Integrating LSTM and BERT for Long-Sequence Data Analysis in Intelligent Tutoring Systems

Preprint · April 2024

DOI: 10.13140/RG.2.2.13638.08007

5 authors, including:

Zhaoxing Li
University of Southampton
25 PUBLICATIONS **94** CITATIONS

SEE PROFILE

Jujie Yang
Durham University
4 PUBLICATIONS **4** CITATIONS

SEE PROFILE

Jindi Wang
Durham University
18 PUBLICATIONS **56** CITATIONS

SEE PROFILE

Lei Shi
Newcastle University
125 PUBLICATIONS **1,518** CITATIONS

SEE PROFILE

Integrating LSTM and BERT for Long-Sequence Data Analysis in Intelligent Tutoring Systems

Zhaoxing Li[1][0000−0003−3560−3461], Jujie Yang[2][0000−0001−6024−2720], Jindi Wang[2][0000−0002−0901−8587], Lei Shi[3][0000−0001−7119−3207], and Sebastian Stein[1][0000−0003−2858−8857]

[1] School of Electronics and Computer Science, University of Southampton, Southampton, UK
[2] Department of Computer Science, Durham University, Durham, UK
[3] Open Lab, School of Computing, Newcastle University, Newcastle upon Tyne, UK
zhaoxing.li@soton.ac.uk,{jujie.yang, jindi.wang}@durham.ac.uk,
lei.shi@ncl.ac.uk, ss2@ecs.soton.ac.uk

Abstract. The field of Knowledge Tracing aims to understand how students learn and master knowledge over time by analyzing their historical behaviour data. To achieve this goal, many researchers have proposed Knowledge Tracing models that use data from Intelligent Tutoring Systems to predict students' subsequent actions. However, with the development of Intelligent Tutoring Systems, large-scale datasets containing long-sequence data began to emerge. Recent deep learning based Knowledge Tracing models face obstacles such as low efficiency, low accuracy, and low interpretability when dealing with large-scale datasets containing long-sequence data. To address these issues and promote the sustainable development of Intelligent Tutoring Systems, we propose a **L**STM **B**ERT-based **K**nowledge **T**racing model for long sequence data processing, namely **LBKT**, which uses a BERT-based architecture with a Rasch model-based embeddings block to deal with different difficulty levels information and an LSTM block to process the sequential characteristic in students' actions. LBKT achieves the best performance on most benchmark datasets on the metrics of ACC and AUC. Additionally, an ablation study is conducted to analyse the impact of each component of LBKT's overall performance. Moreover, we used t-SNE as the visualisation tool to demonstrate the model's embedding strategy. The results indicate that LBKT is faster, more interpretable, and has a lower memory cost than the traditional deep learning based Knowledge Tracing methods.

Keywords: Knowledge Tracing · BERT · Student Modelling · Long-Sequence Data Processing · Technology Enhanced Learning (TEL)

1 Introduction

Technology Enhanced Learning (TEL) has become increasingly important in providing high-quality education to build a more sustainable world. The recent COVID-19 pandemic has significantly impacted traditional classroom education and sparked online learning, enabling teaching and learning remotely. Meanwhile, the development of online learning systems has made it possible to use Intelligent Tutoring Systems (ITS) to store and analyse a sizable amount of student behaviour data to improve intelligent educational services. As one of the widely applied TEL technologies, Knowledge Tracing (KT) has drawn a lot of attention. KT is the field of modelling students' learning trajectories and predicting their sequential actions based on historical interaction data between students and ITS [2].

With the development of ITS, large-scale datasets such as *EdNet* [5] and *Junyi Academy* [4] began to emerge. In these datasets, long-sequence student interaction data were gathered as an increasing number of students used the ITS for an extended period. The long- and short-sequence data in these datasets are unbalanced, which satisfies the long-tail distribution [24]. For instance, within the EdNet dataset, a substantial amount of student action sequences are included, ranging from the shortest sequence that may comprise just a single action to the longest sequence that encompasses 40,157 actions. Notably, the average action sequence length of the EdNet dataset is 121.5, indicating a moderate length of data sequences overall. However, it is important to note that the distribution of sequence lengths is highly skewed, and this unbalanced distribution has an impact on the overall performance of the KT models [12]. Although the quantity of short-sequence data is larger than the long-sequence data, the latter is of more weight than the former in prediction tasks [24].

In general, KT models could be divided into three categories: probabilistic KT models, logistic KT models, and deep learning based KT methods (DKT) [23]. Traditional probabilistic KT models and logistic KT models are forced to confront difficulties such as decreased processing efficiency and increased memory usage as growing amounts of longer sequence data are released. Deep learning based KT models are known to suffer from inefficiencies when processing long-sequence action data problems, including issues related to the accuracy, speed, and memory usage [12, 24]. Therefore, allowing the processing of very long sequence data is key to achieving high performance for next-generation KT models [12]. Moreover, due to the black-box nature of traditional deep learning methods, the current deep learning based KT models also struggle with the lack of interpretability [11].

To address the above issues, in this paper, we propose LBKT, a novel **L**STM **B**ERT **K**nowledge **T**racing model, for processing long sequence data. The model combines the strength of the Bidirectional Encoder Representations from Transformers (BERT) model in capturing the relations of complex data [8] with the strength of the LSTM model in handling long sequential data to improve its performance on large-scale datasets containing long-sequence data (here, the long-sequence data indicates a length longer than 400 interactions). Moreover, we utilise a Rasch model-based embedding method to process the difficulty level information in the historical behaviour data of students. The Rasch model is a classic yet powerful model in psychometrics [30], which could be utilised to construct raw questions and knowledge embeddings for KT tasks [11]. Rasch model based embedding could improve the model's performance and interpretability. The experimental results show that our proposed LBKT outperforms the baseline models in five datasets on metrics ACC and AUC. Moreover, it is faster at processing long-sequence data at two long-sequence datasets we extract from the two large-scale datasets. Furthermore, we use t-SNE as the visualisation tool to demonstrate the interpretability of the embedding strategy.

The main contributions of our paper lie in the following two aspects:

1. We propose LBKT [4], a novel **L**STM **B**ERT **K**nowledge **T**racing model for long sequence data processing. The LBKT leverages the power of BERT, Rasch-based embedding strategies, and LSTM.
2. The experimental results show that LBKT outperforms the baseline models on five ITS datasets on the metric of AUC(assist12, assist17, algebra06, EdNet, and Junyi Academy). Another comparative experiments show the effectiveness of LBKT when processing long-sequence datasets. LBKT model exhibits better interpretability than traditional deep learning based KT models and has advantages in training efficiency.

2 Related Work

2.1 Knowledge Tracing

Knowledge Tracing (KT) is used in Intelligent Tutoring Systems (ITS) to model and predict a student's mastery level of a specific skill or concept over time [1]. It is based on the assumption that a student's knowledge state is a hidden variable that can be inferred from their observable behaviour, such as their responses to questions or tasks related to the skill or concept being measured [7]. Its goal is to provide personalised feedback and support to students by tracking their progress and adapting instruction to meet their individual needs. This can help to improve student learning outcomes and enhance educational effectiveness. Broadly, there are three categories of KT methods: probabilistic KT models, logistic KT models, and deep learning-based KT models [23].

Probabilistic KT models assume that the student's learning process follows a Markov Process, where students' knowledge mastery could be measured by their observed learning performance [7]. Bayesian KT, or BKT, is the earliest and most classic probabilistic model, which was inspired by cognitive mastery learning [6]. BKT models generally use a probabilistic graphical model, such as Hidden Markov Model (HMM)[7] and Bayesian Belief Network [37], to track students' changing learning states. The major shortcoming of BKT is that it assumes a simplistic two-state student modelling framework, where a student's knowledge is either learned or unlearned, and there is no concept of forgetting or decay in the model. However, in reality, a student's knowledge could be complex and multi-faceted and could change over time due to various factors such as decay and interference. Therefore, BKT may not be able to capture the nuances

[4] Source code and datasets are available at https://github.com/******/LBKT

of student learning and may not provide an accurate representation of their knowledge state over time. For example, BKT assumes that each question only required one skill and that the various skills were irrelevant to each other [7, 43]. Therefore, in general, BKT models cannot process complicated problems, including the multiple skills and the complex relationship among the concepts, questions, and skills. To address this limitation, Käser *et al.* proposed Dynamic BKT, or DBKT, based on Dynamic Bayesian Network (DBN), to model the prerequisite hierarchies and dependencies of multiple skills [15]. However, both BKT and DBKT still struggle with processing multiple topics or skills, failing to account for contextual factors that may impact student learning.

The logistic KT models are built on the principle of logistic regression, which is a statistical method used to model the probability of a binary outcome based on one or more predictor variables [23]. In the context of educational data, the predictor variables could include a student's prior performance on a set of related skills or concepts, their response time, and their correctness or incorrectness in answering assessment questions. The output of the logistic regression KT model is a probability estimate of a student's mastery level on a particular skill or concept, which can be used to inform personalized learning interventions and improve student outcomes. There are three logistic models. The Learning Factor Analysis model (LFA) incorporates the initial knowledge state, easiness of knowledge components (KCs), and learning rate of KCs to estimate the student's initial knowledge state, the easiness of different KCs, and the learning rate of KCs [3]. The Performance Factor Analysis (PFA) model is an extension of the LFA model and takes into account the student's performance. PFA considers parameters for previous failures (f) and successes (s) for the KC, in addition to the easiness of KCs [28]. The Knowledge Tracing Machines (KTM) model uses factorization machines (FMs) to extend logistic models to higher dimensions [15].

Inspired by the recent success of deep learning (DL)[16], researchers have applied deep learning technologies into the KT field to develop DL-based Knowledge Tracing [29]. DL-based KT typically models a knowledge tracing task as a sequence prediction problem. With the self-attention architectures applied in the deep learning field, KT models based on the self-attention mechanism began to emerge. For example, SAKT [26] and SAINT+ [31] apply the self-attention mechanism to KT models and achieve higher performance than the traditional DL-based methods. With the development of the self-attention mechanism, Transformer based knowledge tracing models also have been proposed. Ghosh[11] proposes context-aware attentive knowledge tracing (AKT), which introduces a novel monotonic attention mechanism that accounts for the temporal nature of the learning process and the decay of students' knowledge. Nakagawa *et al.* proposed the Graph-based Knowledge Tracing (GKT) model, which incorporates the potential graph structure of KCs into a graph [25]. There were also KT methods based on BERT that had been proposed. MonacoBERT [17] is a BERT-based KT model that incorporates the monotonic convolutional multi-head attention and classical test-theory-based (CTT-based) embedding strategy to improve performance. BEKT [35] is a Bidirectional Encoder representation from the Transformers-based model that predicts student knowledge state by combining historical learning performance.

2.2 Transformer-based Model and Application

Transformer is a prominent neural network model proposed by Vaswani *et al.*, which utilises the self-attention mechanism to extract inherent features [36]. Transformer-based models have achieved significant success in the Deep Learning field, especially in Nature Language Processing (NLP) and image generation tasks [14, 27].

The evolution of Transformer-based models, such as BERT [8] and GPT[10], has achieved outstanding performance in the above tasks. BERT, first proposed by Devlin *et al.*, is a successful application of Transformer [8]. BERT utilises the self-attention mechanism and the masked language model (MLM) to train the Transformer bidirectionally in the NLP fields [8]. BERT is renowned for its exceptional ability to process and comprehend natural language text efficiently. It has consistently outperformed other deep learning models in a broad range of tasks, extending beyond the field of NLP. BERT's success can be attributed to several key features, including its bidirectional context, which allows it to capture the dependencies between both preceding and succeeding tokens in a sequence. Additionally, BERT's large pre-training corpus enables it to learn a robust language representation that can be fine-tuned for downstream tasks with relatively small amounts of labelled data. BERT's transformer architecture, which uses self-attention mechanisms to capture global dependencies between tokens in a sequence, is also a significant factor contributing to its performance. The self-attention mechanism allows BERT to weigh the importance of different tokens in a sequence dynamically, which improves its ability to capture complex patterns and relationships in the data [8]. BERT is also known for its ability to generate high-quality embeddings, which are crucial for many natural language processing tasks [8]. There have been a lot of BERT variants applied in other deep learning fields, demonstrating their outstanding performances. For example, ConvBERT [13] applies the original BERT architecture in the image processing field; BERT4Rec uses BERT model to improve recommendation systems [32]; LakhNES uses BERT model to enhance Music Generation [9]. However, in the Knowledge Tracing field, although some BERT-based models, such as BEKT [19, 20, 22, 18, 21, 42, 40, 38, 41, 39] and BiDKT [34], are proposed to improve performance, they are unable to outperform state-of-the-art KT methods in large-scale datasets containing long-sequence data.

3 Methodology

3.1 Problem Statement

The key to knowledge tracing is to predict the correctness of a student's next answer in a sequence. Let $x_1, \ldots, x_t$ denote the student's actions, and let the

t-th action be represented as $x_t = (q_t, a_t)$, where q_t is the question presented to the student and a_t is the student's response. The goal is to estimate the correctness $P(a_t = 1|x_1, \ldots, x_{t-1})$, that is, the correctness of student's response to the current question, given their previous actions in the sequence.

3.2 Proposed Model Architecture

We propose a novel model, LBKT, for the task of knowledge tracing on large-scale datasets containing long-sequence data. While previous BERT-based KT models have shown remarkable success in capturing the relations of complex data, they also have inefficiencies when dealing with long sequence student action data [35]. On the other hand, LSTM models have been proven to excel in handling long sequential data. In response to these challenges, we propose a novel KT model that combines the strengths of both the BERT and LSTM models to improve performance on large-scale datasets containing long-sequence data (where long-sequence data indicates a length longer than 400 interactions). The Rasch embedding (also known as the 1PL IRT model) is a method to represent questions and concepts in a mathematical space [30]. The embeddings are created using a vector that summarizes the variation in questions covering a concept and a scalar difficulty parameter that controls how far a question deviates from the concept it covers. The embeddings are used as raw embeddings for questions and responses, which is a way to track a learner's knowledge state. By leveraging the strengths of a BERT-based model, Rasch model-based embeddings, and long short-term memory (LSTM) unit, our proposed model architecture has the potential to effectively process and understand relationships among different features in long-sequence data, as illustrated in Fig. 1.

The first component of LBKT is the Rasch model-based embeddings proposed by Ghosh [11]. The Rasch model-based embeddings consist of difficulty level embeddings E_d and question embeddings E_q. These embeddings are multiplied and added to the BERT token embeddings and the *sin* and *cos* positional embeddings to build the final embeddings, as shown in the following equation:

$$E = E_{Rasch} + E_{BERTToken} + E_{Position} \tag{1}$$

where the Rasch model-based embeddings E_{Rasch} are defined as:

$$E_{Rasch} = E_d + E_d \times E_q \tag{2}$$

The segment embeddings, which are typically used to represent information about the segment in the BERT model, are replaced by the Rasch embeddings mentioned above in our model's architecture. Rasch model-based embeddings are able to more accurately estimate students' knowledge states, as explained earlier, making them a key contributor to the effectiveness of LBKT for knowledge tracing tasks.

The second component of LBKT is a BERT-based block, which consists of 12 Transformer blocks. Each includes a multi-head attention mechanism, a feed-forward network (FFN), and sublayer connections. The multi-head attention