Flanagan et al. *Research and Practice in Technology Enhanced Learning* (2024) 19:28
https://doi.org/10.58459/rptel.2024.19028

Research and Practice in Technology Enhanced Learning

RESEARCH Free and Open Access

A human-in-the-loop system for labeling knowledge components in Japanese mathematics exercises

Brendan Flanagan [1] *, Zejie Tian [2], Taisei Yamauchi [2], Yiling Dai [3] and Hiroaki Ogata [3]

*Correspondence:
flanagan.brendanjohn.4n@kyoto-u.ac.jp
Center for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, Kyoto University, Japan
Full list of author information is available at the end of the article

Abstract

Many modern learning systems rely on a data representation of the knowledge that is to be learned to estimate a learner's mastery state and recommend appropriate learning tasks to further improve their acquisition of knowledge and skills. In particular, the rapid development of intelligent tutoring systems (ITS) and standardized curricula has increased the need for information on knowledge structures and their links to learning materials and tasks. However, manually labeling educational data has traditionally been a time-consuming, labor-intensive task, and thus has limited its use by time-constrained teachers and practitioners. In previous research, a range of machine-learning methods have been proposed to address this problem, with only a few of them focusing on Japanese educational datasets from secondary schools. In this paper, to support the labeling of Japanese mathematics exercises by teachers and other domain experts, we apply natural language processing techniques including word-embedding and key-phrase-based exercise-to-exercise similarity methods. We evaluated the proposed method by both the performance of the models when compared to several state-of-the-art methods, and also its effectiveness in supporting humans in the task of labeling educational materials. Through this two-phase evaluation, we found that the proposed method outperformed other methods, and when implemented in a human-in-the-loop system it achieved significantly more accuracy and consumed less time for the task of labeling mathematics exercises.

Keywords: Knowledge labeling, Key-phrase, Nature language processing, Knowledge components, Human-in-the-loop

Introduction

Smart learning systems are increasingly being used in many facets of education, and in particular systems that monitor and estimate learning progress, and recommend learning materials and exercises are often based on a model representing the knowledge components

of a domain. Knowledge tracing in Intelligent Tutoring System (ITS) (Piech et al., 2015; Vie & Kashima, 2019) and personalized recommendation (Hou et al., 2018) are examples of such an environment that is widely being used to support students and promote their learning performance. These systems rely on learning materials and exercises having accurate labels of the knowledge components which has traditionally been performed by domain experts based on their knowledge and experience. However, this task is complex and time-consuming.

Standardized curriculum has also recently been gaining attention, and many governments and organizations are adopting new policies that define broad-ranging curricula that consist of competencies and knowledge components, such as: Common Core State Standards in the USA (National Governors Association, 2010; Porter et al., 2015), the Australian Curriculum (Lingard, 2018), and the Code of Study in Japan (Nakayasu, 2016). As the use of standardized curricula becomes more digitized, teachers and education practitioners who create their own learning materials are increasingly being required to incorporate and adapt resources to these systems. Along with the digitization of the resources, a domain expert is required to assess what parts of the materials address or contain knowledge for specific parts of the curriculum (Churchill, 2007). As a specialized sub-domain of fundamental education, Mathematics has a long history of knowledge model construction that is relevant in both the teaching and studying of mathematics and is a subject that often features in smart learning environments (Carrillo-Yañez et al., 2018). In many schools, learning materials are still manually labeled to integrate instructor-created content into standardized curricula and smart learning environments. However, the manual labeling of these materials takes instructors time and effort, and it can often be a barrier inhibiting the uptake of smart learning environments as shown in the problem overview illustrated in Figure 1.

This problem has drawn the attention of researchers to the task of automatically identifying the knowledge required to solve exercises and then linking it with appropriate knowledge components in mathematics which is a core subject in fundamental education

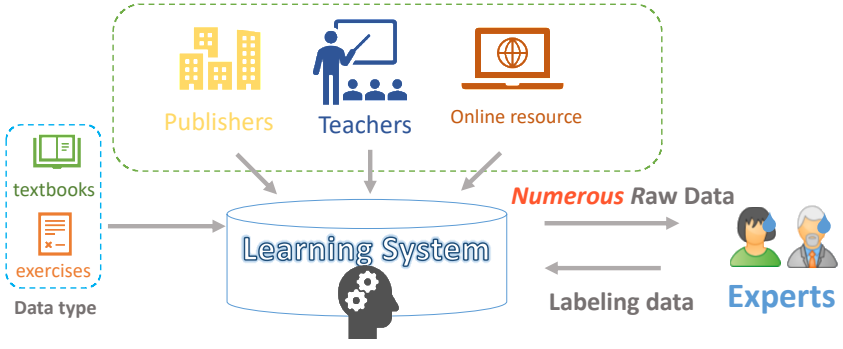


Fig. 1 The bottleneck of knowledge labeling in smart learning systems and the burden it places on domain experts

(Shen et al., 2021). Usually, this task is approached as a multinomial classification problem (Shen et al., 2021), where the contents of the exercise are analyzed and used to predict the most relevant knowledge component label that is assigned to an exercise. Previous research has included supervised learning methods (Hage & Aimeru, 2006; Pardos & Dadu, 2017), unsupervised learning methods (Desmarais, 2012), and deep learning methods (Huang & Li, 2021; Shen et al., 2021). However, these existing methods mainly focus on the automatic labeling of knowledge components in exercises that have been written in English and often require large datasets for training that might not be available or applicable in localized school environments. In addition, few previous research studies have tackled the problem of labeling exercises written in Japanese, and as such there is little work to draw on when investigating reliable linguistic features that may improve the effectiveness of exercise labeling.

Toward these challenges, in this paper, we propose a new model combining a word embedding and key-phrase-based exercise-to-exercise similarity method in an ensemble model. Word embedding models have been shown to be effective in numerous natural language processing tasks including semantic representation of text (Mikolov et al., 2013). In previous research, key-phrase extraction has been shown to be effective in identifying key concepts in learning materials (Chau et al., 2021; Contractor et al. 2015), and the relation of key-phrases has also been examined in the automatic construction of knowledge graphs (Wong et al., 2012). The proposed model's effectiveness was then evaluated and the accuracy was compared to methods from previous research that utilized a range of different machine learning methods for automatic labeling of knowledge components in exercises. We also propose that this method of classification could be implemented in a human-in-the-loop system in which our method is used to recommend labels to domain experts in an effort to increase accuracy and reduce time on task. The present study provides insights for researchers who wish to develop knowledge-based learning systems that can adaptively handle new user-generated mathematics exercises by addressing the following research questions:

RQ1: Can the proposed word-embedding and key-phrase-based exercise-to-exercise similarity model effectively label mathematics exercises in Japanese when compared to previously proposed methods?

RQ2: Can the use of the proposed model reduce the time burden of domain experts when implemented in a human-in-the-loop system for labeling mathematics exercises in Japanese?

RQ3: Can the use of the proposed model increase the accuracy of label mathematics exercises in Japanese when implemented in a human-in-the-loop system?

Related work

Categorizing and labeling exercises

An important task and area of ongoing focus in smart learning systems is the effective management and grouping of exercises and learning materials provided by the system. There is numerous previous research that relies on the results of this task, such as: knowledge tracing (Piech et al., 2015; Vie & Kashima, 2019) and personalized recommendation (Hou et al., 2018; Lin et al., 2021). Traditionally the task of labeling and managing exercises has been accomplished by manual work and highly relies on the knowledge and experience of domain experts. As learning system development continues, there is an increased demand and volume of educational resources, and it becomes necessary to find an automated method for matching exercises with appropriate knowledge components. Each knowledge component usually represents a specific metaknowledge in a domain. Generally, similar exercises serve the same educational purpose and should be identified by the same knowledge components (Del Solato & Du Boulay, 1995). Matching exercises with knowledge components is a similar task to labeling educational contents with knowledge components, except in the case of mathematics there is usually less textual information to utilize. Some previous research matched exercises with the knowledge components of textbooks. Matayoshi and Lechuga (2020) used natural language processing and machine learning methods to match exercises in an ITS system with textbook content from a specific topic. Contractor et al. (2015) utilize external data resources to increase the representation of terms to improve the accuracy of the labeling task.

Some work directly organized exercises by finding similarities between them and then labeling the knowledge components that were found in similar exercises. One supervised machine learning method that is often used to perform this task is the vector space model (VSM). VSM is the model that represents the contents of exercises as a vector and often uses the TF-IDF weight scheme along with methods to calculate the similarity between exercises by text distance methods (Tsinakos & Kazanidis, 2012). In other research, Karlovčec et al. (2012) utilized the support vector machine (SVM) for labeling high dimension data with a large number of knowledge components. Unsupervised learning methods have also been proposed by Desmarais (2012) with applied the Non-negative Matrix Factorization (NMF) based on a Q-matrix to the task and showed promised performance, with the resulting factorization also being interpretable in terms of a Q-matrix. Meng et al. (2016) utilized the LDA method to extract and understand the semantic information of educational contexts. Deep learning methods have also been proposed, with Liu et al. (2018) utilizing multiple features, such as: texts, images, and KCs to extract further semantic information from the training data. Shen et al. (2021) applied a task-adaptive BERT model to reduce problems that were encountered due to the lack of labeled

exercises. Tong et al. (2020) proposed a method based on a knowledge-aware multimodal network, which explored the knowledge hierarchical information to increase the model's accuracy.

A problem often in the high-accuracy methods mentioned above (Liu et al. 2018; Shen et al., 2021; Tong et al., 2020) is that they require a large amount of data to fully train the model and also have less interpretability. In this research, we investigate a method focused on being able to achieve high accuracy while also being flexible with respect to the size of data available. An ensemble model is proposed that combines both word-embedding and key-phrase-based exercise-to-exercise similarity methods. We anticipate that the proposed method will also be effective not only for automated labeling but also in supporting a human-in-the-loop-based method of the exercise labeling task. Previous research into the task of labeling exercises has mainly focused on the algorithms of models, and few research has investigated the effectiveness of assisting time-constrained domain experts, such as teachers who could introduce new exercise contents to the learning system. In this research, to evaluate the effectiveness of the proposed method, we developed a knowledge management system that includes a human-in-the-loop to check and that can also update the model classification results.

Key-phrase extraction

Key-phrases are words or phrases that represent the main topics or ideas in a text (Papagiannopoulou & Tsoumakas, 2020). They can be extracted from a collection of text documents using various techniques, and Siddiqi and Sharan (2015) suggested that there are four major methods for key-phrase extraction: Rule-based linguistic approaches which detect the key-phrases by rules and usually require specific domain knowledge and expert experience for lexical analysis; Statistical approaches based on the frequency within a corpus to filter keywords; Machine learning which usually employ supervised learning methods to automatically detect the key-phrases and required a large of expert labeling data to assure high accuracy; Domain-specific approaches that rely on a knowledge-based or domain knowledge like ontology graphs.

Knowledge components of learning material can be recognized and revealed by a specific key-phrase, and the method is often used to increase the accuracy of models for a number of educational technology tasks. Chau et al. (2021) proposed a machine learning method by using an experts' proposed POS tagger to annotate key-phrases in educational textbooks. Alzaidy et al. (2019) used a deep learning approach based on the Bi-LSTM-CRF model to extract key-phrases from scholarly documents, and it was shown to outperform previous machine learning models. Contractor et al. (2015) extracted key-phrases from learning materials by POS tagger and annotated them according by using external data from the DBPedia spotlight service. Also extracting the relationship of key-phrases has been used

to construct knowledge graphs in previous research (Wong et al., 2012). The construction of educational knowledge graphs can also support other tasks, such as: automatic recommendation and knowledge tracing (Nakagawa et al., 2019; Tong et al., 2020). As with categorizing and labeling exercises, most high accuracy methods require large datasets and often rely on existing methods that have been developed for more general purposes. In this research, we focus on a specific area of mathematics exercises that are well-formed. As such, we reference domain knowledge in the form of key-phrases that have been collected from educational textbook indexes as it is more suitable for domain-based tasks. We also examine the use of the relation of key-phrases by using the linguistic features of mathematics exercises.

Human-in-the-loop

While machine learning models are dominant in many tasks, such as: computer vision, and natural language processing, there are still situations in which systems can benefit from the integration of models and humans within the pipeline, such as: training more accurate reinforcement learning models through human feedback, and verifying if a decisions made by a model is ethical or not before actions are taken (Wu et al., 2022). Recent research has proposed a human-in-the-loop approach to ensure that possible errors aren't left unchecked and to reduce the potential harm that could be caused by incorrect classification. Human-in-the-loop refers to the integration of human decision-making and machine learning algorithms to achieve better results than either could achieve alone. Within the field of education technology, human-in-the-loop is applied to improve the decision-making process of systems to provide better transparency, avoid bias and possible errors, and ensure that educators remain the central decision-makers for instruction and choose how systems are implemented into their work (Ninaus & Sailer, 2022). Human-in-the-loop is also used in learning analytics and educational data mining to improve the accuracy of predictions and recommendations by allowing human feedback to be included in the training and implementation process (Bhutoria, 2022). In essence, a human-in-the-loop system could possibly improve the performance of a model by confirming the automatic predictions with human experience and knowledge while keeping the burden of the task to a minimum (Wu et al., 2022). Usually, the complex task of labeling data is repeated for each new dataset and requires a large amount of manual work (Wu et al., 2022), however, a human-in-the-loop approach has the potential to reduce the time taken and improve accuracy.

Classification models are often constructed from data that has been tagged and labeled by human experts based on their experiences and rely highly on the quality of training datasets. However, this can also create potential problems when there is a gap between the distribution of training datasets and the data that is encountered in real-world tasks (Bengio

et al., 2020). Due to the gap between real-world tasks and training datasets, Liu et al. (2019) proposed to introduce a human-in-the-loop approach where the model first generates pre-labeled data, then this is checked and modified accordingly by the human. The main purpose is to reduce the human labor burden of the task while improving the model performance through the continuous collection of data.

As we have found that few previous research into the labeling of exercises has investigated a human-in-the-loop approach, we aim to investigate if it can reduce the amount of time required for the task by a human, and also if the method can help improve the overall accuracy.

Method

In this research, we examine how to reduce the burden placed on domain experts in the task of labeling mathematics exercises by proposing a classification model to recommend candidate labels. Figure 2 shows an overview of the scenario in which the proposed method is implemented: a knowledge portal system has been developed and is used by domain experts to manage the labeling of exercises. The proposed automatic labeling model provides a list of recommended labels, and the domain expert is a human-in-the-loop who can either select a recommended knowledge component label or search to find a more appropriate label in the system. We anticipate that the proposed method will not only increase the accuracy of the labeling of knowledge components in exercises, but also reduce the time taken by domain experts in carrying out this complex and important task.

Data collection and preprocessing

The method proposed in this paper was implemented as a part of the LEAF learning platform (Flanagan & Ogata, 2018), where mathematics exercises are uploaded to BookRoll the digital learning material reader as PDF files. The PDF files have been

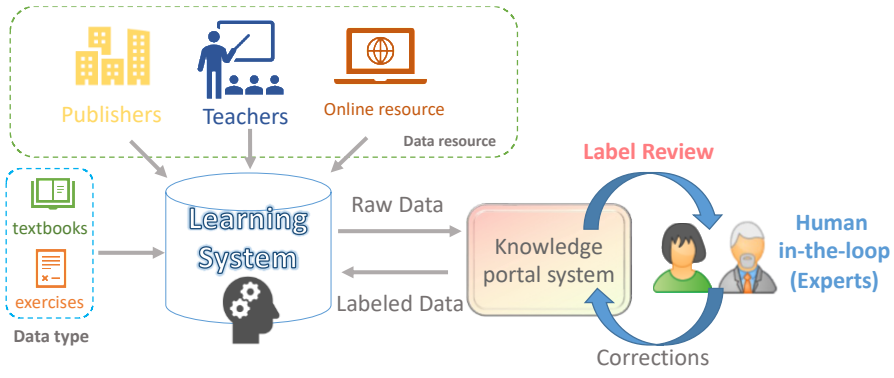


Fig. 2 An overview of the proposed human-in-the-loop learning material and exercise labeling system

provided directly from textbook publishers and can be displayed and answered directly within the BookRoll system using the built-in quiz feature. However, there may be artifacts such as mathematic formulas or functions that are difficult to extract as they can be represented as vector images or utilize special-purpose fonts when compared with purpose-built ePUB or HTML files (Matayoshi & Lechuga, 2020). The text information in Japanese was extracted using the Pdf2text (jalan, 2021) Python library, and then segmentation was performed with the Nagisa (taishi-i, 2020) Python library to extract individual words, a process that is required as there are no word boundaries in the Japanese language (Kitagawa & Komachi, 2018). Additional preprocessing was also performed to remove stop words and noise, such as: those from incomplete formulas, numbers, and functions. We conducted the experiment on 830 Japanese mathematics exercises collected from exercises and textbooks used in junior high school. The datasets contained four main topics: geometry, function, statistics and probability, and 13 detailed knowledge components had been identified by domain experts.

Proposed model

In this paper, we propose an ensemble method combining both word-embedding and key-phrase-based exercise-to-exercise similarity methods. Modern word-embedding was popularized for many tasks in natural language processing with the publication of Mikolov et al.'s (2013) research into Word2Vec which allowed the representation of words or sentences as dense vectors. In this research, we utilize the FastText method (Joulin et al., 2016) to transform the exercises into a vector space for understanding the semantic meaning of exercises. Moreover, the key-phrase-based exercise-to-exercise similarity can help to indicate the key knowledge components of each exercise. The overall structure of our model is shown in Figure 3. The whole process can be simplified as a task that involves inputting the exercise text and outputting the most relevant knowledge components. For each of the models, the output is the vector of the probability of a knowledge component being contained in each exercise. We combine both the word embedding and key-phrase-based exercise-to-exercise similarity model results as the input feature for the final WE-KE ensemble model. The details of these methods will be introduced in the following sections.

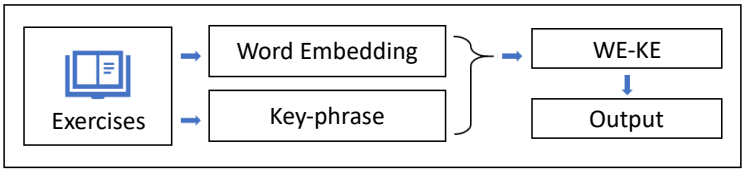


Fig. 3 The overview of the proposed word embedding and key-phrase-based exercise-to-exercise similarity ensemble model for exercise labeling

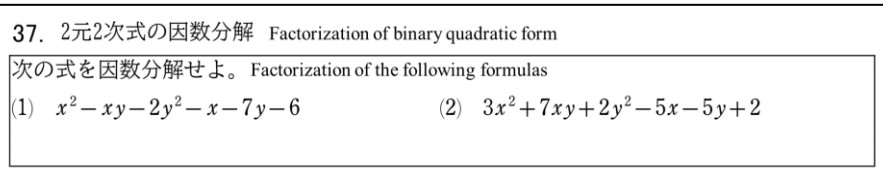
Word embedding method

To transform the words of the exercises into word embeddings we utilized the FastText method (Joulin et al., 2016). In this research, we used the publicly available multilingual pre-trained FastText word embeddings consisting of 300-dimensional representations of Japanese words trained on a large corpus. A tool is also provided to reduce the dimensionality of the embeddings, however, research has shown that higher-dimensional word embeddings can capture more nuanced semantic information compared to lower-dimensional embeddings, so we decided to use the original pre-trained word embeddings (Umer et al., 2023). As each word is represented as a 300-dimension vector, we used the average of all the word vectors for representing the whole exercise sentences. An XGBoost model was trained on the 300-dimension vector representation of the exercises and output the probability for each knowledge label.

Key-phrase-based exercise-to-exercise similarity method

During data preprocessing, words are extracted from the Japanese mathematics exercise sentences, and the role that each word has in the sentence is tagged with appropriate parts of speech. Often in this process, compound words are divided into individual components and may lose their overall semantic context. While preprocessing Japanese junior high school mathematics exercise text, we found that important words relating to key knowledge concepts might be segmented, such as “因数分解” (factorization) as shown in Figure 4 reveals that it is a question about factorization. However, this compound word is segmented into “因数” (factor) and “分解” (decomposition), therefore partially losing semantic context. To overcome this problem, we compiled a list key-phrase of key mathematics concepts from the indexes of textbooks and supplementary learning material. This was then used to detect the key-phrases contained in each exercise and determine whether it had specific math concepts. As for the example question, the key-phrase list is [‘2 元 2 次式 (binary quadratic)’, ‘因数分解’(factorization)].

The method proposed by Wu et al. (2012) for extracting the relationship between key-phrase was also examined in the proposed key-phrase model. We extracted the relationship of knowledge concepts by utilizing basic Japanese linguistic rules. The Japanese word ‘の’ (‘no’) often plays the role of possessor and modifier, and could possibly provide insight


37. 2元2次式の因数分解 Factorization of binary quadratic form

次の式を因数分解せよ。Factorization of the following formulas

(1) $x^2-xy-2y^2-x-7y-6$ (2) $3x^2+7xy+2y^2-5x-5y+2$

Fig. 4 The math exercises about factorization in the junior high school math

Table 1 Examples of linguistic pattern rules

Pattern	Example
_の_ (_'no'_)	**多項式の次数** (Degree of polynomial)
_と_の_ (_'to'_'no'_)	**乗法と除法の混じった計算** (Calculation with a mixture of product and division)
_の_と_ (_'no'_'to'_)	**多項式の加法と減法** (Adding and subtracting polynomials)
_を_ (_'wo'_)	**同類項をまとめ** (Summarize similar terms)

further detail into the knowledge contained, such as "整数の加法"(integer addition) which shows that the exercise is about the addition of integers. The Japanese word 'を' ('wo') usually is used to mark the object of the sentence, which often shows the purpose of exercises such as "式を計算" (calculation of a function). The Japanese word 'と' ('to') is a cumulative coordinating conjunctive that is similar in meaning to 'and' in English and is used to connect to similar items together. Examples of the linguistic pattern rules used to identify relations of key-phrases are shown in Table 1.

We assume that if exercises contain the same knowledge components, then they should have similar key-phrases and key-phrase relations. Based on this assumption, we calculate the similarity score between exercises as follows:

$$similarity(E_1, E_2) = \frac{keywords(E_1) \cap keywords(E_2) + relationship(E_1) \cap relationship(E_2)}{keywords(E_1) \cup keywords(E_2) + relationship(E_1) \cup relationship(E_2)}$$

The $keywords(x)$ function represents the set of key concepts for exercise x and $relationship(x)$ function represents the set of key concepts relationship for exercise x. We calculate the average of similarity scores between exercises in the test set and a batch of exercises belonging to certain knowledge components in the training set and store the final results into a matrix.

Ensemble model

The ensemble model proposed in this research is constructed from the results of a word embedding-based method to capture the semantic meaning of exercises from text, and a key-phrase-based exercise-to-exercise similarity method that targets the key topics of exercises in the form of keywords. It is hypothesized that similar exercises will contain similar semantic meanings and key concepts. An overview of the proposed ensemble model is shown in Figure 5 and was designed based on the simply weighted average method proposed by Zhou (2012).

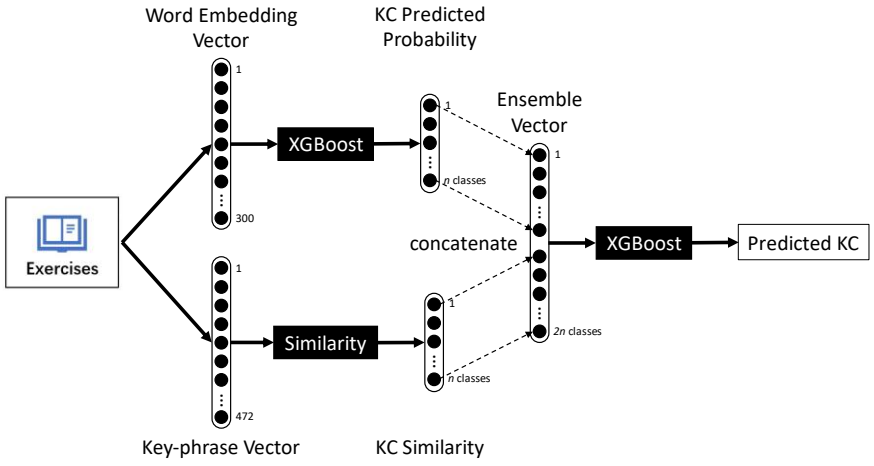


Fig. 5 An overview of the ensemble method process

It is an extension of the simple voting ensemble method and applies fixed weights to each of the models in the ensemble depending on the importance the model has in determining the correct final result. Similar methods have also been found to be effective in the labeling of natural language data into discrete classes by taking advantage of known key characteristics (Flanagan & Hirokawa, 2018). First, the word embedding model and key-phrase-based exercise-to-exercise similarity model are trained from the training dataset, with each model outputting a vector of the prediction for each class. In the case of the word embedding model, softmax is applied to the XGBoost output, and this results in a vector of the predicted probability for each target KC. The key-phrase-based exercise-to-exercise similarity model calculates the similarity between the input test data and the training dataset for which the KC labels are known, and results in a vector of the normalized similarity of the test data with each group of training data that has the same KC label. The model output vectors are then concatenated to form the input for the XGBoost model at the final stage of the ensemble model.

Experiment

The XGBoost model was used for the classification by word embedding-based methods and the final state of the ensemble method. We also evaluated the following methods as baselines for classifying question text that have been proposed in previous research to measure the compare the effectiveness of the proposed methods:

- Vector space model (VSM) (Tsinakos & Kazanidis, 2012) transformed exercises into a vector space and compared the cosine similarity between them.
- Support vector machine (SVM) (Karlovčec et al., 2012) has shown good results when classifying high dimensional input features and is suitable for the classification of exercises with dense concepts.

- XGBoost (Chunamari et al., 2022) is a scalable machine learning model that is built on the tree-boosting method and has been shown to have high performance in many kinds of classification problems.
- Neural network (NN) (Patikorn et al., 2019) is a classic machine learning model and has shown promising performance in the classification of exercises.

Evaluation

Although the model applied in this section provides some indicators to evaluate model performance, it does not provide any accuracy indicator. Therefore, we applied a range of metrics, following the concept of prediction accuracy proposed by Huang and Fang (2013) to design indicators to evaluate prediction performance. In particular, we measure the model performance with the macro F1-score which is the unweighted F1 mean across all target classes, and Accuracy. The equations for measurement are shown below, where TP = true positive, FP = false positive, P = positive, TN = true negative, and N = negative when comparing the gold standard class with test data predictions by the model as described by Japkowicz and Shah (2011).

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{P}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + TN}{P + N}$$

The performance of each model was evaluated using 5-fold cross-validation to ensure generalizability. The dataset was split at the KC level for cross-validation, in which the exercises of each KC were split into 5 distinct groups, and the training and testing data would therefore contain the same number of KCs.

Results

The experiment results are shown in Table 2. According to the results, the proposed model achieved a macro F1-score of 0.7897 and an accuracy of 0.7957 for the labeling of knowledge components in exercises, which outperformed the other baseline methods from previous research. Other top performing models that should be mentioned are XGBoost

Table 2 Results of the evaluation of the proposed model compared with models from previous research

Evaluation	Model				
	VSM	XGBoost	SVM	NN	WE-KE
Macro F1-score	0.5559	0.7137	0.6964	0.7127	**0.7897**
Accuracy	0.5579	0.7287	0.7073	0.7287	**0.7957**

with a macro F1-score of 0.7137 and accuracy of 0.7287, and NN with a macro F1-score of 0.7127 and accuracy of 0.7287.

In summary, our model outperforms other models across a range of junior high school Japanese mathematics exercises contained in the datasets. The results show that the ensemble model effectively combines both semantic meaning that is represented by word embedding and key-phrase and relations to achieve effective labeling of knowledge components. In the next section, we will examine the effectiveness of the individual components of the proposed method in an ablation study.

Ablation study

In order to understand the importance of different parts of the proposed model that contribute to the performance, we conducted an ablation study that evaluates the individual parts and their impact on the model (Tian et al., 2021). The proposed model consists of two main feature sets: word embedding features, and key-phrase-based exercise-to-exercise similarity features. The results of the evaluation of the individual feature sets are shown in Table 3 along with the proposed model that combines both feature sets. All of the parts that were evaluated were constructed using the same XGBoost model and hyper-parameters so no other variables could affect the outcome. The results show that the individual feature sets on their own have promising results, with macro F1-score and accuracy equivalent to that of other high-performing models that were evaluated. However, combining these features in an ensemble model provides superior results. Word embedding-based methods often excel at representing semantic information and key-phrase-based exercise-to-exercise similarity can identify labels by domain keywords.

Assisting human domain experts

To evaluate the effectiveness of the proposed label classification model, we designed and conducted an experiment to examine if the proposed method can improve the speed and

Table 3 Results of the ablation study

Evaluation	Model		
	WE	KE	WE-KE
Macro F1-score	0.7137	0.7350	**0.7897**
Accuracy	0.7287	0.7195	**0.7957**

Table 4 The accuracy of top *k*

Top *k* Recommendations	Model Accuracy				
	VSM	XGBoost	SVM	NN	WE-KE
k = 1	0.5579	0.7287	0.7073	0.7287	**0.7957**
k = 2	0.6829	0.8445	0.8445	0.8994	**0.8994**
k = 3	0.7378	0.9146	0.9132	0.9421	**0.9543**

accuracy of human domain experts in the labeling task. While the proposed model provides relatively high accuracy in classifying the knowledge contained, we also evaluated the accuracy of the model in relation to top *k* classifications to see if it could provide greater accuracy. This design also allows the domain experts to participate as a human-in-the-loop and decide the appropriate label from a short list of candidate labels. For this evaluation, a classification from the proposed model was recorded as correct if a correct label was classified within the top *k* label classifications from the prediction model. The results of the top *k* accuracy are shown in Table 4, and it can be seen that the proposed WE-KE model has high accuracy over different top *k*'s.

While increasing the number of *k* classifications would most likely result in higher accuracy and increase in the perceived autonomy through choice by the human domain expert, it could also potentially increase their cognitive load in the labeling task (Schneider et al., 2018). This informed the design of the label assignment interface of the knowledge management system, and it was decided that three labels would be shown to the human domain expert to allow choice from recommendations with higher accuracy, which is anticipated will result in the labeling task taking less time. The interface of the label assignment user interface of the knowledge management system is shown in Figure 6, with the user being presented with a list of quiz titles that can be clicked on to verify the contents of the quiz, a list of three knowledge component labels as recommended by the system,

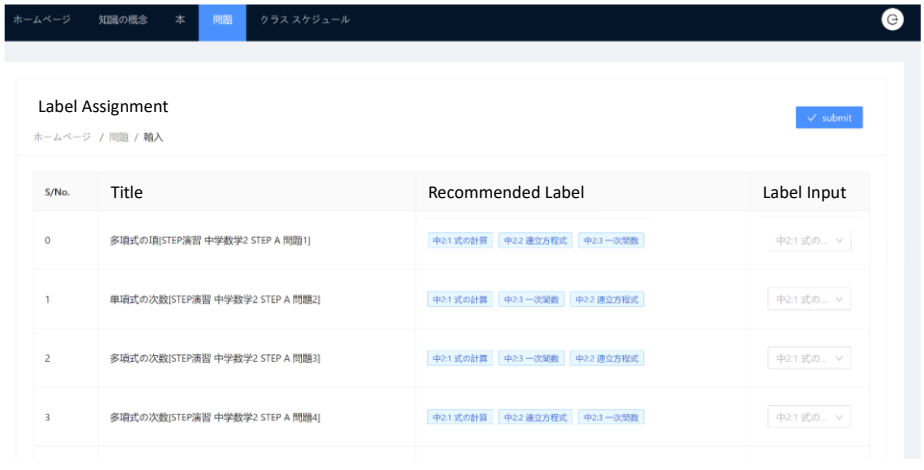


Fig. 6 A screenshot of the label assignment UI in the Knowledge Management System

and a searchable label input field on the right. The user can select a recommended label by clicking directly on the label and this will be assigned to the label input field. When participants were using the system without the model classification treatment, the recommended label interface was still shown, except the labels shown were static for all of the items, and therefore the interface features between different conditions do not differ.

Experiment design

To verify the effectiveness of the proposed model in the human-in-the-loop system, we designed an experiment and recruited participants who were Ph.D. or master's degree students from a top national university and who had good self-reported mathematics ability. As these participants were required to be suitable for the task of labeling mathematics exercises in the role of a domain expert, each participant was given the same tutorial to explain the functions of the system, and to ensure that all of the participants understood the task, and a preliminary test was given to check their ability to identify the meaning of sample quiz exercises and the knowledge components that are contained. This resulted in 10 candidates who were suitable to participate in the experiment. To evaluate the effectiveness of the proposed classification model in providing recommendations for the task of labeling mathematics exercises, we employed a crossover design as proposed by Laska et al. (1983). This experimental design involves subjecting two groups to two treatments in different sequences, such as: treatment A followed by treatment B for group one, and treatment B followed by treatment A for group two. One treatment followed by another treatment can be influenced by the carryover effect, where the first treatment impacts the results of the second treatment, so it is important to confirm and measure the size of the effect across different sequences of treatments.

An overview of the crossover experiment design is shown in Figure 7, with the first step consisting of randomly assigning the participants to two different experiment groups that would undergo different sequences of the two treatments over two periods. Due to the limited participant sample size, we adopted a balanced group design where equal numbers of participants were assigned to each group, as it has been reported in previous research to be generally robust for repeated-measures analysis (Keselman et al., 1996; Oberfeld & Franke, 2013; Sullivan et al., 2016). In each period, the labeling accuracy and time taken were measured so that differences in the effectiveness of the treatments could be compared. In treatment A the participants labeled the knowledge component contained in a quiz without the assistance of the label recommendation based on the proposed model. In treatment B the knowledge management system provided three label recommendations for each item that was labeled as part of the task. The participants were given a user perception questionnaire after the first period of the experiment for the qualitative evaluation of the knowledge management system. We randomly selected 50 exercises from the dataset for

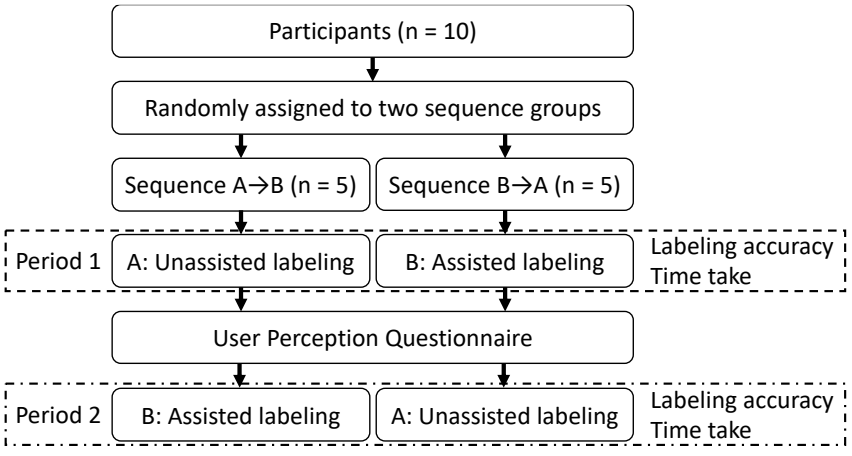


Fig. 7 Crossover experimental design consisting of two sequence groups

the labeling task, where 25 items were labeled by the participants in each period of the experiment.

The accuracy of the labels assigned by participants and the time taken to perform the task were recorded by the knowledge management system for each period. The descriptive statistics of these variables are shown in Table 5, contrasting the treatment of, A) unassisted labeling in which the system did not provide any recommended labels, and B) assisted labeling in which the system provided recommended labels from the proposed model.

Regardless of the sequence in which the treatments were applied, for treatment B the accuracy increases, and the time taken decreases when compared to treatment A. A plot of the descriptive statistics is shown in Figure 8, where it can be seen that the change in accuracy and time is less pronounced in the AB sequence of treatments when compared to the BA sequence. The crossover design of the experiment can introduce possible carryover effects which will be examined in the latter of this section to see if they have a significant impact on the results.

The correlation between the accuracy and time measurements taken for different periods in the experiment are shown in Table 6. While there are some weak positive and negative correlations between the two treatments of time and accuracy, it should be noted that none

Table 5 Descriptive statistics of labeling accuracy for treatments A and B over the different sequences of the crossover experiment

Treatment	Sequence	Labeling Accuracy		Time Taken (s)	
		Mean	SD	Mean	SD
A: Unassisted labeling	AB (n = 5)	0.728	0.034	784	135
	BA (n = 5)	0.744	0.051	831	238
B: Assisted labeling	AB (n = 5)	0.813	0.041	722	207
	BA (n = 5)	0.869	0.016	594	159

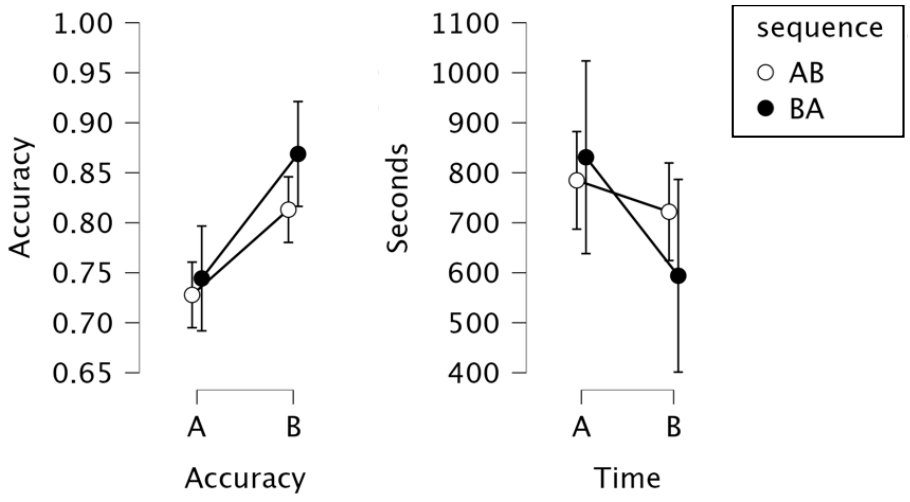


Fig. 8 Descriptive plots of the differences in treatments and the effect of the sequence in which they were applied

Table 6 Correlation between the variables of time and accuracy in the period (A) where no treatment was applied, and period (B) where the label recommendation treatment was applied

	Time (A)	Time (B)	Accuracy (A)	Accuracy (B)
Time (A)				
Time (B)	0.48896803			
Accuracy (A)	-0.3751799	-0.3332943		
Accuracy (B)	0.04167775	-0.5120429	0.24544877	

of these correlations were revealed to be significant ($p > 0.05$). We conducted repeated measures analysis of variance (ANOVA) to examine the differences in the accuracy and time taken to in the labeling task undertaken by participants.

Results of assisting human domain experts

Levene's test revealed homogeneity of variance among the labeling accuracy of the two groups ($F = 0.339$; $p = 0.567 > 0.05$). As indicated in Table 7, the repeated-measures ANOVA revealed significant intertreatment differences ($F = 44.351$; $p < 0.001 < 0.05$) in accuracy scores, indicating that the label recommendation model is effective in increasing accuracy. Also, it was revealed that there were no significant intergroup differences ($F = 1.551$, $p = 0.248 > 0.05$), indicating that the carryover effect from the sequence of the

Table 7 Results of repeated-measures ANOVA: labeling accuracy

Cases	Sum of Squares	DF	Mean Square	F	p	η^2
Accuracy	0.055	1	0.055	44.351	< 0.001	0.637
Accuracy ✻ Sequence	0.002	1	0.002	1.551	0.248	0.022

Table 8 Results of repeated-measures ANOVA: time taken for labeling

Cases	Sum of Squares	DF	Mean Square	F	p	η^2
Time	112350.050	1	112350.050	7.426	0.026	0.153
Time ✻ Sequence	37932.050	1	37932.050	2.507	0.152	0.052

treatment is minimal. It should also be noted that the performance from the human-in-the-loop method yielded higher accuracy in the labeling task then just the individual parts of the method, namely: only human labeling accuracy, and only model labeling accuracy.

Levene's test revealed homogeneity of variance among the time taken for labeling of the two groups ($F = 0.004$; $p = 0.948 > 0.05$). As indicated in Table 8, the repeated-measures ANOVA revealed significant intertreatment differences ($F = 7.426$; $p = 0.026 < 0.05$) in time taken for labeling, indicating that the label recommendation model is effective in reducing the amount of time taken to assign labels. Also, it was revealed that there were no significant intergroup differences ($F = 2.507$, $p = 0.152 > 0.05$), indicating that the carryover effect from the sequence of the treatment is minimal.

User perception of knowledge label recommendation

In the previous section, it was shown quantitively that the system can effectively support accurate labeling of knowledge, however, it is also important to conduct a qualitative evaluation of the effectiveness of the knowledge management system. A questionnaire was given to the participants after the first period of the experiment to assess if there was a perceived difference in the assistance from the system while conducting the knowledge labeling task. The questionnaire was designed based on the design proposed by Hwang et al. (2011), and consisted of six questions in total on a 10-point Likert scale: one question on the participant's confidence of their performance in assigning labels, and five questions on the perceived satisfaction with the knowledge management system. To measure the overall reliability of the questionnaire, we calculated Cronbach's alpha and it was found to be of excellent internal consistency (6 items; $\alpha = 0.90$). Table 9 shows the descriptive statistics of the questionnaire items. The t-test results show for question item 1 that participants did not perceive any significant difference in their confidence in their performance in the labeling task whether they were being assisted by the system or not. Question items 2 through to 6 show that the participants were overall significantly satisfied with the assistance provided by the knowledge management system. The results of this perception questionnaire confirm that the proposed system and label classification model is effective in supporting human-in-the-loop domain experts in assigning appropriate labels to describe knowledge components in quiz exercises.

Table 9 Participant perception questionnaire results

Question item	Treatment	Mean	P-value
1. How confident you are about your classification results?	A	7.0	0.63
	B	6.4	
2. The rank option in the selection box is helpful for classification?	A	1.4	0.00
	B	8.4	
3. I am satisfied with the rank option in the selection box result?	A	2.2	0.00
	B	7.6	
4. The rank option in the selection box can help me to find the right knowledge component?	A	2.2	0.00
	B	8.2	
5. If given the chance, I am willing to re-use the system to classify the quizzes?	A	3.6	0.00
	B	8.6	
6. Generally speaking, I am satisfied with the current system?	A	3.6	0.00
	B	8.2	

Discussion and conclusion

In this research, we proposed an ensemble model that consisted of a word embedding and key-phrase-based exercise-to-exercise similarity method. The following sections discuss the results presented in this paper in relation to the research questions and echoing related studies.

Can the proposed word-embedding and key-phrase-based exercise-to-exercise similarity model effectively label mathematics exercises in Japanese when compared to previously proposed methods? (RQ1)

The first research question concerns with the performance of the proposed model when compared to previous research. It was assumed that the word embedding based method could provide a semantic representation of the exercises through the analysis of dense word vectors. The key-phrase-based exercise-to-exercise similarity method and the use of information about the relation of key-phrases could provide insight into the key topics of each exercise. We performed an experiment to compare the proposed model that combines these two methods in an ensemble with state-of-the-art baseline models from previous research. The evaluation results show that the proposed method outperforms other baseline models for both Macro F1-score and Accuracy metrics. An ablation study was conducted to understand the importance of different parts of the proposed model, and how they contribute to the overall performance of the ensemble model. It was found that the individual components' performance was inferior to that of the ensemble model. Similar results have also been obtained in previous work where ensembles of models were combined to realize increased model prediction performance (Flanagan & Hirokawa, 2018).

Although the present research shows effective performance for the labeling knowledge components in Japanese mathematics exercises, there are several challenges and limitations that still need to be examined. First, parsing a PDF file of a mathematics exercise results

in the loss of information that might otherwise be retained in other formats, such as: HTML, XML, or LaTeX. The mathematical functions and special characters can be well represented in these formats, however, conversion to and from PDF format may result in the loss of such information due to loose representation restrictions. Especially for mathematics exercises, the math characters and functions could potentially be useful information for understanding the meaning of exercises (Patikorn et al., 2019). Therefore, the extraction and utilization of this information remains an important topic for future research. Second, graphics and plots are often integral parts of exercises in the topic of geometry in a mathematics course and could provide crucial information about the knowledge components contained in the exercises (Liu et al., 2018). However, the method proposed in this research does not utilize this information and therefore may cause lower labeling accuracy for geometry exercises. Moreover, the overfitting problem as described by Patikorn et al. (2019) also needs to be considered, as exercises are usually generated from a common set of templates, and there are a large number of near-identical problems. The model may only identify some specific words or structures, but not truly understand the meaning of exercises or the knowledge components contained. Additionally, the proposed model has been designed to address the labeling task as a multiclassification problem where one specific knowledge component is assigned to an exercise. However, it is possible that one exercise could contain several knowledge components, as is often the case in real-world mathematics exercises. Therefore, methods for labeling all of the possible knowledge components contained in exercises should be addressed in future research.

Can the use of the proposed model reduce the time burden of domain experts (RQ2) and increase accuracy (RQ3) when implemented in a human-in-the-loop system for labeling mathematics exercises in Japanese?

The second and third research questions concern how effective the implementation of the proposed model in a human-in-the-loop system was in reducing the time burden of domain experts and increasing the accuracy of labeling mathematics exercises respectively. A crossover design experiment was conducted with 10 domain expert participants split into two groups. One group was tasked with labeling exercises firstly unassisted, and then with assistance from the system, and the other group was given the treatment in the reverse order. Repeated measures ANOVA analysis showed that there was no significant difference between the results of the experiment due to carryover effects that can be present when using a crossover design. It was found that there was a significant improvement in both the time taken and the accuracy of the labels assigned by participants when assisted by the proposed model in the human-in-the-loop system.

While it was shown to be effective in both increasing accuracy and reducing time taken, we did not test the retraining and model based on the feedback from domain experts during the experiment. It is possible that the proposed model could improve in accuracy by utilizing the information provided by humans using the system and requires further investigation. Furthermore, the human-in-the-loop experiment involved 10 participants, which may have a negative impact on the reliability of the results due to possible variances in the participants. There are two main factors that would influence the result in this study and could provide an explanation for the difference between the results of the different treatment sequences: firstly, the participants within each of the groups, and secondly, carry-over due to the order in which the treatment was administered (Johnson, 2010). To minimize the possible difference between the group's abilities and ensure fitness, we performed a preliminary test to assess if the candidate possessed the required skills. For the order of treatment, it could be possible that the difference between these two numbers is due to carry-over effects, however, the carry-over effects between the groups based on sequence were found to be insignificant as reported for intergroup differences in Accuracy and Time in the results section. In particular, the sequence in question A->B first starts with unassisted labeling (treatment A), after which assisted labeling (treatment B) is conducted. Therefore, it is plausible that a participant might get more proficient at a task as repetitions increase. Further research could address these issues by extending the period of the experiment and the number of participants to help verify the results reported, however, it is beyond the scope of this current research.

In conclusion, we proposed an ensemble method for the automatic labeling of knowledge components contained in Japanese junior high school mathematics exercises. It consists of word-embedding for the representation of the semantic meaning of texts and a key-phrase-based exercise-to-exercise similarity extraction model to identify domain knowledge and can be easily interpretable. An experiment was conducted to compare the accuracy of the proposed model with other state-of-the-art models from previous research and it was shown to outperform. We developed a knowledge portal management system and implemented the proposed method in a simple interface to recommend three candidate knowledge component labels to domain experts in a human-in-the-loop design of the labeling task. It was shown that the accuracy of the human-in-the-loop method proposed in this research significantly outperformed both human-only labeling and model-only labeling, while significantly reducing the amount of time taken on the task.

Abbreviations

ANOVA: Analysis of Variance; ITS: Intelligent Tutoring Systems; NMF: Non-negative Matrix Factorization; NN: Neural Network; SVM: Support Vector Machine; VSM: Vector Space Model.

Acknowledgements

We would like the thank the participants of the human test for their time in helping to evaluate the performance of the method proposed in this research. Without their time and cooperation, this research would not be possible.

Authors' contributions
BF, ZT, TY, YD, and HO contributed to the research conceptualization and methodology. BF and ZT developed models and systems, conducted experiments, and wrote the manuscript. BF analyzed the experiment results. BF, TY, and YD provided comments to improve the manuscript. All authors read and approved the final manuscript.

Authors' information
Brendan Flanagan is an Associate Professor at the Center for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, and the Graduate School of Informatics at Kyoto University. He received a bachelor's degree from RMIT University and master's and Ph.D. degrees from the Graduate School of Information Science and Electrical Engineering, Kyushu University. His research interests include: Learning Analytics, Educational Data Science, Educational Data Mining, NLP/Text Mining, Machine Learning, Computer Assisted Language Learning, and the Application of Blockchain in Education.

Zejie Tian is a master's student at the Graduate School of Informatics, Kyoto University. His research focuses on transformer-based knowledge tracing and the automatic extraction of metadata from learning materials.

Taisei Yamauchi is currently a master's student at the Graduate School of Informatics, Kyoto University. His research interests include assigning topics to learning materials automatically and improving students' learning engagement or achievement with topic-based learning pattern analysis.

Yiling Dai is a Program-Specific Researcher at the Academic Center for Computing and Media Studies, Kyoto University. She received a Bachelor's degree from Zhejiang University, a Master's degree from the Graduate School of Business, Rikkyo University, and a PhD degree from the Graduate School of Informatics, Kyoto University. Her research interests include: Information Retrieval, Knowledge Discovery, Educational Data Mining and Learning Analytics.

Hiroaki Ogata is a full Professor at the Academic Center for Computing and Media Studies, Kyoto University. His research interests include: Learning Analytics, Evidence-Based Education, Educational Data Mining, Educational Data Science, Computer Supported Ubiquitous and Mobile Learning, and CSCL.

Funding
This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) JP20H01722 and JP23H01001, (Exploratory) JP21K19824, (A) JP23H00505, and NEDO JPNP20006.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Center for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, Kyoto University, Kyoto, Japan.

[2] Graduate School of Social Informatics, Kyoto University, Kyoto, Japan.

[3] Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan.

Received: 31 January 2023 Accepted: 12 November 2023
Published online: 1 January 2024 (Online First: 20 November 2023)

References
Alzaidy, R., Caragea, C., & Giles, C. L. (2019). Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents. In L. Liu & R. White (Eds.), *Proceedings of WWW '19 The World Wide Web Conference* (pp. 2551–2557). ACM. https://doi.org/10.1145/3308558.3313642

Bengio, Y., Deleu, T., Rahaman, N., Ke, N. R., Lachapelle, S., Bilaniuk, O., Goyal, A., & Pal, C. (2020). A meta-transfer objective for learning to disentangle causal mechanisms. In *Proceedings of the International Conference on Learning Representations 2020*.

Bhutoria, A. (2022). Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence*, *3*, 100068. https://doi.org/10.1016/j.caeai.2022.100068

Carrillo-Yañez, J., Climent, N., Montes, M., Contreras, L. C., Flores-Medrano, E., Escudero-Ávila, D., Vasco, D., Rojas, N., Flores, P., Aguilar-González, Á., Ribeiro, M., & Muñoz-Catalán, M. C. (2018). The mathematics teacher's

specialised knowledge (MTSK) model. *Research in Mathematics Education*, *20*(3), 236–253. https://doi.org/10.1080/14794802.2018.1479981

Chau, H., Labutov, I., Thaker, K., He, D., & Brusilovsky, P. (2021). Automatic concept extraction for domain and student modeling in adaptive textbooks. *International Journal of Artificial Intelligence in Education*, *31*(4), 820–846. https://doi.org/10.1007/s40593-020-00207-1

Chunamari, A., Yashas, M., Basu, A., Anirudh, D. K., & Soumya, C. S. (2022). Quora question pairs using XG Boost. In N. R. Shetty, L. M. Patnaik, H. C. Nagaraj, P. N. Hamsavath & N. Nalini (Eds.), *Emerging Research in Computing, Information, Communication and Applications. Lecture Notes in Electrical Engineering, vol 790* (pp. 715–721). Springer, Singapore. https://doi.org/10.1007/978-981-16-1342-5_55

Churchill, D. (2007). Towards a useful classification of learning objects. *Educational Technology Research and Development*, *55*, 479–497. https://doi.org/10.1007/s11423-006-9000-y

Contractor, D., Popat, K., Ikbal, S., Negi, S., Sengupta, B., & Mohania, M. K. (2015). Labeling educational content with academic learning standards. In *Proceedings of the 2015 SIAM International Conference on Data Mining* (pp. 136–144). Society for Industrial and Applied Mathematics.

Del Solato, T., & Du Boulay, B. (1995). Implementation of motivational tactics in tutoring systems. *Journal of Artificial Intelligence in Education*, *6*(4), 337–378.

Desmarais, M. C. (2012). Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explorations Newsletter*, *13*(2), 30–36. https://doi.org/10.1145/2207243.2207248

Flanagan, B., & Hirokawa, S. (2018). An automatic method to extract online foreign language learner writing error characteristics. *International Journal of Distance Education Technologies*, *16*(4), 15–30. https://doi.org/10.4018/IJDET.2018100102

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning: An International Journal*, *10*(4), 469–484. https://doi.org/10.34105/j.kmel.2018.10.029

Hage, H., & Aimeru, E. (2006, March). ICE: A system for identification of conflicts in exams. In *Proceedings of the IEEE International Conference on Computer Systems and Applications 2006* (pp. 980–987). IEEE. https://doi.org/10.1109/AICCSA.2006.205207

Hou, Y., Zhou, P., Xu, J., & Wu, D. O. (2018). Course recommendation of MOOC with big data support: A contextual online learning approach. In *Proceedings of the IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 106–111). IEEE. https://doi.org/10.1109/INFCOMW.2018.8406936

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, *61*, 133–145. https://doi.org/10.1016/j.compedu.2012.08.015

Huang, T., & Li, X. (2021). An empirical study of finding similar exercises. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Workshop on Math AI for Education (MATHAI4ED)*.

Hwang, G. J., Wu, C. H., Tseng, J. C., & Huang, I. (2011). Development of a ubiquitous learning platform based on a real-time help-seeking mechanism. *British Journal of Educational Technology*, *42*(6), 992–1002. https://doi.org/10.1111/j.1467-8535.2010.01123.x

jalan. (2021). *Pdftotext (2.2.2)*. https://github.com/jalan/pdftotext

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. Cambridge University Press.

Johnson, D. E. (2010). Crossover experiments. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(5), 620–625. https://doi.org/10.1002/wics.109

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Karlovčec, M., Córdova-Sánchez, M., & Pardos, Z. A. (2012). Knowledge component suggestion for untagged content in an intelligent tutoring system. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent Tutoring Systems. ITS 2012. Lecture Notes in Computer Science, vol 7315* (pp. 195–200). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-30950-2_25

Keselman, J. C., Lix, L. M., & Keselman, H. J. (1996). The analysis of repeated measurements: A quantitative research synthesis. *British Journal of Mathematical and Statistical Psychology*, *49*(2), 275–298. https://doi.org/10.1111/j.2044-8317.1996.tb01089.x

Kitagawa, Y., & Komachi, M. (2018). Long short-term memory for Japanese word segmentation. In S. Politzer-Ahles, Y.-Y. Hsu, C.-R. Huang & Y. Yao (Eds.), *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation* (pp. 279–288). Association for Computational Linguistics.

Laska, E., Meisner, M., & Kushner, H. B. (1983). Optimal crossover designs in the presence of carryover effects. *Biometrics*, 1087–1091.

Lin, Y., Feng, S., Lin, F., Zeng, W., Liu, Y., & Wu, P. (2021). Adaptive course recommendation in MOOCs. *Knowledge-Based Systems*, *224*, 107085. https://doi.org/10.1016/j.knosys.2021.107085

Lingard, B. (2018). The Australian curriculum: A critical interrogation of why, what and where to?. *Curriculum Perspectives*, *38*(1), 55–65. https://doi.org/10.1007/s41297-017-0033-7

Liu, Q., Huang, Z., Huang, Z., Liu, C., Chen, E., Su, Y., & Hu, G. (2018). Finding similar exercises in online education systems. In Y. Guo & F. Farooq (Eds.), *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1821–1830). ACM. https://doi.org/10.1145/3219819.3219960

Liu, Z., Wang, J., Gong, S., Lu, H., & Tao, D. (2019). Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6122–6131). IEEE. https://doi.org/10.1109/ICCV.2019.00622

Matayoshi, J., & Lechuga, C. (2020). Automated matching of ITS problems with textbook content. In S. Sosnovsky, P. Brusilovsky, R. Baraniuk & A. Lan (Eds.), *Proceedings of the Second International Workshop on Intelligent Textbooks 2020* (pp. 17–28). CEUR-WS.org.

Meng, R., Han, S., Huang, Y., He, D., & Brusilovsky, P. (2016). Knowledge-based content linking for online textbooks. In *Proceedings of 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 18–25). IEEE. https://doi.org/10.1109/WI.2016.0014

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (Eds.), *Proceedings of Advances in Neural Information Processing Systems 26 (NIPS 2013)* (pp. 3111–3119). ACM.

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2019). Graph-based knowledge tracing: Modeling student proficiency using graph neural network. In P. Barnaghi, G. Gottlob, Y. Manolopoulos, T. Tzouramanis & A. Vakali (Eds.), *Proceedings of WI '19: IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 156–163). ACM. https://doi.org/10.1145/3350546.3352513

Nakayasu, C. (2016). School curriculum in Japan. *The Curriculum Journal*, *27*(1), 134–150. https://doi.org/10.1080/09585176.2016.1144518

National Governors Association. (2010). *Common core state standards*. National Governors Association.

Ninaus, M., & Sailer, M. (2022). Closing the loop–The human role in artificial intelligence for education. *Frontiers in Psychology*, *13*, 956798. https://doi.org/10.3389/fpsyg.2022.956798

Oberfeld, D., & Franke, T. (2013). Evaluating the robustness of repeated measures analyses: The case of small sample sizes and nonnormal data. *Behavior Research Methods*, *45*, 792–812. https://doi.org/10.3758/s13428-012-0281-2

Papagiannopoulou, E., & Tsoumakas, G. (2020). A review of keyphrase extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(2), e1339. https://doi.org/10.1002/widm.1339

Pardos, Z. A., & Dadu, A. (2017). Imputing KCs with representations of problem content and context. In M. Bielikova, E. Herder, F. Cena & M. Desmarais (Eds.), *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (pp. 148–155). ACM. https://doi.org/10.1145/3079628.3079689

Patikorn, T., Deisadze, D., Grande, L., Yu, Z., & Heffernan, N. (2019). Generalizability of methods for imputing mathematical skills needed to solve problems from texts. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren & R. Luckin (Eds.), *Artificial Intelligence in Education. AIED 2019. Lecture Notes in Computer Science, vol 11625* (pp. 396–405). Springer, Cham. https://doi.org/10.1007/978-3-030-23204-7_33

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In C. Cortes, D. D. Lee, M. Sugiyama & R. Garnett (Eds.), *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1* (pp. 505–513). MIT Press.

Porter, R. E., Fusarelli, L. D., & Fusarelli, B. C. (2015). Implementing the common core: How educators interpret curriculum reform. *Educational Policy*, *29*(1), 111–139. https://doi.org/10.1177/0895904814559248

Schneider, S., Nebel, S., Beege, M., & Rey, G. D. (2018). The autonomy-enhancing effects of choice on cognitive load, motivation and learning with digital media. *Learning and Instruction*, *58*, 161–172. https://doi.org/10.1016/j.learninstruc.2018.06.006

Shen, J. T., Yamashita, M., Prihar, E., Heffernan, N., Wu, X., McGrew, S., & Lee, D. (2021). Classifying math knowledge components via task-adaptive pre-trained BERT. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin & V. Dimitrova (Eds.), *Artificial Intelligence in Education. AIED 2021. Lecture Notes in Computer Science, vol 12748* (pp. 408–419). Springer, Cham. https://doi.org/10.1007/978-3-030-78292-4_33

Siddiqi, S., & Sharan, A. (2015). Keyword and keyphrase extraction techniques: A literature review. *International Journal of Computer Applications*, *109*(2), 18–23. https://doi.org/10.5120/19161-0607

Sullivan, L. M., Weinberg, J., & Keaney Jr, J. F. (2016). Common statistical pitfalls in basic science research. *Journal of the American Heart Association*, *5*(10), e004142. https://doi.org/10.1161/JAHA.116.004142

taishi-i. (2020). *Nagisa (0.2.7)*. https://github.com/taishi-i/nagisa

Tian, Z., Zheng, G., Flanagan, B., Mi, J., & Ogata, H. (2021). BEKT: Deep Knowledge Tracing with Bidirectional Encoder Representations from Transformers. In M. M. T. Rodrigo et al. (Eds.), *Proceedings of the 29th International Conference on Computers in Education* (pp. 544–553). Asia-Pacific Society for Computers in Education.

Tong, S., Liu, Q., Huang, W., Huang, Z., Chen, E., Liu, C., Ma, H., & Wang, S. (2020). Structure-based knowledge tracing: An influence propagation view. In *Proceedings of 2020 IEEE International Conference on Data Mining* (pp. 541–550). IEEE. https://doi.org/10.1109/ICDM50108.2020.00063

Tong, W., Tong, S., Huang, W., He, L., Ma, J., Liu, Q., & Chen, E. (2020). Exploiting knowledge hierarchy for finding similar exercises in online education systems. In *Proceedings of 2020 IEEE International Conference on Data Mining* (pp. 1298–1303). https://doi.org/10.1109/ICDM50108.2020.00167

Tsinakos, A., & Kazanidis, I. (2012). Identification of conflicting questions in the PARES system. *International Review of Research in Open and Distributed Learning*, *13*(3), 297–313.

Umer, M., Imtiaz, Z., Ahmad, M., Nappi, M., Medaglia, C., Choi, G. S., & Mehmood, A. (2023). Impact of convolutional neural network and FastText embedding on text classification. *Multimedia Tools and Applications*, *82*(4), 5569–5585. https://doi.org/10.1007/s11042-022-13459-x

Vie, J. J., & Kashima, H. (2019). Knowledge tracing machines: Factorization machines for knowledge tracing. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)* (Vol. 33, No. 01, pp. 750–757). Association for the Advancement of Artificial Intelligence.

Wong, W., Liu, W., & Bennamoun, M. (2012). Ontology learning from text: A look back and into the future. *ACM Computing Surveys*, *44*(4), 1–36. https://doi.org/10.1145/2333112.2333115

Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012). Probase: A probabilistic taxonomy for text understanding. In K. S. Candan, Y. Chen, Ri. Snodgrass, L. Gravano & A. Fuxman (Eds.), *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (pp. 481–492). ACM. https://doi.org/10.1145/2213836.2213891

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems, 135,* 364–381. https://doi.org/10.1016/j.future.2022.05.014

Zhou, Z. H. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press.

Publisher's Note