IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 28, 2018, accepted May 14, 2018, date of publication May 22, 2018, date of current version June 29, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2839607

KnowEdu: A System to Construct Knowledge Graph for Education

PENGHE CHEN[1], YU LU[2], VINCENT W. ZHENG [3], XIYANG CHEN[1], AND BODA YANG[1]

[1]Advanced Innovation Center for Future Education, Beijing Normal University, Beijing 100875, China
[2]Faculty of Education, School of Educational Technology, Beijing Normal University, Beijing 100875, China
[3]Advanced Digital Sciences Center, Singapore 138602

Corresponding author: Yu Lu (luyu@bnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61702039 and in part by the Humanities and Social Sciences Foundation, Ministry of Education of China, under Grant 17YJCZH116.

ABSTRACT Motivated by the vast applications of knowledge graph and the increasing demand in education domain, we propose a system, called *KnowEdu*, to automatically construct knowledge graph for education. By leveraging on heterogeneous data (e.g., pedagogical data and learning assessment data) from the education domain, this system first extracts the concepts of subjects or courses and then identifies the educational relations between the concepts. More specifically, it adopts the neural sequence labeling algorithm on pedagogical data to extract instructional concepts and employs probabilistic association rule mining on learning assessment data to identify the relations with educational significance. We detail all the abovementioned efforts through an exemplary case of constructing a demonstrative knowledge graph for mathematics, where the instructional concepts and their prerequisite relations are derived from curriculum standards and concept-based performance data of students. Evaluation results show that the F1 score for concept extraction exceeds 0.70, and for relation identification, the area under the curve and mean average precision achieve 0.95 and 0.87, respectively.

INDEX TERMS Educational knowledge graph, instructional concept, educational relation, pedagogical data, learning assessment, educational data mining.

I. INTRODUCTION

Knowledge graph serves as an integrated information repository that interlinks heterogeneous data from different domains. Google's Knowledge Graph [1] is such a prominent example that represents real world entities and relations through a multi-relational graph. Existing generic knowledge graphs have demonstrated their advantages in supporting a large number of applications, typically including semantic search (e.g., Google's Knowledge Panel), personal assistant (e.g. Apple's Siri [2]) and deep question answering (e.g., IBM's Watson [3] and Wolfram Alpha [4]). However, those generic knowledge graphs usually cannot well support many domain-specific applications, because they require deep domain information and knowledge. Education is one of such domains, and in this work, we mainly focus on how the knowledge graph for education can be automatically constructed.

In education domain, knowledge graphs are often used for subject teaching and learning in school, where they are also called concept maps. Moreover, popular massive open online course (MOOC) platforms, such as Khan Academy [5], also adopt them for concept visualization and learning resource recommendation. Such knowledge graphs are usually constructed by experienced teachers or domain experts in a manual way. However, such a manual construction process is actually time-consuming and not scalable to large number of concepts and relations. What's more, the number of courses and subjects grows fast on MOOC platforms, so it is much more difficult, or even impossible, to manually construct knowledge graphs for each new course. On the other hand, the manual construction approach is error-prone: according to the pedagogical research, there often exists *expert blind spot* [6], which means expert's cognition and learner's cognition on the same concept often do not well align. In other words, even the domain experts or experienced teachers may easily misunderstand learners' cognitive process. As a result, those manually created knowledge graphs can be suboptimal or misleading for learners.

Motivated by the increasing demands for knowledge graph in education domain and the limitations of manual

construction approach, we propose this system, called *KnowEdu*, to automatically construct educational knowledge graphs that can be used for teaching and learning on school subjects and online courses. Compared to existing generic knowledge graphs, construction of educational knowledge graphs faces several major challenges. Firstly, the desired nodes in educational knowledge graphs represent *instructional concepts* in subjects or courses rather than common real world entities in generic knowledge graphs. An *instructional concept* is a basic concept that learners need to fully understand and grasp (e.g., ''linear equation'' in mathematics or ''photosynthesis'' in biology). The extraction requires authoritative data from education domain and new entity taggers (rather than the traditional taggers, e.g., Person, Location and Organization). Existing NLP tools, including Stanford NLP software [7] and Apache OpenNLP [8], mainly employ the traditional entity taggers, and thus a new *instructional concept extractor* needs to be independently designed and trained. Secondly, the relations between instructional concepts reflect learner's cognitive and educational process, and thus are usually abstract and implicit, e.g., the learning order between two math concepts ''rational numberąś and ''fraction''. Such relations are relatively difficult to identify without proper analysis and modeling on the specific educational data. Comparatively, in the case of generic knowledge graphs, relations among node entities are more detailed and explicit, e.g., the relation between United States and Barack Obama can be explicitly inferred from text semantics. Hence, we particularly select learning assessment and activity data to identify educational relations, because such data can help to capture learners' cognitive and knowledge acquisition processes.

The proposed *KnowEdu* system endeavors to tackle the above challenges, and it principally makes the following key contributions:

- We propose a novel and practical system to automatically construct knowledge graphs for education, which utilizes heterogeneous data, typically including pedagogical data and learning assessment data, to extract instructional concepts and identify significant educational relations.
- Considering the educational purpose of instructional concepts, we propose to apply recurrent neural network models on pedagogical data (e.g., the curriculum standards and textbooks) to accomplish the instructional concept extraction task. To the best of our knowledge, this is the first work of applying neural sequence labeling on entity extraction for education domain.
- We argue that the desired educational relations are substantially different from the traditional relations in generic knowledge graphs that can be properly identified from text corpus. In this work, we particularly utilize the concept-based student assessment data, on which we perform the probabilistic association rule mining to infer the desired relations.
- We demonstrate an exemplary case by constructing a knowledge graph for a subject, with conducting comprehensive and empirical experiments to evaluate the proposed system.

Moreover, the proposed system and the built knowledge graphs can be integrated into intelligent tutoring systems (ITS) [9] and MOOC platforms to support personalized teaching services and adaptive learning solutions.

The rest of this paper is organized as follows: Section II introduces the related work. Section III briefly depicts the proposed system. Section IV and V describe the concept extraction and relation identification processes respectively. Section VI demonstrates an exemplary case and the evaluation results. We conduct discussion in Section VII and conclude in Section VIII.

II. RELATED WORK

Besides Google's knowledge graph, a variety of generic knowledge graphs, such as Freebase [10], Reverb [11], Google Vault [12] and Microsoft's Probase [13], have been constructed by industry and academia, mainly utilizing data collected from the Internet. In educational realm, few studies focus on systematic construction of domain-specific knowledge graphs, but there are some recent works investigating different relation extractions between certain known educational entities: Wang *et al*. [14] extract concepts hierarchies from the textbooks; Chaplot and Koedinger [15] induce structures of multiple units in a course; and Liang *et al*. [16] recover prerequisite relations from university course dependencies. The most relevant work to our research is carried out by Carnegie Mellon University: the researchers utilize observed relations among courses to create a directed concept graph [17], but the relations are assumed to be known in advance. In educational industry, MOOC providers, like Khan Academy [5], have built some dedicated knowledge graphs for their online courses, but most are undirected graphs built by domain experts. All these pioneer studies and efforts demonstrate the increasing interests and pressing needs of knowledge graph construction in education domain.

Entity recognition, as a key step of knowledge graph construction, aims to extract concepts of interest from structured or unstructured data. Among different models for entity recognition task, one main group of models is based on conditional random field (CRF) [18], which has been widely applied in terminology recognition [19] and entity recognition in Chinese [20]. Another popular group is based on neural networks, where different neural architectures are exploited, typically including gated recurrent unit [21] and long short-term memory [22]. Our system mainly adopts these neural network models, and to our best knowledge, it is the first work of applying neural sequence labeling on educational entity extraction.

Relation identification is another key step of knowledge graph construction, and usually leverages on the semantic meaning of data. The distant supervision approach [23] has

been well studied, which can be combined with the attention models [24], [25]. Moreover, the knowledge base completion [26] and refinement [27] techniques can also be used to identify or predict the undetected and missing relations. Compared to generic knowledge graphs, educational relations are usually more abstract and implicit, thus are hard to be directly identified using the popular approaches and data. Our system thus adopts data mining techniques (e.g., probabilistic association rule mining) on learning assessment data to accomplish the task of educational relation identification.

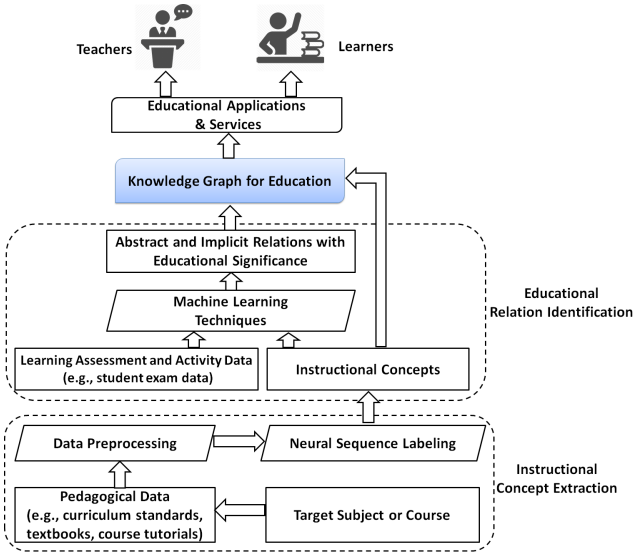


FIGURE 1. Block Diagram of the KnowEdu System.

III. SYSTEM OVERVIEW

The block diagram of proposed *KnowEdu* system is illustrated in Figure 1. Its hierarchical architecture mainly consists of two modules: *Instructional Concept Extraction Module* and *Educational Relation Identification Module*. Their general descriptions are given as follows:

- *Instructional Concept Extraction Module*: the main objective of this module is to extract instructional concepts for a given subject or course. This module mainly utilizes the pedagogical data, typically including the curriculum standards, textbooks and course tutorials, which are usually for teaching purposes and collected from the education domain. They may need to be firstly converted from printed documents into machine-readable text format. After the data selection and format conversion, named entity recognition techniques, especially neural sequence labeling, can be deployed to extract instructional concepts. The key outputs of this module are the extracted concepts which are cornerstones of the built knowledge graphs.
- *Educational Relation Identification Module*: the main objective of this module is to identify the educational relations that interlink instructional concepts to help the learning and teaching process directly. Since educational relations are more implicit and abstract, this module mainly utilizes the learning assessment and activity data that can reflect learners' cognitive and knowledge acquisition process, and adopts the latest data mining techniques, such as the probabilistic association rule mining. Finally, those identified relations connect instructional concepts to formulate the desired knowledge graphs for education, which can be used to support a variety of applications and services for both learners and teachers.

In the following two sections, we will elaborate our design for these two modules respectively.

IV. INSTRUCTIONAL CONCEPT EXTRACTION

A. DATA SOURCE AND PREPROCESSING

As mentioned earlier, the desired nodes in our educational knowledge graphs represent instructional concepts that should be mastered by learners. The input data is thus mainly collected from education domain and pedagogical practices, such as curriculum standards, textbooks and course tutorials. The input data can be in different formats, typically including printed text, audio and video. Thus conversion into machine-readable format may be needed and a variety of format conversion techniques can be applied. For example, the optical character recognition (OCR) [28] technique can be used to handle the printed documents (e.g., textbooks or course tutorials). The speech to text (STT) [29] technique can be employed to manage the audio data (e.g., teacher's voice records during lectures). After this pre-processing step of transforming pedagogical data into machine-readable text, the proposed system can perform the instructional concept extraction.

B. CONCEPT EXTRACTION

Given converted machine-readable text, this concept extraction task can be naturally regarded as a word sequence labeling problem. For example, given a word sequence "*understand the meaning of rational number*", the main objective is to annotate each word with a label specifying whether the word is part of an instructional concept.

We thus define three labels for concept extraction: 1) *B-CP* (meaning "beginning of a concept"); 2) *I-CP* (meaning "inside a concept") and 3) *O* (meaning "outside a concept"). Both *B-CP* and *I-CP* represent instructional concepts. Table 1 illustrates a correct concept extraction result for the given word sequence, where $x = \{x_1, x_2, \ldots, x_T\}$ denotes the input word sequence and $y = \{y_1, y_2, \ldots, y_T\}$ denotes the corresponding output labels. Both x and y are with length T.

TABLE 1. An Example of concept extraction by word sequence labeling.

y	O	O	O	O	B-CP	I-CP
x	Understand	the	meaning	of	rational	number

Different from common entities in generic knowledge graphs, instructional concepts are usually well

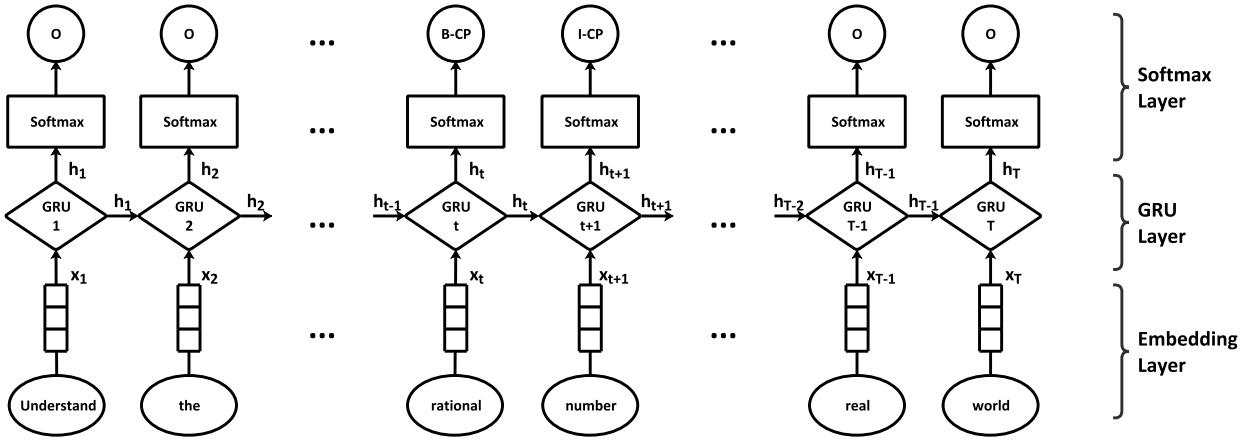


FIGURE 2. Architecture of the GRU Network.

defined or described with little ambiguity in the pedagogical data sources, especially in curriculum standards or textbooks. They usually co-occur with some *signal* words, such as "*know*", "*understand*" and "*definition*", which can be leveraged to capture the desired concepts. Considering the sequential nature of educational text data and their internal word dependencies, we firstly describe the traditional CRF model, and then propose neural network models to accomplish this task.

1) CRF MODEL

Briefly speaking, CRF model can be regarded as the conditional probability distributions on an undirected graphical model, which is commonly used to label the observation sequence data. Since the sequential structure of the input text data and output labels, the CRF model with a linear-chain structure can be utilized. Specifically, given x and y as defined earlier in this section, the model defines a distribution $p(y|x)$ that takes the form

$$p(y|x) = \frac{1}{Z(x)} \exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x, t)), \quad (1)$$

where $\lambda = \{\lambda_k\} \in \Re^K$ is a parameter vector, $F = \{f_k(y_{t-1}, y_t, x, t)\}_{k=1}^{K}$ is a set of real-valued feature functions, t is the position in word sequence, and $Z(x)$ is an input-independent normalization function defined as

$$Z(x) = \sum_{y} \exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x, t)). \quad (2)$$

Simply speaking, feature function set F mainly captures the state transition from y_{t-1} to y_t and their dependencies on input sequence x. For example, one feature function can be defined as

$$f_k(y_{t-1}, y_t, x, t) = \begin{cases} 1 & y_{t-1} = B\text{-}CP \text{ and } y_t = I\text{-}CP \\ & \text{and } x_t = \text{"}number\text{"} \\ 0 & otherwise. \end{cases} \quad (3)$$

Given a set of feature functions and labeled training data sequences, the CRF model training process is to find the model parameter vector λ using maximum likelihood estimation. In practice, we employ the classic L-BFGS algorithm [30] to learn λ, and adopt Viterbi algorithm [31] to infer the optimal label sequence y. Both algorithms have been well studied and the details can be found in [18] and [32].

As shown in equation 3, the traditional CRF model requires a large amount of efforts on explicitly engineering features. To overcome such a limitation, we propose to adopt the neural network model, which is able to learn features automatically.

2) NEURAL NETWORK MODEL

Among different neural network models, we choose recurrent neural network (RNN), as it cannot only obviate the feature engineering step but also well capture dependencies in sequential data. Specifically, we adopt gated recurrent unit (GRU) network [21], and the simplified architecture of the proposed model is illustrated in Figure 2.

The proposed model mainly consists of three layers: embedding layer, GRU layer and softmax layer. The embedding layer utilizes the word2vec [33] algorithm to generate embedding vector for each word, in which the text corpus is built upon input data. The GRU layer is a series of GRU units, which control the addition and removal of information through a carefully designed gate structure, typically including update gate and output gate. Figure 3 shows the architecture of the GRU unit. In the GRU layer, the model recurrently utilizes the GRU unit on each input word, with following implementation:

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (4)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (5)$$
$$\tilde{h}_t = tanh(W_h x_t + U_h(r_t * h_{t-1})) \quad (6)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (7)$$

where x_t and h_t are the input and output of GRU unit, where t is the position in word sequence. $W_z, W_r, W_h, U_z, U_r, U_h$

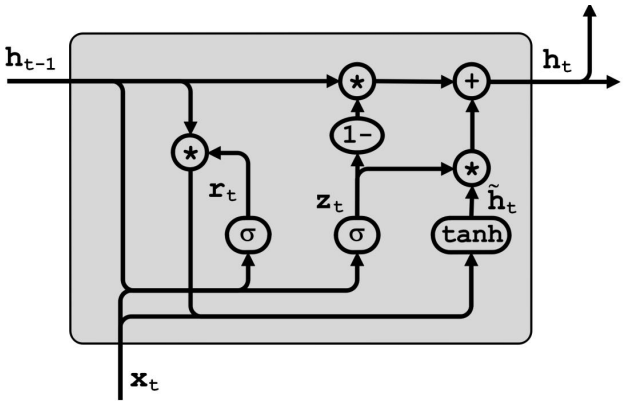


FIGURE 3. Architecture of GRU unit.

are weight matrices, $\sigma()$ is the element-wise sigmoid function, and "$*$" is the element-wise product. Note that GRU can be regarded as a variant of the classic long short-term memory (LSTM) [22] network, with simplifying LSTM unit by combining LSTM's forget gate and input gate. More details about the GRU unit can refer to [34] and [35].

The softmax layer applies the softmax function on each GRU unit output h_t to produce a L-dimensional vector, which gives the probability of each label defined in the concept extraction task. Given a weighting vector ω and the current GRU output h_t, the predicted probability of the j^{th} label is

$$P(y = j|h_t) = \frac{e^{h_t^T \omega_j}}{\sum_{l=1}^{L} e^{h_t^T \omega_l}}, \tag{8}$$

where L is the number of label types. L is currently set to 3, as we define three labels in our model, namely *B-CP*, *I-CP* and *O*.

Given the labeled text data sequences, the neural network model training process is to find proper weight matrices for the GRU layer and weighting vector for the softmax layer. In practice, the model is trained by minimizing the cross-entropy loss [36], where the Adam algorithm [37] and the back-propagation through time (BPTT) [38] algorithm are utilized.

Besides the GRU units, we also attempt to adopt the LSTM units under the same architecture shown in Figure 2. All the evaluation results on the GRU-based model, LSTM-based model and CRF model summarize in section VI.

V. EDUCATIONAL RELATION IDENTIFICATION

A. RELATION TYPE AND DATA SELECTION

As mentioned earlier, the main task of this module is to identify the relations interlinking instructional concepts. These relations can be any logic connections that can aid learning and teaching process directly. In the education domain, a number of relations are critical to teachers and learners, such as inclusion relation, causal relation, progressive relation and prerequisite relation. In this section, we mainly focus on identifying prerequisite relation, while briefly introduce inclusion relation.

1) PREREQUISITE RELATION

Among the above mentioned relations, prerequisite relation is the most implicit one and thus relatively hard to identify. Prerequisite relation is in accordance with the knowledge space theory [39], which argues that prerequisite exists as a natural dependency between concepts in human cognitive process. Specifically, a prerequisite relation from concept A to concept B means that a learner should master concept A first before proceeding to concept B.

Hence, the identified prerequisite relations can help teachers design proper pedagogical strategy and help learners study effectively. For example, when students encounter difficulties in learning math concept "quadratic equation", teachers can utilize prerequisite relations to identify the possible reasons for this learning obstacle. In addition, students can also utilize prerequisite relations to determine their concept learning or revision order for subjects or courses. Furthermore, prerequisite relations can also be used in today's MOOC platforms and online tutoring systems to support adaptive learning and personalized teaching.

2) DATA SELECTION FOR PREREQUISITE RELATION IDENTIFICATION

To identify a prerequisite relation between key concepts, educators traditionally use a specific strategy based on learners' performance: when concept γ is mastered by learners, all its prerequisites should have also been mastered by learners; meanwhile, when any prerequisite of concept γ has not been mastered by learners, it is hard for learners to master γ.

Inspired by the above strategy, we can utilize the learning assessment data to identify prerequisite relations automatically. In fact, such data can be easily collected from MOOC or online tutoring platforms on a large number of learners and in the form of concept-based pretest or posttest results.

3) INCLUSION RELATIONS

Besides the prerequisite relation, other relations, such as inclusion relation, are also important for education. Inclusion relation indicates one concept belongs to another one, which is commonly used by educators to build concept hierarchies for courses or subjects. Comparing with prerequisite relation, inclusion relation is relatively easy to identify, as such information is usually preserved in the original structure of textbooks or tutorials. Our system can thus adopt similarity-based method to extract the hierarchical structure from the table of content (TOC) of textbooks, and then specify inclusion relations among concepts.

In the rest of this section, we mainly introduce how to identify prerequisite relation using students' performance data. Specifically, this system adopts the probabilistic association rule mining algorithm which conveniently implements the above mentioned strategy for prerequisite relation identification and well handles uncertainties in students' performance data.

B. PREREQUISITE RELATION IDENTIFICATION

1) PRELIMINARY OF ASSOCIATION RULE MINING

Association rule mining [40] is a simple, yet effective data mining technique for discovering interesting relations hidden in large databases. It is originally used to discover and analyze the relations between sold products in supermarkets. Denote I as a set of items and D as a set of transactions, each transaction in D contains a subset of items in I. The uncovered relations can be represented in the form of association rules, each of which is an implication of the form $X \Rightarrow Y$, where $X, Y \subset I$. Two key measures, namely *support* and *confidence*, are commonly used for determining association rules. The *support* of $X \Rightarrow Y$, denoted as $supp(X \Rightarrow Y)$, is the co-occurrence of X and Y within data. The *confidence* of $X \Rightarrow Y$, denoted as $conf(X \Rightarrow Y)$, is the percentage of transactions in data containing X that also containing Y. Given threshold values for the two measures, denoted as *minsupp* and *minconf*, an association rule $X \Rightarrow Y$ can be considered interesting and strong if the following condition is satisfied:

$$supp(X \Rightarrow Y) \geq minsupp \quad \text{AND} \quad conf(X \Rightarrow Y) \geq minconf. \tag{9}$$

Assuming instructional concepts as items and students' performance data as transactions, association rule mining is a natural way to implement the educators' strategy for identifying prerequisite relations.

2) PREREQUISITE RELATION IDENTIFICATION

As mentioned earlier, if concept i is a prerequisite of concept j, learners who do not master i very likely do not master j either. Meanwhile, learners who master concept j very likely master concept i too. In the perspective of association rule mining, such a prerequisite relation from concept i to concept j is deemed to exist if the following pair of association rules can be determined simultaneously:

$$S_j \Rightarrow S_i \quad AND \ \overline{S_i} \Rightarrow \overline{S_j}, \tag{10}$$

where S_i and S_j denote learners have mastered concepts i and j respectively, $\overline{S_i}$ and $\overline{S_j}$ denote learners have not mastered concepts i and j yet. Accordingly, the system needs to estimate the interestingness of these two association rules with the knowledge state information (i.e., whether concepts i and j are mastered or not) from multiple learners.

However, a learner's mastery on a concept usually cannot be directly observed because knowledge state is typically a latent variable. It is a common and feasible way to infer one learner's knowledge state using his or her academic (exam) performance data. However, uncertainties exist in exam results, especially due to slipping (i.e., making an error despite having mastered that concept) and guessing (i.e., giving a right answer despite not understanding that concept), a learner's mastery on one concept is usually regarded as a random variable. Hence, probabilistic association rule mining [41] is required, which is an extension of the association rule mining for handling uncertainties in data. Specifically, given the *support* and *confidence* measures in probabilistic data as random variables, the deterministic association rule $S_j \Rightarrow S_i$ is formulated as $P(S_j \Rightarrow S_i)$, and the rule holds if $P(S_j \Rightarrow S_i)$ is larger than a given threshold *minprob*:

$$P(S_j \Rightarrow S_i) \geq minprob. \tag{11}$$

By taking *support* and *confidence* into account, we instantiate equation (11) as

$$P\{supp(S_j \Rightarrow S_i) \geq minsupp \quad AND \ conf(S_j \Rightarrow S_i) \geq minconf\} \geq minprob. \tag{12}$$

According to equation (10), in order to determine a prerequisite relation from concept i to j, both rules need to be held, so we require:

$$P(S_j \Rightarrow S_i) * P(\overline{S_i} \Rightarrow \overline{S_j}) \geq minprob. \tag{13}$$

To calculate the probability in equation (13), we adopt the p-Apriori [41] algorithm which is specifically designed for probabilistic association rule mining. Note that it is possible that a pair of concepts are determined to be the prerequisite of each other, and such a symmetric relation means the two concepts need to be learned together.

In short, we mainly introduce how to employ the probabilistic association rule mining on students' exam data to identify prerequisite relations, while different techniques and data can be further introduced and applied for different relation identification problems. To demonstrate the performance of both concept extraction and relation identification in our system, we construct an exemplary knowledge graph for a subject and conduct the comprehensive evaluations using the real-world educational data.

VI. EXEMPLARY CASE AND SYSTEM EVALUATION

To evaluate the proposed *KnowEdu* system, we construct an exemplary knowledge graph for mathematics, which demonstrates the instructional concept extraction and the educational relation identification processes. Comprehensive evaluations are conducted to assess the system performance.

A. CONCEPT EXTRACTION

1) DATASET AND PREPROCESSING

As mentioned earlier, different from generic knowledge graphs, datasets for instructional concept extraction are usually from the pedagogical and educational sources, such as curriculum standards, textbooks and course manuals. These materials are usually used as official guidance for teaching and pedagogical practice. We choose the national curriculum standards of mathematics for primary and secondary schools, which are published by the ministry of education of China, as the main data source.

For the data preprocessing step, the system firstly uses Tika [42] to extract text from the official version of the curriculum standards, and then conducts sentence segmentation based on the specific symbols for sections, paragraphs and punctuation. Moreover, non-text information such as

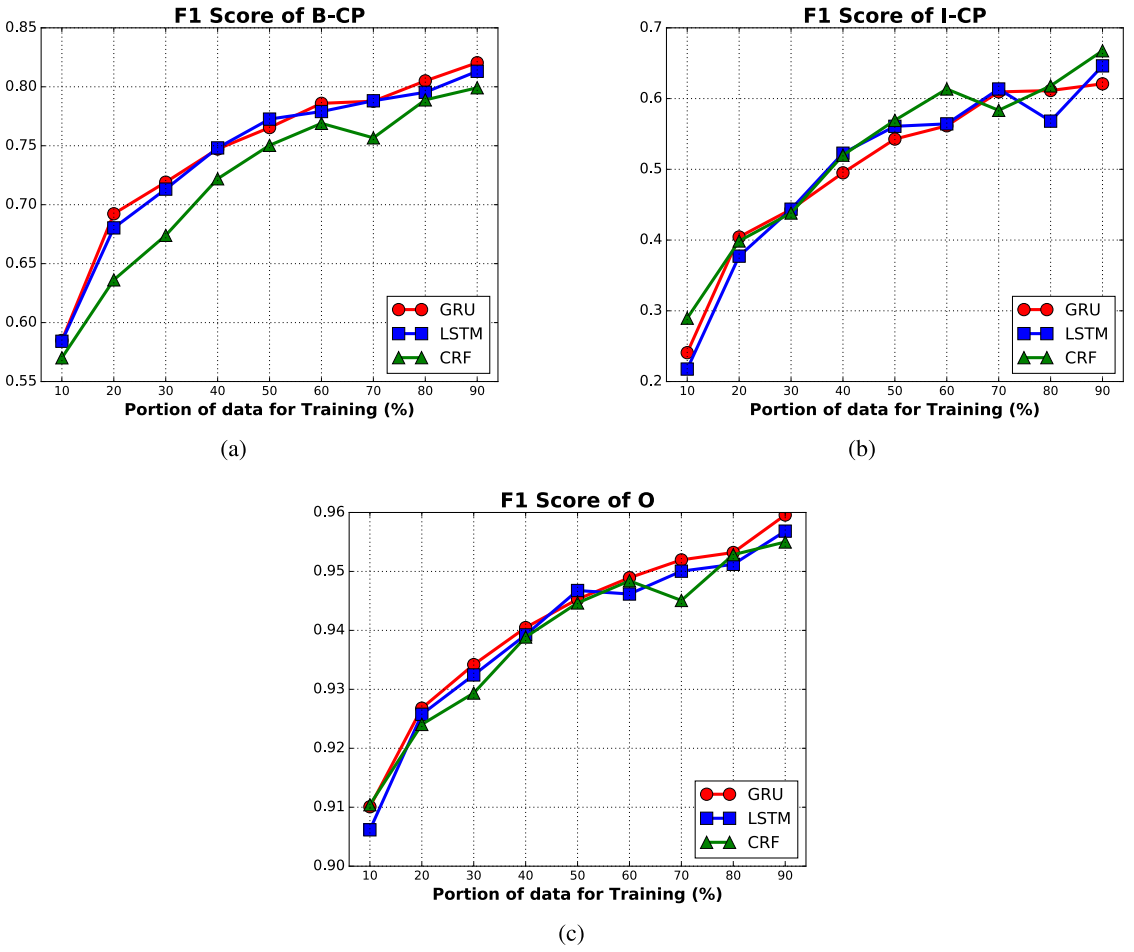


FIGURE 4. Evaluation results for concept extraction. (a) B-CP: Begin-Concept. (b) I-CP: Inside-Concept. (c) O:Non-Concept.

images and table boundaries are automatically removed. Subsequently, the system conducts word segmentation utilizing *ICTCLAS* [43] which is an open source library for Chinese word segmentation. In the end, 1,847 sentences and 36,697 words are obtained from the raw dataset.

2) EVALUATION FOR CONCEPT EXTRACTION

To obtain the ground truth for model evaluation, we invite two domain experts from Beijing Normal University, who were involved in drafting the national curriculum standards, to label all the instructional concepts. Totally 4,251 words are labeled as *B-CP* and 969 words are labeled as *I-CP*. The two experts achieved a high consistency between their labeling and the corresponding kappa value is 0.945.

To fully evaluate the model performance, we randomly split the dataset into two parts for training and testing respectively, and we gradually increase the percentage of training examples from 10% to 90%. We repeat each experiment for 10 times and report the average results in evaluation.

For the proposed GRU-based and LSTM-based neural network models, the dimension of each unit output h_t is set to 128. For the Adam algorithm used for training the models, its initial learning rate and its iteration number is set to 0.01 and 1000 respectively with parameters $\beta 1 = 0.9$, $\beta 2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$. For the CRF model, L-1 regularization is used for conducting the maximum likelihood estimation. In the L-BFGS algorithm used for training the CRF model, the number of iterations is set to 50 with parameters $\epsilon = 1 \times 10^{-5}$ and $\delta = 1 \times 10^{-5}$ respectively.

Figure 4 summarizes the evaluation results, where F1 scores of the three models (namely GRU, LSTM and CRF models) are reported for each of the three types of labels (namely *B-CP*, *I-CP* and *O*). In general, all the F1 score curves grow with the portion of data for training increasing. Figure 4a shows that for the *B-CP* (beginning of concept) extraction, both the GRU-based and LSTM-based neural network models outperform the CRF model, and no significant difference between the two neural network models. Meanwhile, Figure 4b and Figure 4c illustrate that, for the *I-CP* and Non-Concept extraction, all the three models have similar performance. Moreover, by comparing Figure 4a with Figure 4b, we see that for all the three models, the *I-CP* extraction usually has lower F1 score than *B-CP*, which indicates that the *I-CP* extraction is more difficult than

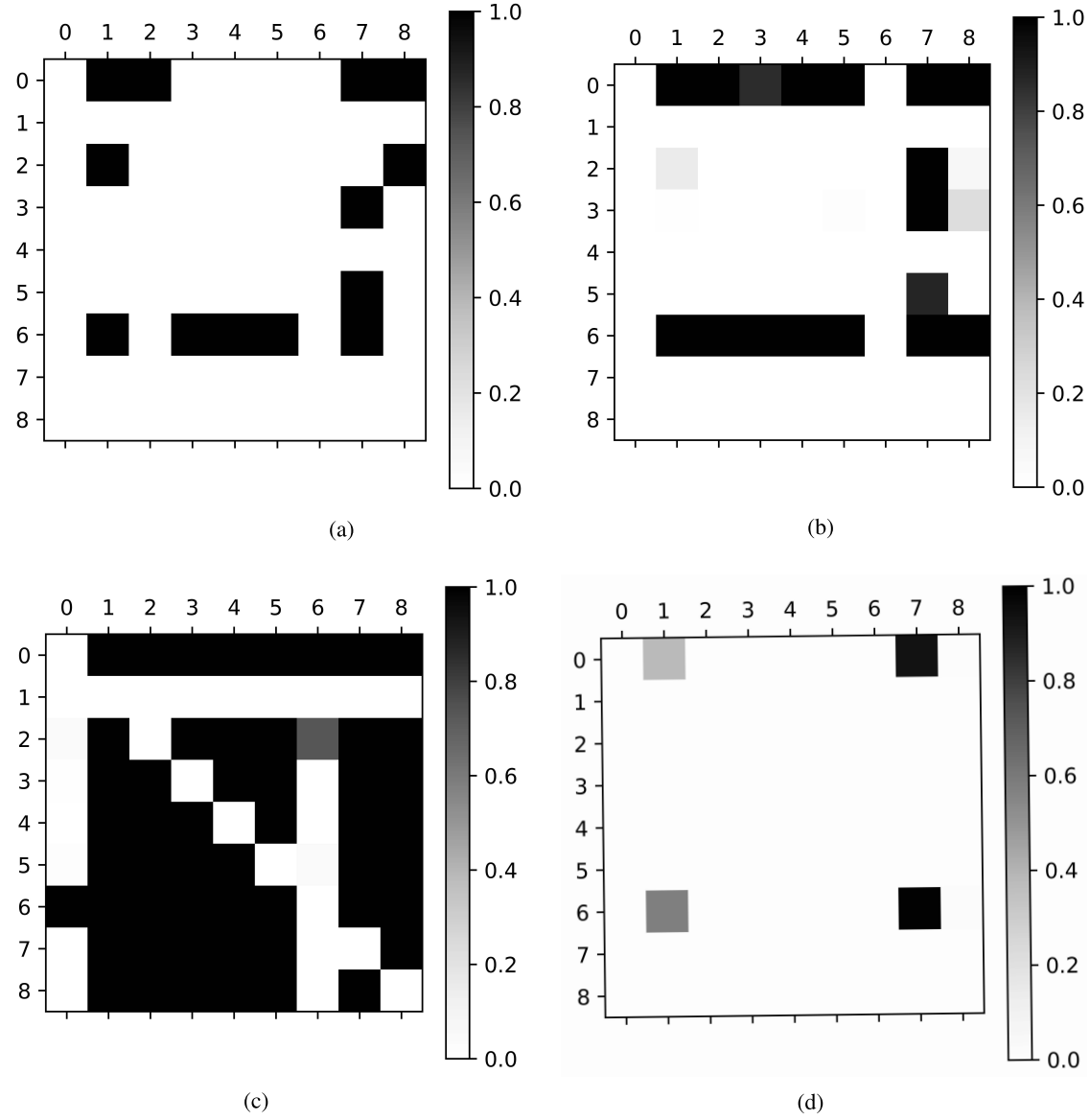


FIGURE 5. A Comparison of relation identification results. (a) Ground truth. (b) minsupp = 400, minconf = 0.4. (c) minsupp = 1000, minconf = 0.7.

B-CP extraction. It is reasonable, as a correct *I-CP* labeling conditions on a correct *B-CP* labeling. Table 2 gives the precision, recall and F1 score on *B-CP* extraction of the three models, when 50% data are used for training. We see that for this case, the precision of the CRF model is even slightly higher than the two neural network models, but its low recall eventually results the worst F1 score among the three models. In short, the above evaluation results verify the feasibility and effectiveness of proposed models for the instructional concept extraction task.

TABLE 2. System Performance on B-CP.

	Precision	Recall	F1-Score
CRF	0.81	0.70	0.75
LSTM	0.79	0.76	0.77
GRU	0.77	0.76	0.76

B. RELATION IDENTIFICATION

1) DATASET

The system collects students' exam data from unit tests and uses it to identify the prerequisite relations between instructional concepts. Each unit test consists of multiple questions for the same concept, and the corresponding score rate is accordingly used to represent the knowledge state of examinees on that concept. To ensure the statistical

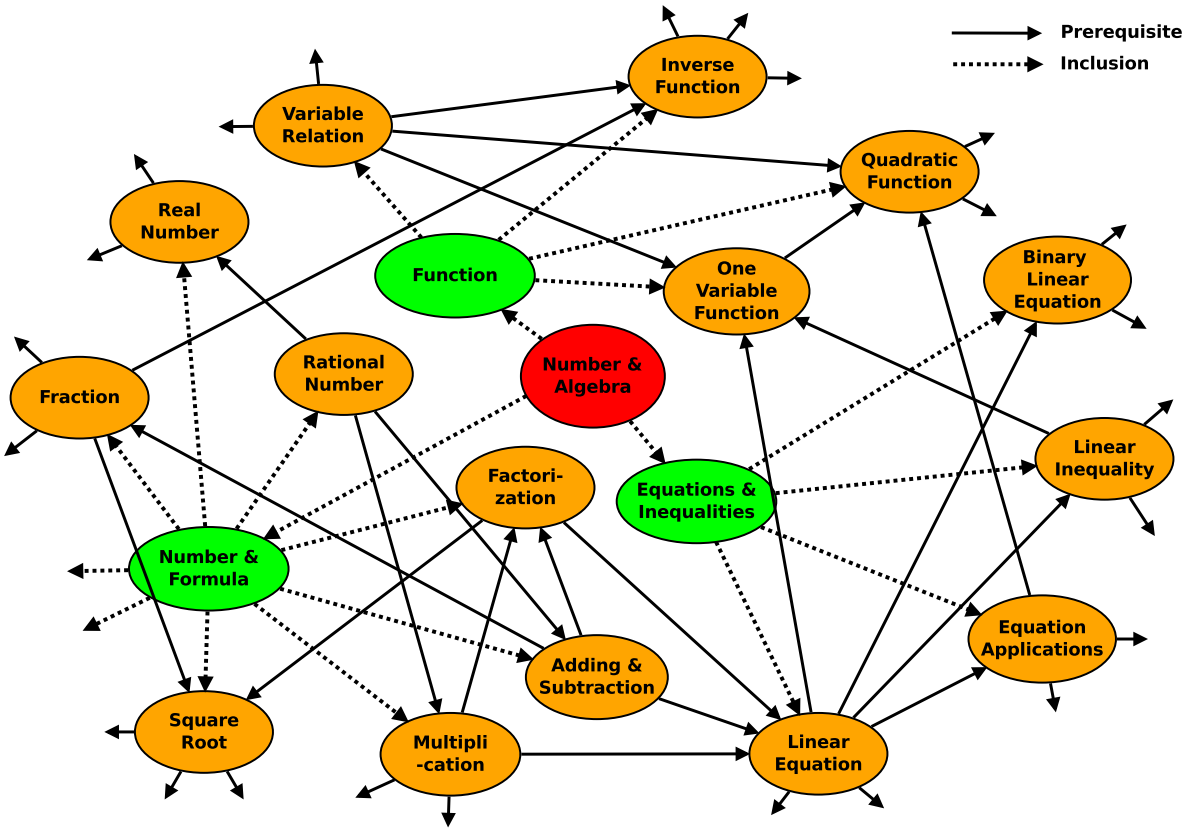


FIGURE 6. A snapshot of built knowledge graph for mathematics.

significance of collected data, the official online assessment platform [44] is used to collect exam data from 31 middle schools and totally 4,488 students in Beijing of China, who are mainly the 7^{th} grade middle school students.

2) EVALUATION

To obtain the ground truth of the prerequisite relations between the instructional concepts, two domain experts are invited to annotate the relations: if a prerequisite relation exists from concept A to concept B, we call it a positive relation, while if no prerequisite relation exists between them, we call it a negative relation. A positive relation is determined only when both experts annotate it positive, and the kappa value is 0.896.

Similar to the traditional approach in educational data mining, the system uses score rate as the estimated probability of learners' mastery on each concept. Accordingly, each concept is regarded as one item, and the estimated knowledge states from 4,488 students are regarded as 4,488 transactions in the context of association rule mining. For each prerequisite relation candidate, the system calculates its probability of being positive, where the key parameters are the two thresholds *minsupp* and *minconf*.

To properly determine the two key parameters and evaluate the performance of this probabilistic association rule mining algorithm, we use Area Under ROC curve (AUC) [45] and macro-averaged Mean Average Precision (MAP) [46] as the main metrics. Tables 3 and 4 summarize the AUC and MAP values for different pairs of *minsupp* and *minconf* respectively. We see that the *minconf* and *minsupp* pairs (0.6, 600) and (0.6, 800) have a significant higher AUC and MAP than other pairs. Figure 5 further shows a comparison of relation identification results in the form of heat map, where 9 randomly selected concepts and 3 groups of typical parameters are used. Figure 5a shows the ground truth of all the prerequisite relations between the 9 concepts. The $(i, j)^{th}$ entry denotes a prerequisite relation from concept j to concept i. The color in heat maps indicates the probability of a prerequisite relation according to the left part of equation (13), where a darker color means a higher probability.

TABLE 3. AUC with different parameter pairs.

AUC		*minsupp*					
		400	600	800	1000	1200	1400
minconf	0.3	0.623	0.690	0.645	0.483	0.510	0.477
	0.4	0.689	0.756	0.722	0.518	0.525	0.478
	0.5	0.868	0.874	0.838	0.623	0.559	0.493
	0.6	0.953	**0.953**	**0.954**	0.803	0.692	0.546
	0.7	0.836	0.836	0.836	0.840	0.840	0.688
	0.8	0.850	0.850	0.850	0.853	0.858	0.756
	0.9	0.735	0.735	0.735	0.735	0.747	0.682

We can see clearly that Figure 5b is much more similar to the ground truth in Figure 5a, when *minconf* = 0.6 and *minsupp* = 800 are used as the threshold parameters in the system. Meanwhile, we see that Figure 5c and Figure 5d show the worse performance, which is mainly due to the improper threshold parameters and results a low precision and a low recall respectively.

TABLE 4. MAP with different parameter pairs.

MAP		*minsupp*					
		400	600	800	1000	1200	1400
minconf	0.3	0.566	0.627	0.627	0.505	0.521	0.535
	0.4	0.594	0.656	0.661	0.518	0.525	0.535
	0.5	0.737	0.802	0.727	0.564	0.534	0.538
	0.6	0.877	**0.877**	**0.863**	0.778	0.595	0.550
	0.7	0.818	0.818	0.818	0.816	0.766	0.660
	0.8	0.823	0.823	0.823	0.822	0.814	0.742
	0.9	0.788	0.788	0.788	0.788	0.801	0.785

Figure 6 further demonstrates a snapshot of the built knowledge graph in English, where each circle represents one concept and two key relations, namely the prerequisite relation and inclusion relation, are marked on the graph using solid line and dash line respectively.

VII. DISCUSSION

In general, the proposed system mainly focuses on instructional concepts and their internal relations, other educational components, which are also significant to learners and teachers, can also be considered and included as new categories of entities, such as learning resources and pedagogical objectives. Accordingly, new relations between such novel entities need to be properly defined and identified. Moreover, the current design only aims to identify the intra-concept relations within one subject or one course, while the inter-course and inter-subject relations can be further explored.

For the instructional concept extraction, both CRF model and neural network models require a relatively large number of training data to achieve a high performance. Some semi-supervised learning models can be considered to utilize unlabeled data for training, where the Wikipedia and other online encyclopedia data may serve as an important role in the entity extraction step. Moreover, constructing a knowledge graph for mathematics or other science subjects is relatively easy, but such a task would be harder for the subjects or courses falling into languages and literatures. It is probably caused by the complexity in human emotion and the ambiguity in human expression, which accordingly impose new challenges for both entity extraction and relation identification tasks.

VIII. CONCLUSION

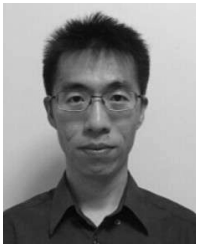
We have introduced and implemented the *KnowEdu* system, which automatically constructs knowledge graph for education. It extracts instructional concepts and implicit educational relations from heterogenous data sources, mainly including standard curriculum data and learning assessment data. For the instructional concept extraction, neural network models are employed, and for the prerequisite relation identification, the probabilistic association rule mining is introduced. We demonstrate the promise of this system via building a knowledge graph for mathematics, where the F1 score on B-CP extraction exceeds 0.75 when 50% data for training, and the AUC achieves 0.95 for the prerequisite relation identification.

On a broader canvas, this *KnowEdu* system has demonstrated the feasibility and effectiveness to automatically construct dedicated knowledge graphs for different subjects or courses. A variety of personalized teaching and learning services, such as online diagnosis of learning obstacles and intelligent recommendation of learning resources, can be developed using such dedicated knowledge graphs, especially for the next generation of MOOC platforms.

REFERENCES

[1] A. Singhal, "Introducing the knowledge graph: Things, not strings," Official Google Blog, CA, USA, Tech. Rep., 2012.
[2] Apple Inc. (2017). *Apple Siri*. [Online]. Available: https://www.apple.com/ios/siri/
[3] (2017). *IBM Waston*. [Online]. Available: https://www.ibm.com/watson/
[4] (2017). *Wolfram Alpha*. [Online]. Available: https://www.wolframalpha.com/
[5] (2017). *Khan Academy*. [Online]. Available: http://en.www.khanacademy.org/
[6] M. J. Nathan and A. Petrosino, "Expert blind spot among preservice teachers," *Amer. Edu. Res. J.*, vol. 40, no. 4, pp. 905–928, 2003.
[7] (2017). *Stanford Natural Language Processing*. [Online]. Available: https://nlp.stanford.edu/software/
[8] (2017). *Apache OpenNLP*. [Online]. Available: https://opennlp.apache.org/
[9] E. Wenger, *Artificial Intelligence and Tutoring Systems: Computational and Cognitive Approaches to the Communication of Knowledge*. San Mateo, CA, USA: Morgan Kaufmann, 2014.
[10] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2008, pp. 1247–1250.
[11] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*. Edinburgh, Scotland: Association for Computational Linguistics, 2011, pp. 1535–1545.
[12] X. Dong *et al.*, "Knowledge vault: A Web-scale approach to probabilistic knowledge fusion," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 601–610.
[13] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2012, pp. 481–492.
[14] S. Wang *et al.*, "Concept hierarchy extraction from textbooks," in *Proc. ACM Symp. Document Eng.*, 2015, pp. 147–156.
[15] D. Chaplot and K. R. Koedinger, "Data-driven automated induction of prerequisite structure graphs," in *Proc. Edu. Data Mining (EDM)*, 2016, pp. 318–323.
[16] C. Liang, J. Ye, Z. Wu, B. Pursel, and C. L. Giles, "Recovering concept prerequisite relations from university course dependencies," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4786–4791.
[17] H. Liu, W. Ma, Y. Yang, and J. Carbonell, "Learning concept graphs from online educational data," *J. Artif. Intell. Res.*, vol. 55, pp. 1059–1090, Apr. 2016.
[18] J. Lafferty *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn., (ICML)*, vol. 1, 2001, pp. 282–289.
[19] R. Leaman, C.-H. Wei, and Z. Lu, "tmChem: A high performance approach for chemical named entity recognition and normalization," *J. Cheminform.*, vol. 7, no. 1, p. S3, 2015.

[20] A. L.-F. Han, D. F. Wong, and L. S. Chao, "Chinese named entity recognition with conditional random fields in the light of chinese characteristics," in *Language Processing and Intelligent Information Systems*. Springer, 2013, pp. 57–68.
[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. (2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling." [Online]. Available: https://arxiv.org/abs/1412.3555
[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[23] C. Quirk and H. Poon. (2016). "Distant supervision for relation extraction beyond the sentence boundary." [Online]. Available: https://arxiv.org/abs/1609.04873
[24] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proc. ACL*, vol. 1, 2016, pp. 2124–2133.
[25] G. Ji, K. Liu, S. He, and J. Zhao, "Distant supervision for relation extraction with sentence-level attention and entity descriptions," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 3060–3066.
[26] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 926–934.
[27] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic Web*, vol. 8, no. 3, pp. 489–508, 2017.
[28] S. M. Beitzel, E. C. Jensen, and D. A. Grossman, "Retrieving OCR text: A survey of current approaches," in *Proc. Symp. Document Image Understand. Technol. (SDUIT)*, 2003, pp. 1–6.
[29] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.
[30] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, no. 1, pp. 503–528, 1989.
[31] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. TIT-13, no. 2, pp. 260–269, Apr. 1967.
[32] C. Sutton *et al.*, "An introduction to conditional random fields," *Found. Trends Mach. Learn.*, vol. 4, no. 4, pp. 267–373, 2012.
[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[34] K. Cho *et al.* (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation." [Online]. Available: https://arxiv.org/abs/1406.1078
[35] K. Yao, T. Cohn, K. Vylomova, K. Duh, and C. Dyer. (2015). "Depth-gated recurrent neural networks." [Online]. Available: https://128.84.21.199/abs/1508.03790v2
[36] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, 2005.
[37] D. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980
[38] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct. 1990.
[39] J.-P. Doignon and J.-C. Falmagne, "Spaces for the assessment of knowledge," *Int. J. Man-Mach. Stud.*, vol. 23, no. 2, pp. 175–196, 1985.
[40] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*, 1994, pp. 487–499.
[41] L. Sun, R. Cheng, D. W. Cheung, and J. Cheng, "Mining uncertain data with probabilistic guarantees," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 273–282.
[42] (2017). *Apache Tika*. [Online]. Available: http://tika.apache.org/
[43] (2017). *ICTCLAS*. [Online]. Available: http://ictclas.nlpir.org/
[44] (2017). *Smart Learning Partner*. [Online]. Available: http://slp.101.com/
[45] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[46] K. Kishida, "Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments," Nat. Inst. Inform., Tokyo, Japan, Tech. Rep., 2005.

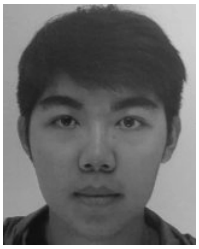


PENGHE CHEN received the B.S. and Ph.D. degrees in computer science from the National University of Singapore. He was with the Advanced Digital Sciences Center, Singapore. He is currently a Principle Researcher with the Advanced Innovation Center for Future Education, Beijing Normal University, China. His research interests include knowledge graph construction, data mining, and learning analytics.

YU LU received the Ph.D. degree from the National University of Singapore in 2012. He was a Research Scientist with the Agency for Science, Technology and Research, Singapore. He is an Associate Professor with the Faculty of Education, School of Educational Technology, Beijing Normal University, where he also serves as the Director of the Artificial Intelligence Laboratory, Advanced Innovation Center for Future Education. His recent research interests include educational data mining, learning analytics, and educational robotics.

VINCENT W. ZHENG received the Ph.D. degree in computer science from The Hong Kong University of Science and Technology in 2011. He is currently a Senior Research Scientist with the Advanced Digital Sciences Center, Singapore, and a Research Affiliate with the University of Illinois at Urbana–Champaign. His research interests include graph mining, information extraction, ubiquitous computing, and machine learning.

XIYANG CHEN received the M.S. degree in software engineering from Peking University in 2017. He is currently a Research Engineer with the Advanced Innovation Center for Future Education, Beijing Normal University. His research interests include knowledge graph construction and data mining.



BODA YANG received the M.S. degree in computer science from the University of St Andrews in 2017. He is currently a Research Engineer with the Advanced Innovation Center for Future Education, Beijing Normal University. His research interests include learning analytics and educational data mining.