Transformer モデルによる自由記述文からの成績予測

EP20050 小池正基　　指導教授：山下隆義

1. はじめに

教育環境のデジタル化に伴い，デジタル教材に対する学習者の操作ログから成績予測することで，教育をサポートする取り組みが始まっている．ログデータを収集するには，システム構築に多大なコストが必要であることに加え，成績に関連しない不要な操作などのノイズが多く含まれており，データの前処理を必要とする．そこで本研究では，記録するためのシステムが簡潔で記録のコストが低いアンケートに注目する．アンケートの回答文には生徒の理解度に繋がるキーワードが文脈に含まれるため，成績予測手法に有効であると考える．提案手法では，Term Frequency–Inverse Document Frequency (TF-IDF) を用いて回答文を分析し，重要単語を成績予測に活用することで，成績予測精度の向上を図る．

2. 講義アンケートを利用した従来の学習支援手法

Tamura らは，講義受講後の自由記述感想文や課題の合否結果を記録したログデータから，BERT モデルを用いて生徒の学習意欲の予測を行っている [1] ．しかし，本研究はアンケート文から成績を予測するものではなく，アンケートを用いた成績予測の有効性は明らかでない．

3. 提案手法

本研究では，講義後アンケートから成績を予測する手法を提案する．TF-IDF により推定した重要単語を基に，自然言語モデル RoBERTa [2] の Attention weight を修正し，単語の傾向を考慮した成績予測モデルを構築する．

3.1. TF-IDF による成績別重要単語の推定

回答文における成績ごとの特徴的な単語を強調するために，TF-IDF を用いて回答文内の単語の重要度を評価する．TF-IDF は，文章に含まれる各単語の出現回数を用いて，単語の重要度を TF-IDF スコアとして表す尺度である．全文章に含まれる成績 i の文章集合 G_i に対して，各文章 $g \in G_i$ における各単語 t の TF-IDF スコア S は式 (1)，式 (2) により求める．

$$S(t, G_i) = \frac{\sum_{g \in G_i} \mathrm{TF\text{-}IDF}(t, g, G_i)}{|G_i|} \tag{1}$$

$$\mathrm{TF\text{-}IDF}(t, g, G_i) = \log(1 + c(t, g)) \cdot \log\left(\frac{|G_i|}{df(t)}\right) \tag{2}$$

ここで，$c(t, g)$ は文章 g 中の単語 t の出現回数，$|G_i|$ は成績 i の文章数，$df(t)$ は G_i で単語 t を含む文章の数である．各成績の各単語 TF-IDF スコアが他の成績のスコアの平均よりも 2 倍以上であるとき，その単語を重要単語とする．図 1 に各成績の重要単語を示す．

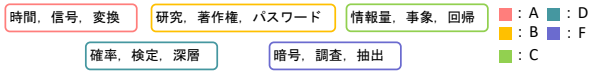


図 1: TF-IDF によって求められた重要単語

3.2. 重要単語の Attention weight の修正

RoBERTa は Transformer モデルのエンコーダ部分を用い，文章の穴埋めを行うタスク (MLM) によって事前学習した自然言語モデルである．RoBERTa のアテンション機構は，入力トークンをクエリとし，キーとの内積をとることで Attention weight を算出する．クエリと重要単語を比較し，一致する部分の Attention weight にバイアスを加算することで，アテンション機構において重要単語が強く反映されるよう修正する．入力トークン $\mathbf{Q}$ について重要単語の Attention weight A を求める式を式 (3) に示す．

$$A(\mathbf{Q}, \mathbf{K}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathrm{bias}\right)$$

$$\mathrm{bias} = \begin{cases} 0.1 & \text{if } S(\mathbf{Q}, G_i) > 2\dfrac{\sum_{j \neq i}^{i=|\mathbf{G}|-1} S(\mathbf{Q}, G_j)}{|G|-1} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

ここで，$\mathbf{K}$ はキー，d_k はモデルの次元数，$|G|$ は G の成績の数，G_j は i 以外の成績 j の文章群，bias は Attention weight に加算するバイアスを示す．

4. 評価実験

本実験では，成績予測精度を比較することで，提案手法の有効性を調査する．

4.1. 実験条件

本実験では，九州大学の情報科目の講義で実施したアンケートの回答文と，回答者の成績が与えられたデータセットを学習する．アンケートは 15 回にわたり，709 名の生徒に対して 5 つの質問を行い，回答文を収集する．アンケートの質問内容を表 1 に示す．回答が無い場合を除き，得られたデータ数は合計 28,669 であった．回答文を事前学習済み RoBERTa モデルに入力し，得られた特徴ベクトルから MLP によって 5 段階の成績（A, B, C, D, F）に分類する．全体の 8 割の生徒 (552 人) の回答を学習データ，2 割の生徒 (139 人) の回答を評価データとして使用する．

表 1: 質問内容と未回答率

質問	質問内容	未回答率
Q1	講義を自分なりの言葉で説明してください	22.7%
Q2	講義で分かったことを書いてください	24.0%
Q3	講義で分からないことを書いてください	35.7%
Q4	質問があれば書いてください	64.2%
Q5	講義の感想や反省を書いてください	25.6%

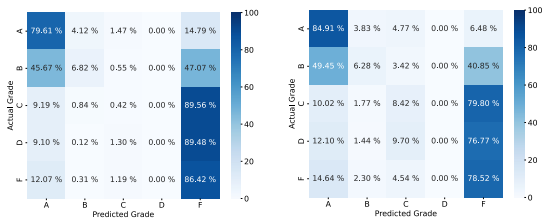
4.2. 実験結果

表 2 に既存の RoBERTa モデル (Baseline) と RoBERTa に提案手法を導入した場合 (Ours) の予測精度を示す．表 2 より，Attention weight を修正することで，Accuracy が 1.1pt 向上している．しかし，F1-score は Baseline の方が高いため，一部の成績で精度が向上していると考えられる．

図 2 に各成績の予測精度を混同行列で示す．図 2 より，Ours では，成績 A，C の精度がそれぞれ 5.3pt，8.0pt 向上し，成績 F の精度が 7.9pt 減少している．一方で，モデルが成績 F と誤分類した結果を比較すると，Ours では誤分類が低下している．この結果から，偏った予測を行うモデルにおいて，提案手法の有効性を示すことができた．

表 2: 予測精度 [%] の比較

	Baseline	Ours
Accuracy	44.1	**45.2**
F1-score	**39.6**	38.1



(a) Baseline　　(b) Ours

図 2: 混同行列による比較

5. おわりに

本研究では，データ収集が容易である自由記述アンケートをデータセットに用いた成績予測モデルの有効性を評価した．また，TF-IDF を用いて成績ごとの重要単語を算出し，Attention weight を重要単語を基準に修正する手法を提案した．その結果，提案手法の Accuracy は提案手法なしと比較して 1.1pt 上昇したが，F1-score は 1.5pt 低下した．今後は，大規模言語モデルによる成績予測，合否判定を予測するモデルの作成，成績の判断根拠の可視化を試みる．

参考文献

[1] M. Tamura, *et al.*, “オンライン演習下における自由記述感想文からの学習意欲の推論モデル”, 情報教育シンポジウム論文集, 2020.

[2] Y. Liu, *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, arXiv: 1907.11692, 2019.