END-TO-END DEEP KNOWLEDGE TRACING BY LEARNING BINARY QUESTION EMBEDDING

Hiromi Nakagawa, Kaoru Nasuno, Yusuke Iwasawa, Katsuya Uenoyama & Yutaka Matsuo
The University of Tokyo
Japan
{nakagawa,nasuno,iwasawa,uenoyame,matsuo}@weblab.t.u-tokyo.ac.jp

ABSTRACT

Recent advancements in computer-assisted learning systems have increased the research of *knowledge tracing*, which estimates student proficiency. In this context, the method called Deep Knowledge Tracing (DKT) shows remarkable performance; however, existing DKT requires human labeling of required skills to solve a question. This limits the optimization of modeling student proficiency and application to real-world data, which are often not well-organized. In this paper, we propose an end-to-end DKT model, which does not depend on any human labeling. Using two datasets, we empirically validated that the learned tags show the same or better performance on DKT and have an information-efficient structure. These results show the potential of our proposed method to enhance the applicable scope and effectiveness of DKT, which could help improve the learning experience of students in more diverse environments.

1 INTRODUCTION

Recent advancements in computer-assisted learning systems have increased the research of knowledge tracing (Corbett & Anderson (1994)), which estimates student proficiency based on their past exercise performance. Piech et al. (2015) reported that the method called Deep Knowledge Tracing (DKT), which leverages recurrent neural networks (RNN) (Williams & Zipser (1989)), performs significantly better than other methods previously proposed.

However, existing DKT models have an essential problem; they need *skill-tags* predefined by human experts that show the required skills to solve each question. In existing DKT, question-space answers are translated into tag-space answers based on a human-defined rule and input into the DKT model. Such a knowledge tracing method, which implicitly depends on the human labeling, presents several problems. One is that the skill-tag quality affects the model's performance, because if the skill-tags are not well-organized, DKT cannot model student proficiency well. Another problem is that DKT cannot be applied to data that have no skill-tags, which is often the case with real-world data.

In this paper, we propose a first end-to-end DKT model, which does not depend on human-predefined skill-tags. Regarding the translation of questions to tags as multiplication of a binary matrix, we introduce a new Q-Embedding Model, which learns the matrix to help predict student proficiency purely from the student question-answer logs only. In addition to the above extension, this paper also presents two techniques to learn a better question-embedding matrix: reconstruction regularization of question-space and tag-space and sparse regularization of the question-embedding matrix. In this study, we empirically validate that effective question-embedding is learnable using two open datasets of math exercise.

The main contributions of this work are as follows: 1)We proposed an end-to-end DKT model, which requires no human labeling. The model enables modeling student proficiency independently of human-defined skill-tag quality and performing knowledge tracing with data that have no human labeling. 2)We proposed two techniques to learn a better question-embedding matrix. Using two datasets, we showed that the techniques make it possible to perform knowledge tracing using the learned tags with the same or better performance compared to the human-predefined skill-tags.

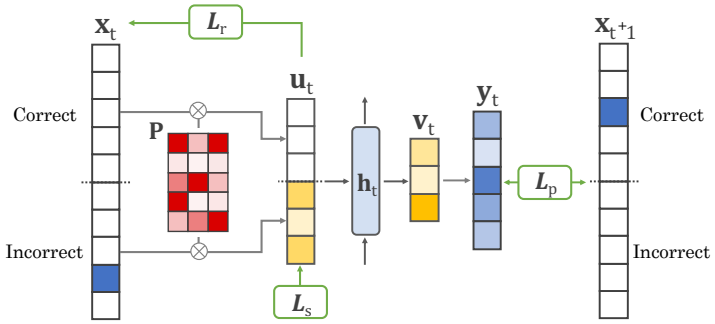


Figure 1: Architecture of Q-Embedding Model

2 Proposed Method: Q-Embedding Model

In existing DKT, a student's question-answer logs are translated into tag-answer logs based on a human-predefined rule and input into the model. For datasets with M unique questions and N unique skill-tags, this translation process can be formulated by obtaining tag-ID vector $\mathbf{s}_t$ by multiplying a binary matrix $\mathbf{P}$ with a size of $M \times N$ by question-ID one-hot vector $\mathbf{q}_t$, where $\mathbf{P}_{i,j} = 1$ if question i is associated with tag j and $\mathbf{P}_{i,j} = 0$ otherwise. Existing DKT implicitly presupposed this matrix as given; however, we learn this matrix from students' question-answer logs.

In order to learn the question-embedding matrix $\mathbf{P}$, we introduce the Q-Embedding Model. We present the architecture of the model in Figure 1. In the Q-Embedding Model, a student's question-answer logs are directly used as the model's input $\mathbf{x}_t$, and the output $\mathbf{y}_t$ is the predicted probability of the student answering each question correctly the next time. In addition, to learn the matrix that translates input question-space to low-dimensional tag-space, we add two hidden layers: $\mathbf{u}_t$ and $\mathbf{v}_t$ with a size of $2N'$ and N', respectively. Here, N' is the dimension of the tag-space and $\mathbf{P}$ is a sigmoid-activated matrix with a size of $M \times N'$. After training the model, we extract $\mathbf{P}$ and binarize it to 0 and 1 on a certain condition so that we can use it for translating question-ID to tag-ID.

We train the model on the following objective function:

$$L = \alpha L_p + \beta L_r + \gamma L_s \tag{1}$$

$$L_p = \alpha(\sum_t l(\mathbf{y}_t^T \tilde{\delta}(\mathbf{q}_{t+1}), \mathbf{a}_{t+1})), L_r = \beta(\sum_t l(\mathbf{x'}_t^T \tilde{\delta}(\mathbf{q}_t), \mathbf{a}_t)), L_s = \gamma(\sum_t (0.5 - |\mathbf{u}_t - 0.5|))$$

where α, β, and γ are arbitrary nonnegative real numbers; $\tilde{\delta}(\mathbf{q}_{t+1})$ is a one-hot encoding of which exercise is answered at time $t+1$; $\mathbf{a}_{t+1}$ is a vector of whether the exercise is answered correctly or incorrectly (1 or 0) at time $t+1$; l is the binary cross entropy; and $\mathbf{x'}_t$ is the vector reconstructed from the first half of $\mathbf{u}_t$ by the same translation as that from $\mathbf{v}_t$ to $\mathbf{y}_t$. L_p is the negative log-likelihood of the observed sequence of a student's responses under the model, which aims to learn $\mathbf{P}$ to help predict student proficiency. L_r is the reconstruction regularization of question-space and tag-space, which aims to reflect the assumption to the training that a student's response to questions is estimable from the student's understanding of each concept corresponding to tag-space. L_s is the sparse regularization, which aims to make $\mathbf{P}$ near 0 or 1 and suppress the information loss when binarizing $\mathbf{P}$ after training the model.

3 Experiments

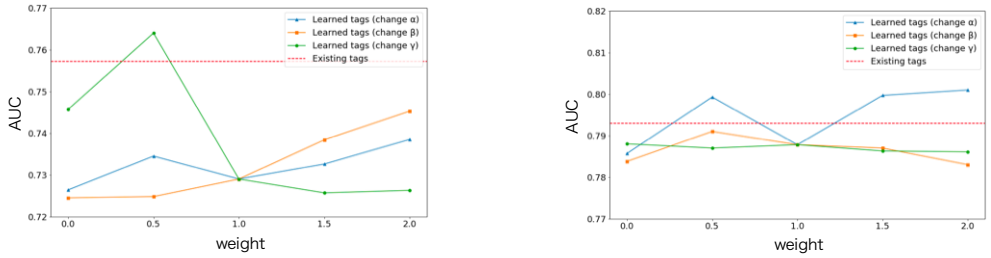
3.1 Settings

We used two open datasets of students' math exercise logs: ASSISTments 2009-2010 'skill_builder'[1] (hereinafter called 'ASSISTments') and Bridge to Algebra 2006-2007 (Stamper et al. (2010)) (here-

[1] https://sites.google.com/site/assistmentsdata/home/assistment-2009-2010-data/skill-builder-data-2009-2010

Table 1: Comparison of existing tags and learned tags

Dataset	Statistics				Results				
	Students	Questions	Skill-tags	Logs	Tags	AUC	Flow hierarchy	GRC	σ
ASSISTments	3,410	2,635	55	129,317	Existing	0.75	0.47	0.51	3140.94
					Learned	**0.76**	**0.92**	**0.93**	**1543.96**
KDDCup	1,136	3,439	192	606,819	Existing	0.79	0.72	0.70	9701.57
					Learned	**0.80**	**0.88**	**0.87**	**3674.71**



(a) ASSISTments

(b) KDDCup

Figure 2: Effect of loss function weight

inafter called 'KDDCup'). In both datasets, human-predefined skill-tags are associated with questions. The statistics of the preprocessed datasets are shown in Table 1. We unified the Q-Embedding Model's tag-space dimension with the number of the existing tags: $N' = 55$ in ASSISTments and $N' = 192$ in KDDCup. After training the Q-Embedding Model, we extracted $\mathbf{P}$ and binarized it as 0 and 1 based on equation 2, searching the threshold θ as a hyperparameter.

$$\mathbf{P'}_{i,j} = \begin{cases} 1 & \text{if } \mathbf{P}_{i,j} = \max(\mathbf{P}_i) \text{ or } \mathbf{P}_{i,j} \geq \theta \\ 0 & \text{else} \end{cases} \tag{2}$$

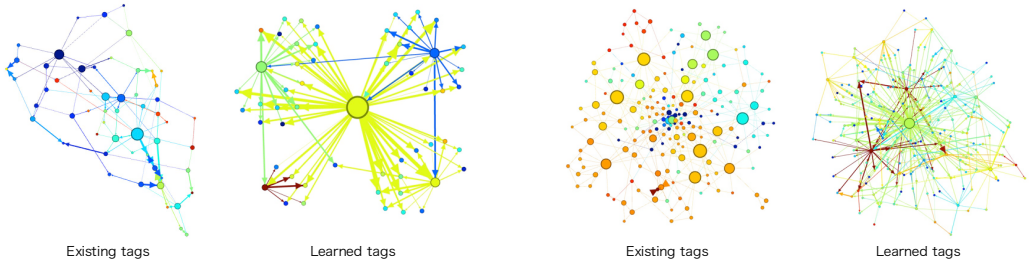
where i, j corresponds to the index of rows and columns of $\mathbf{P}$. Using this binary question-embedding matrix $\mathbf{P'}$, we translated question-space answers to tag-space answers and applied them to DKT.

3.2 Results

DKT models student interaction based on the tags associated with the questions; thus, we validated the quality of the learned tags by comparing the performance of DKT with the existing tags. We present the DKT results in Figure 2 and the best one in Table 1, showing the high score in bold for each dataset. In both datasets, the learned tags recorded the same or higher AUC score than the existing tags and this suggests that the proposed model can learn tags that are as effective for knowledge tracing as the human-defined skill-tags, without using any human labeling. In addition, we can see that α, the weight of L_p, improved prediction performance in both datasets. Although β and γ, the weights of L_r and L_s, improved performance in ASSISTments but had little effect in KDDCup, the settings where any one of the weights is 0 did not show good performance in both datasets; thus, it is considered that each loss function had a good influence on learning a question-embedding matrix to some extent.

Next, we constructed exercise influence graphs (Piech et al. (2015)) from the trained DKT model. The graph can be regarded as representing the relationships between knowledge, thus we measured its hierarchy based on flow hierarchy (Luo & Magee (2011)) and global reaching centrality (GRC) (Mones et al. (2012)).We show the graphs of both tags in Figure 3 and the comparison of flow hierarchy and GRC of each network in Table 1, showing the high index in bold for each dataset. We can see that the network of the learned tags is more hierarchical than that of the existing tags, which can be suited for representing math knowledge and has a potential to improve student learning efficiency (Block & Airasian (1971); Cohen & Hyman (1979); Abelson (2008)).

Finally, we compared the distribution of the number of times each tag appeared in the answer log, which can directly affect the learning of each unit of DKT. We present the standard deviation of the distribution as σ in Figure 1, showing the low value in bold. We can see that the distribution of the



(a) ASSISTments (b) KDDCup

Figure 3: Exercise influence graphs

learned tags is less-variant than that of the existing tags. Since the proposed method are learned with the optimization of neural networks, it seems that the information of each tag is evenly distributed so that the neural network is easy to model the student interaction with the learned tags.

4 Conclusion

In this paper, we highlighted the limitations of existing DKT that they need predefined human labeling, and proposed an end-to-end DKT model, which does not depend on any human labeling and learns binary question-embedding automatically. Using two open datasets, we empirically validated its effectiveness and showed that the learned tags have information-efficient structure for DKT. Our future works are extending the model's structure to improve the performance, investigating the model's applicable scope on various datasets, and making the obtained tags interpretable by humans. We believe our proposed method could help improve the learning experience of students in more diverse environments.

Acknowledgments

References

Hal Abelson. The creation of opencourseware at mit. *Journal of Science Education and Technology*, 17(2):164–174, 2008.

James H Block and Peter W Airasian. *Mastery learning: Theory and practice*. Holt Rinehart & Winston, 1971.

S Alan Cohen and Joan S Hyman. Learning for mastery: Ten conclusions after 15 years and 3,000 schools. *Educational Leadership*, 37(2):104–9, 1979.

Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

Jianxi Luo and Christopher L Magee. Detecting evolving patterns of self-organizing networks by flow hierarchy measurement. *Complexity*, 16(6):53–61, 2011.

Enys Mones, Lilla Vicsek, and Tamás Vicsek. Hierarchy measure for complex networks. *PloS one*, 7(3):e33799, 2012.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, pp. 505–513, 2015.

J. Stamper, A. Niculescu-Mizil, S. Ritter, G.J. Gordon, and K.R Koedinger. Bridge to algebra 2006-2007. development data set from kdd cup 2010 educational data mining challenge. http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp, 2010.

Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.