様式第 6 号

令和 5 年度実績報告書

令和　6年　3月　21日

公立千歳科学技術大学
学長　宮永　喜一　様

公立千歳科学技術大学特別研究等助成要綱第 7 条に基づき、下記のとおり報告いたします。

報告者	所属	共通教育科	職名	助手
	氏名	上野　春毅	ふりがな	うえの　はるき
研究課題名	深層学習技術を活用した演習問題の難易度を推定するシステムの開発			
本研究費による発表論文、著書　など	次年度以降に予定			

研究成果報告

高等教育において知識の定着や活用を図るために、個々人に適した難易度の演習問題に取り組めることが重要である。しかし、これまでに大学等の教育機関で利活用が進んでいる Learning Management System（LMS）を通じて提供される演習問題には難易度が暗黙的に設定されているものの明示的に付与されていないことがある。これらの暗黙的に設定された難易度をあきらかにして学修者に提示することで、既存の演習問題をより有効に活用できる可能性がある。この分野の研究では、難易度推定のアプローチはおおきくは 2 つに分けられる。１つは正答・誤答の回答履歴から難易度を算出する方法である。もう一つは、文章の平均文長や単語の難易度などの特徴から回帰式を用いて推定する方法である。前者は、項目反応理論をベースに解答履歴から難易度を推定するのが一般的であるが、当理論を適応するには統制された環境において大規模な数の被験者と解答データ数が要求されるために一部の専門的な機関での実施に留まっており一般的な教育機関での実施は容易ではない。後者は、回答履歴を用いずに問題文の文章から難易度を測っている。しかし、学年単位などの難易度というスケールが大きいものが多い。授業への適用を考えると、学習単元を着実に身につけるために単元内の教材に難易度が付与されていることが求められる。本研究では、学習単元を段階的に学ぶ教材を想定して、解答履歴ではなく問題の文章から難易度を推定できるモデルの構築を試みる。

本研究ではこれまでに Python の入門内容の教材を対象に深層学習による予測モデルを構築して検証してきた。専門家が難易度を判別する方法としてドメイン知識における重要な単語や文の構成（文脈）を参照していると仮定してこの特徴量を扱えるモデルを選定した。予測モデルには Bidirectional Encoder Representations from Transformers（BERT）とした。BERT は Transformer と呼ばれるアーキテクチャが使われており、Attention 機構により単語とその文脈を考慮して予測を可能としている。検証結果からデータセットのテストデータから難易度を７割程度当てることができた。一定程度の推定ができる可能性が示唆された。外れたデータからその特性を調べると、７段階の難易度を大きくはずすことがなく、例えば最低難易度を最高難易度ということがなく、予測範囲はおおよそ設定されるべき難易度の近辺にあった。この結果から深層学習を用いて単語や文脈を考慮して推定することが有効な可能性があると考えられる。

本報告では、Attention 機構を有する深層型ニューラルネットワークの可能性をさらに探るために、言語処理の観点でパラメータ数の増加がより問題文章を解釈できるという仮定の下に、大規模言語モデルの適応を図る。具体的には、近年急速に発展する大規模言語モデル（LLM; Large Language Model）である OpenAI の ChatGPT を用いる。ChatGPT は BERT 同様に Transformer の Attention 機構を有しており、より膨大なデータセットとパラメータから構成される。生成系 AI と呼ばれており、入力を工夫することで任意の出力が得られる。入力のプロンプトには、Zero-Shot や Few-Shot などの方法によって入力を工夫することで難易度を推測させることとした。プロンプトの構成を示す。まず、役割を設定した。学習単元を段階的に学ぶ教材を想定しているので、その想定を具体的に記述した。次に、難易度の考え方を設定した。専門家があらかじめ設定した難易度の構成を記述した。さらに一部の問題を例示した。そして、演習問題の問題文を設定して、難易度を出力する指示を設定した。今回の検証で用意した演習問題の数は 67 件である。演習問題は本学の CIST-Solomon の CBT の教材として、難易度は 7 段階とした。これを GPT-3.5-Turbo モデルで予測させて検証した。検証結果（表 1）から、1～3 の低・中のレベルは一定程度の予測ができたことがわかった。しかし、4 以降の中・高の難易度をあてることができなかった。出力には理由が述べられており、理由を分析した。分析の結果から、高レベルのものであっても、簡単な問題と認識している旨が見られた。これはデータセットの特徴が現れている可能性が考えられる。つまり、GPT の訓練時に世界中のさまざまなデータを学習しており、本研究で扱う演習問題は初学者向けであるために高い難易度とは判定されなかったという可能性が考えられる。これに対して、プロンプトで例示する数を増やすか（Few-Shot のアプローチ）、ファインチューニングで扱う領域に調整する手段が考えられる。

表 1 検証結果

正解値＼予測値	1	2	3	4	5	6	7
1	2	2	1	0	0	0	0
2	3	4	3	0	0	0	0
3	0	2	8	0	0	0	0
4	0	1	8	0	0	1	0
5	0	0	9	0	0	1	0
6	0	1	8	2	0	0	0
7	0	0	0	1	1	0	0