部分と青色で示される解説部分のまとめて問題と定義する。文章部分に関しては原則として句点で区切ったものを一文と処理している。数式部分に関しては、原則として等号で区切ったものを一文としている。

3.2　数 学 用 語

数学 IA・IIB の問題集に記載されている各単元の用語まとめから抽出した全 186 単語を数学用語と定めた。その数学用語の例を示したものが表 1 である。

表 1　数学用語の例

余事象	2 次不等式	加法定理
条件付き確率	円周角の定理	共役な複素数

また、トークナイザーに数学用語を追加する処理を行った。

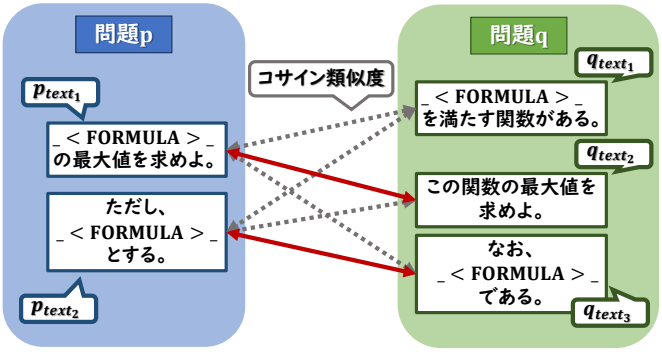

図 2　類似度算出のイメージ

この処理を行うことで、表 2 のようにトークナイズした際においても数学用語をひとまとまりとして捉えることができる。

表 2　トークナイズの例

元の文	余弦定理を適用する
数学用語の追加なし	余弦/定理/を/適用/する
数学用語の追加あり	余弦定理/を/適用/する

3.3　文章部分の類似度

3.3.1　文章部分の事前学習

文章部分の類似度を求めるにあたり事前学習として数学 IA・IIB の問題集に記載されている例題全 348 題を用いて、教師なし SimCSE を行った。この際、例題の中に含まれる数式を『_<FORMULA>_』という 1 単語に置換をした。ベースモデルに関しては SentenceBERT [7] を使用しており、トークナイザーに数学用語と『_<FORMULA>_』の合計 84 単語を新たに追加している。また、それに伴ってベースモデルとトークナイザーの word_embeddings を 32084×768 に変更した。

3.3.2　文章部分の片方向の類似度

まず問題 p からみた問題 q の文章部分の類似度 $sim_{text}(p \Rightarrow q)$ を算出する。このとき 3.3.1 節で作成した学習済みモデルを使用し、数学の問題内の各文をベクトル化する。

問題 p,q は、それぞれ n,m 個の文から構成されており、その文をそれぞれ p_{text_i}, q_{text_j} とすると $sim_{text}(p \Rightarrow q)$ は式 2 と表す。

$$sim_{text}(p \Rightarrow q) = \frac{1}{m}\sum_{i=1}^{m}\max_{1\leq j\leq n}\cos(p_{text_i}, q_{text_j}) \quad (2)$$

この算出方法の具体例を示したものが図 2 となる。この例では問題 p には p_{text_1} と p_{text_2} が存在するため、それぞれに対応する文章を問題 q から検出する。p_{text_1} とコサイン類似度が最も大きい文章は赤線で結ばれている q_{text_2} であるため、このコサイン類似度の値を抽出する。p_{text_2} でも同様の操作を行い、コサイン類似度の最大値を抽出する。これらのコサイン類似度の最大値らを平均した値が $sim_{text}(p \Rightarrow q)$ となる。

また同様の手順で問題 q からみた問題 q の文章部分の類似度

$sim_{text}(q \Rightarrow p)$ を表現すると、式 3 となる。

$$sim_{text}(q \Rightarrow p) = \frac{1}{n}\sum_{j=1}^{n} \max_{1 \leqq i \leqq m} \cos(p_{text_i}, q_{text_j}) \quad (3)$$

3.3.3 数学用語の集合間類似度

人間が数学の問題を似ていると判断する際には、解法だけでなく共通する数学用語の存在も大きいといえる。そこで、3.3.2 で求めた類似度に加えて数学用語の集合間類似度も算出し、文章部分の類似度算出に使用する。この集合間類似度には、包含関係も考慮できる Simpson 係数を用いる。問題 p,q に含まれる数学用語集合をそれぞれ A_p,A_q とすると、問題 p,q の数学用語の集合間類似度 $sim_{words}(p, q)$ は式 4 と表すことができる。

$$sim_{words}(p, q) = \frac{|A_p \cap A_q|}{\min(|A_p|, |A_q|)} \quad (4)$$

なお、$|A_p|$,$|A_q|$ は問題 p,q にそれぞれ含まれる数学用語集合の要素数を示す。

3.3.4 文章部分の類似度算出方法

3.3.2、3.3.3 節で求めた値を用いて、文章部分の類似度 $sim_{text}(p, q)$ を算出する。具体的な算出方法は、式 5 の通りである。

$$sim_{text}(p, q) = \frac{1}{2}(sim_{text}(p \Rightarrow q) + sim_{text}(q \Rightarrow p)) \times sim_{words}(p, q) \quad (5)$$

なお、片方向の類似度の平均をとることで双方向の類似度を表現する。

3.4 数式部分の類似度

3.4.1 数式部分の事前学習

数式部分の類似度を求めるにあたり事前学習として数学 IA・IIB の問題集に記載されている各単元の要点まとめに含まれる 385 個の数式を使用して教師なし SimCSE を行った。ベースモデルとトークナイザーには MathBERT を使用し、エポック数は 1 とした。

3.4.2 数式部分の片方向の類似度

まず文章部分の片方向の類似度と同様で問題 p からみた問題 q の数式部分の類似度 $sim_{math}(p \Rightarrow q)$ を算出する。このとき 3.4.1 で作成した学習済みモデルを使用し、数学の問題内の各文をベクトル化する。

問題 p、q は、それぞれ k、l 個の式から構成されており、その式をそれぞれ p_{math_i}, q_{math_j} とすると $sim_{math}(p \Rightarrow q)$ は式 6 と表す。なお、式については原則イコールで区切り、1 つの式と認識している。

$$sim_{math}(p \Rightarrow q) = \frac{1}{l}\sum_{i=1}^{l} \max_{1 \leqq j \leqq k} \cos(p_{math_i}, q_{math_j}) \quad (6)$$

また同様の手順で問題 q からみた問題 q の数式部分の類似度 $sim_{formula}(q \Rightarrow p)$ を表現すると、式 7 となる。

$$sim_{math}(q \Rightarrow p) = \frac{1}{k}\sum_{j=1}^{k} \max_{1 \leqq i \leqq l} \cos(p_{math_i}, q_{math_j}) \quad (7)$$

3.4.3 数式部分の類似度算出方法

3.4.2 節で求めた値を用いて、数式部分の類似度 $sim_{text}(p, q)$ を式 8 のように示す。

$$sim_{math}(p, q) = \frac{1}{2}(sim_{math}(p \Rightarrow q) + sim_{math}(q \Rightarrow p)) \quad (8)$$

文章部分と同様に、片方向の類似度の平均をとることで双方向の類似度を表現している。

4 実　　験

本論文では 2 種類の実験を行った。1 つ目は最適なパラメータの調査であり、式 1 におけるパラメータ α の最適値を算出した。2 つ目はベースラインと提案手法の精度比較である。

4.1 実験概要

4.1.1 評価方法

本研究では青チャート数学 IA [8]・IIB [9] を使用し、事前学習や実験を行った。この青チャートに含まれている問題のうち、入試問題が記載されている EXERCISES から 85 題を抽出し、それに対応する例題とのペアをテストデータとした。対応する例題に関しては、青チャート内で明記されているものに準じている。この EXERCISES に記載されている任意の問題の入力に対し、全 85 題の例題との問題間の類似度 (式 1) を算出し、順位付けを行った。その際に、上位 n 位内に対応する例題が検出しているかを示す $HitRate@n$ で評価をした。$HitRate$ の算出方法に関しては、式 9 に示す。

$$HitRate@n = \frac{\text{上位 } n \text{ 位以内に対応する例題が含まれていた数}}{\text{全例題数 (85 題)}} \quad (9)$$

この $HitRate@1$ と $HitRate@5$ を用いて、提案手法とベースラインの精度比較を行った。なお、提案手法の類似度算出には式 1 を使用し、パラメータ α は実験で求めた最適なパラメータを適用する。

4.1.2 ベースライン

ベースラインとして用いたのは MathBERT、TF-IDF、Simpson 係数の 3 種類である。

a） MathBERT

MathBERT に関しては、事前学習をしないそのままのモデルを使用した。また、日本語の入力に対応していないため、数式のみの入力とした上で式 8 と同じ算出方法で類似度を算出した。

b） TF-IDF

テストデータの青チャートコーパス内の問題 q に含まれる単語 w の出現回数を $n_{q,w}$ とし、問題 q 内に含まれる全単語の出現回数の総和を N_q としたとき、単語 w の問題 q における出現回数 $TF(w,q)$ は式 10 と表すことができる。

$$TF(w,q) = \frac{n_{q,w}}{\sum_{i=1}^{N_q} n_{q,w_i}} \tag{10}$$

また、青チャートコーパス内の全問題数を $|C|$ とし、単語 w が出現した問題数を $DF(w,C)$ としたときに、青チャートコーパス内の単語 w の逆文書頻度 $IDF(w,C)$ は式 11 と表すことができる。

$$IDF(w,C) = \log \frac{|C|}{DF(w,C)} + 1 \tag{11}$$

以上の式 10、11 を用いて $TF-IDF$ は式 12 と表される。

$$TF-IDF(w,q,C) = TF(w,q) \times IDF(w,C) \tag{12}$$

c） Simpson 係数

3.3.3 節で算出した Simpson 係数のみを用いて、問題の類似度算出を行った。そのため、数式部分に関しては一切考慮されていない類似度となっている。

4.2 最適なパラメータ

4.2.1 結　　果

式 1 に含まれる α について最適な値を設定するために、$0 \leqq \alpha \leqq 1$ において α を 0.01 ずつ増加させて $HitRate$@1、$HitRate$@5 の推移を可視化させたものが図 3 である。

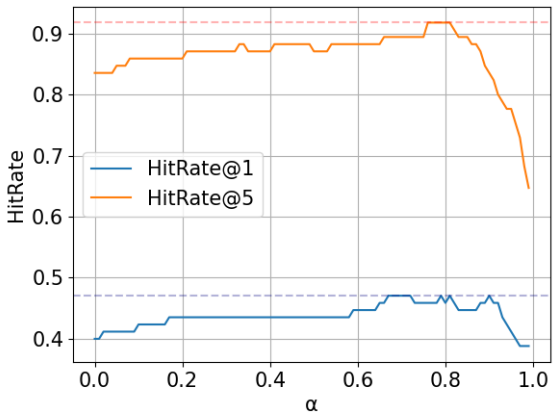

図 3　各パラメータにおける HitRate の推移

図 3 より、$HitRate$@1、$HitRate$@5 ともに $\alpha = 0.79$ が最適なパラメータであるといえる。

4.2.2 考　　察

α の値は増加させることで $HitRate$ が向上するものの、$\alpha = 0.8$ を超えると低下することがうかがえた。特に $HitRate$@5 の精度低下が著しいことから、数学の問題の類題を考えるにあたって具体的な解法だけでなく、文章部分といった表面的な類似性も一定数の影響を与えていると考えられる。

4.3 ベースラインとの比較

4.3.1 結　　果

4.1.2 節で述べた 3 種類のベースラインと提案手法の $HitRate$ の比較を行った。なお、提案手法においては、最適なパラメータである $\alpha = 0.79$ とした上で式 1 の類似度を算出している。

表 3　HitRate@n の比較

	HitRate@1	HitRate@5
MathBERT	0.47	0.74
tf-idf	0.39	0.73
Simpson 係数	0.33	0.39
提案手法 ($\alpha = 0.79$)	0.47	0.92

$HitRate$@1 と $HitRate$@5 ともに提案手法の精度が最もよいという結果となった。特に $HitRate$@5 の精度が 9 割を超えており、十分な精度であるといえる。

4.3.2 考　　察

提案手法の $HitRate$@5 が著しく良かったことに対して、問題のおおまかな性質を捉えるために文章部分と数式部分のそれぞれの特徴を反映することは重要であると考えられる。また、ベースラインである MathBERT と $HitRate$@1 では同等の精度だったものの、$HitRate$@5 においては精度が大きく上昇していることから提案手法の有用性がうかがえる。また、Simpson 係数のみでは $HitRate$ が低いことについては、ある問題に含まれる数学用語が限られており、同じ Simpson 係数を取る問題ペアが非常に多いことが原因であると考えられる。

4.4 単元別の最適なパラメータ

4.2 節では、どの問題でも対応できる汎用性の高いパラメータを求めたが、ここでは提案手法の特徴を捉えるために単元別の最適パラメータについて記す。具体的には、式 1 内の α の値を 0.05 ずつ増加させ、最適なパラメータを単元ごとに算出した。そして、その単元別の最適なパラメータをまとめたものが表 4 である。なお、ここでの単元とは高等学校数学の指導要領 [10] を元に設定した。

表 4 より、数学と人間の活動や複素数の方程式といった式が他と異なる性質持つ単元は特にパラメータの値が大きくなっている。そのため、このような単元は類似度を考える場合に数式部分が重要な意味を持つと考えられる。しかしながら、式変形が多いと考えられる数と式や二次関数の単元はパラメータの値が小さい方が $HitRate$@1 が増加した。この単元に登場する数式は他の単元でも登場する基本的な概念であるため、本来の類題を検出するためには文書部分を考慮する必要性があると考えられる。また、図形と計量やデータの分析といった、数式以外の別の部分に情報がある問題においてもパラメータの値が小さくなっていると考えられる。

ここで、表 4 で求めた各単元の最適なパラメータを使用した場合と、全体を対象とした最適なパラメータ $\alpha = 0.79$ を使用した場合における $HitRate$@1 の差をまとめたものが図 4 である。図 4 には、差が 0 となった単元は記載していない。なお、ある単元 u の $HitRate_u$ の差は以下の式 13 と定義する。ここ

表 4　単元別の最適なパラメータ α

	HR@1 が最大	HR@5 が最大	HR@1 と HR@5 の平均が最大
数と式	0.01	0.79	0.01
集合と命題	0.79	0.79	0.79
2 次関数	0.01	0.79	0.79
図形と計量	0.05	0.79	0.79
データの分析	0.03	0.79	0.79
場合の数	0.79	0.20	0.20
確率	0.79	0.79	0.79
図形の性質	0.79	0.79	0.79
数学と人間の活動	0.90	0.79	0.90
式と証明	0.79	0.79	0.79
複素数の方程式	0.90	0.79	0.90
図形と方程式	0.83	0.79	0.83
三角関数	0.79	0.79	0.79
指数関数と対数関数	0.79	0.79	0.79
微分法	0.79	0.79	0.79
積分法	0.79	0.79	0.79
数列	0.79	0.79	0.79
統計的な推測	0.79	0.79	0.79

※ HR は HitRate の略

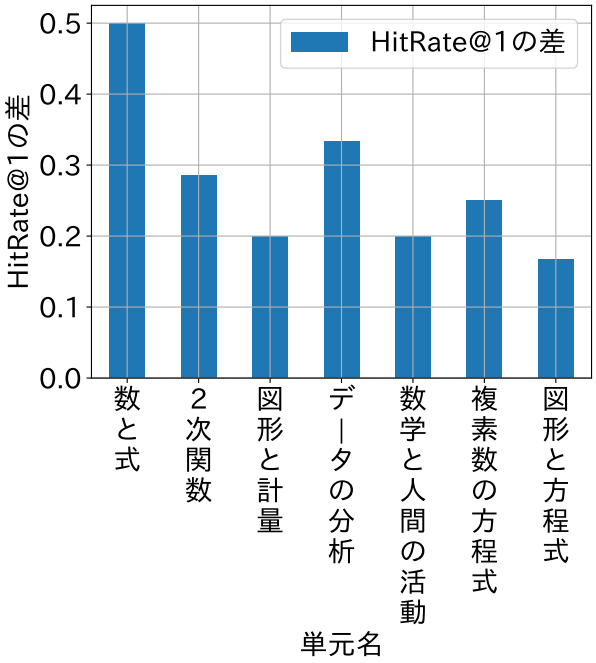

図 4　$\alpha = 0.79$ と各単元の最適パラメータによる HitRate@1 の差

での α_u は単元 u における最適なパラメータ、α^* は全体の最適なパラメータである 0.79 を示す。

$$HitRate_u\text{の差} = (\alpha_u\text{での } HitRate_u) - (\alpha^*\text{での } HitRate_u) \tag{13}$$

図 4 より、各単元において最適なパラメータを用いることで $HitRate@1$ がさらに向上した。$HitRate@5$ に関しては、場合の数のみが 0.3 増加する結果となった。これらのことより、単元別に最適なパラメータを使用して類似度を求めることによって、同じ単元でもより似ている問題を判定できると考えられる。一方、$HitRate@5$ がそこまで改善しなかったため、問題の大まかな特徴を反映するには $\alpha = 0.79$ で十分対応可能であるといえる。しかしながら、唯一 $HitRate@5$ が上昇した場合の数においては、場合の数の単元における最適なパラメータの値も非常に低く、別の特徴を持っているといえる。この理由としては、数式部分が確率などの分野と似てしまっており、文章部分のキーワードの違いが判断材料になっていると考えられる。

4.5　提案手法で検出できなかった問題

提案手法を用いて対応する類題が検出できなかった問題を表 5 に示す。

表 5　検出できなかった問題一覧

問題番号	類題の順位	sim_{math}	sim_{text}
0	18	0.75	0
1	50	0.48	0
24	60	0.16	0.45
26	47	0.58	0
27	32	0.58	0.42
39	16	0.51	0.43
47	14	0.79	0

表 5 の中でも文章部分の類似度が極端に低いため全体の類似度が低くなっている問題とそうでない問題の 2 種類が存在することがうかがえる。そこで、文章部分の類似度が 0 である問題の内容をまとめたものが表 6 である。

表 6　$sim_{text} = 0$ である問題の概要

問題番号	問題の概要
0	式を展開した際の各係数を求める問題
1	式を展開して値を求める問題
26	アルファベットの辞書順の並べ方の問題
47	恒等式の問題

文章部分の類似度が 0 である全ての問題において、数学の集合間類似度が 0 となっていることが原因であった。問題番号 0, 1 の問題は数と式の問題で非常に基礎的な例題と対応している問題のため、問題文にほとんど日本語が含まれておらず数学用語自体が検出されなかった。問題番号 47 に関しても、式と証明の単元であるため式変形が多く、文章部分から数学用語を抽出することができていなかった。対して、問題番号 26 の問題は文章は少なくないものの、キーワードとなる順列などといった数学用語が解説には直接登場しなかったことが原因であった。これらに対して、数学の問題に出現するやや汎用的な用語も数学用語集合に追加することで改善が期待できる。

文章部分の類似度が 0 ではなかった問題の内容をまとめたものが表 7 である。また、全問題ペアにおいて文章部分の類似度

表 7　$sim_{text} \neq 0$ である問題の概要

問題番号	問題の概要
24	整数集合の問題
32	硬貨を投げた結果によって座標が定まる確率の問題
16	図を参照して角度を求める問題

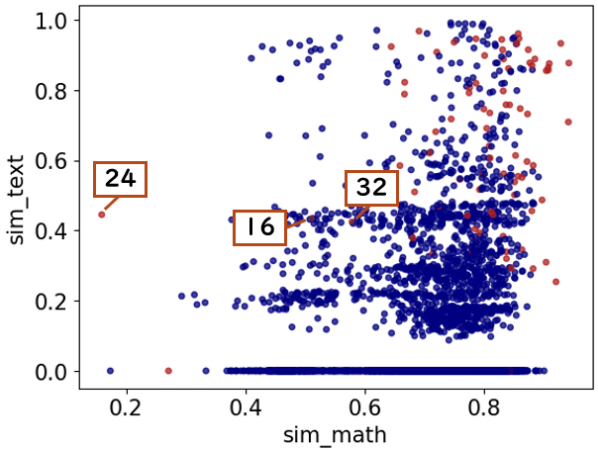

図 5　全問題ペアの類似度の分布

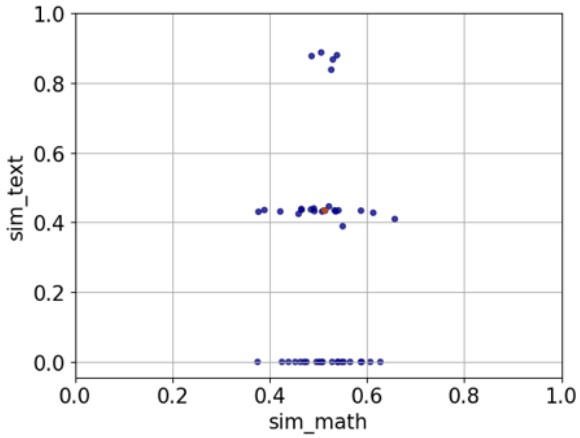

図 6　問題番号 16 と例題 85 題の散布図

と数式部分の類似度の分布を示したものが図 5 である。なお、実際の類題ペアである 85 ペアを赤色、その他の非類題ペアを青色で示している。そして、文章部分の類似度が 0 ではなかった問題 3 題の正解ペアに対してのみ問題番号を散布図上に記載している。

表 7、図 5 より整数集合の問題である 24 番は数式部分の類似度が極端に低いため、例題として検出できていないといえる。この問題に登場する表現は、積集合や和集合を多用するものであり、入力の一文が過度に長くなっていた。そのため、似た表現であっても数式部分の類似度が低くなっている可能性が考えられる。

一方、16 番と 32 番の問題は全体の中でどちらかの類似度が極端に低い値を取っているわけではないため、それぞれ個別に分布を調査したものが図 6、7 である。

図 6 より、数式部分の文章部分の類似度は 0.4~0.6 の範囲に多く分布している。これは、この問題が図を参照して問題を解

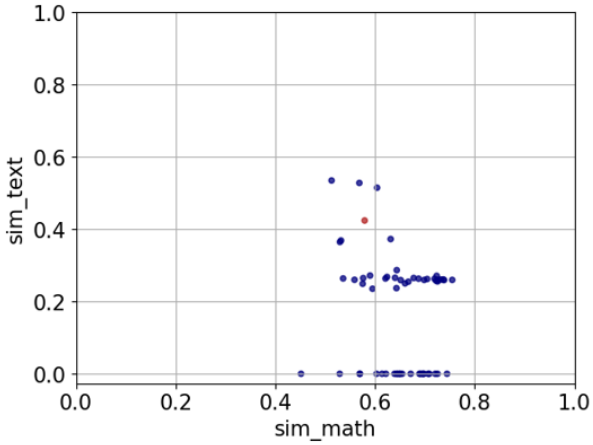

図 7　問題番号 32 と例題 85 題の散布図

くため、その図に含まれる情報量を上手く反映できていない可能性が考えられる。

図 7 からは、数式部分の類似度が全体的に高い傾向にあることがうかがえる。その理由としては、数式部分は汎用的な確率の計算が中心であるため、多くの問題と類似度が高くでている可能性が考えられる。しかしながら、文章部分については、特徴的な表現が多いことにより値が大きくなっているため、パラメータを上手く調節することによって正しく検知できると推測される。

5　おわりに

本論文では高校数学の問題に含まれる文章と数式それぞれに対して教師なし SimCSE を行った。そして、その事前学習済みモデルを使用し、問題に含まれる各文と式のベクトル化をして求めたコサイン類似度を問題間の類似度とした。この際、文章部分の類似度に関しては、数学用語の Simpson 係数を用いることでベクトルベースの類似度と比較して数学の問題の性質をより反映させた。そして、この文章部分の類似度と数式部分の類似度にパラメータ α を掛け合わせたものを問題間の類似度とした。実験より、このパラメータの最適値は $\alpha = 0.79$ となった。この最適なパラメータを適用した上で提案手法の類似度算出方法を使用してある問題の順位づけを行うと、上位 5 件内に該当する類題が入っているかのタスクにおいてベースラインよりも大幅に精度が向上した。また、単元別にパラメータの最適値を求めたところ、この上位 5 件内に該当する類題が入っているかのタスクの精度がさらに向上した。

謝　　辞

本研究は JSPS 科学研究費助成事業 24K15195 による助成を受けたものです。ここに記して謝意を表します。

文　　献

[1]　上江洲 弘明, 谷口 哲也, 高井 勇輝, 西岡 圭太, 中川 勇人, 数学問題の類似性評価, 第 38 回ファジィシステムシンポジウム 講演論文集 (FSS2022 オンライン), 2022

[2] Shuai Peng, Ke Yuan, Liangcai Gao, Zhi Tang, MathBERT: A Pre-Trained Model for Mathematical Formula Understanding, arXiv:2105.00377v1 [cs.CL], 2021
[3] Tianyu Gao, Xingcheng Yao, Danqi Chen, SimCSE: Simple Contrastive Learning of Sentence Embeddings, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp.6894–6910, 2021
[4] Kamal Raj Kanakarajan, Bhuvana Kundumani, Abhijith Neil Abraham Malaikannan Sankarasubbu, BioSimCSE: BioMedical Sentence Embeddings using Contrastive learning, https://aclanthology.org/2022.louhi-1.10/, 2022
[5] Mpho Mokoatle, Vukosi Marivate, Darlington Mapiye, Riana Bornman and Vanessa. M. Hayes, A review and comparative study of cancer detection using machine learning: SBERT and SimCSE application, https://doi.org/10.1186/s12859-023-05235-x, 2023
[6] Hayato Tsukagoshi, Ryohei Sasano, Koichi Takeda, Japanese SimCSE Technical Report, arXiv:2310.19349v1, 2023
[7] https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens
[8] チャート研究所, 新課程 チャート式 基礎からの数学 I+A, 2022
[9] チャート研究所, 新課程 チャート式 基礎からの数学 II+B, 2022
[10] 文部科学省, 高等学校学習指導要領（平成 30 年告示）解説 数学編 理数編, 2023, https://www.mext.go.jp/content/20230217-mxt_kyoiku02-100002620_05.pdf