Predicting Prerequisite Relations for Unseen Concepts

Yaxin Zhu and Hamed Zamani
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
{yaxinzhu, zamani}@cs.umass.edu

Abstract
Concept prerequisite learning (CPL) plays a key role in developing technologies that assist people to learn a new complex topic or concept. Previous work commonly assumes that all concepts are given at training time and solely focuses on predicting the unseen prerequisite relationships between them. However, many real-world scenarios deal with concepts that are left undiscovered at training time, which is relatively unexplored. This paper studies this problem and proposes a novel alternating knowledge distillation approach to take advantage of both content- and graph-based models for this task. Extensive experiments on three public benchmarks demonstrate up to 10% improvements in terms of F1 score.

1 Introduction
As the amount of online educational data rapidly grows, it is more important than ever to develop information access systems that assist users to learn new complex topics or concepts (Gwizdka et al., 2016; Collins-Thompson et al., 2017; Eickhoff et al., 2017; Urgo and Arguello, 2022). A fundamental step towards developing these systems is concept prerequisite learning (CPL)—the task of building a concept graph by structuring open knowledge in prerequisite relations. A prerequisite is a directed relation between two concepts, e.g., Heap Tree is a prerequisite of Heap Sort.
CPL was first introduced by Talukdar and Cohen (2012) with the aim of formulating probabilistic planning problems for machine learning solvers. A number of CPL models use various kinds external resources including Wikipedia links (Liang et al., 2015), textbook structures (Wang et al., 2016), and course dependencies (Liang et al., 2017; Liu et al., 2016) in two ways: content-based (Pan et al., 2017; Roy et al., 2019; Gasparetti, 2022) and graph-based (Liang et al., 2015). There also exists research on active learning (Liang et al., 2018a,b), unsuper-vised learning (Li et al., 2020), and domain adaptation (Li et al., 2021) to meet data insufficiency challenges in CPL problems. However, most existing approaches assume the system to reconstruct concept prerequisite paths with vague knowledge (i.e. incomplete relations) of each concept, or to transfer graph structure information to a new domain. In practical scenarios, knowledge is updated with both new concepts and relations introduced.
To jump out of the offline setting of graph completion with given concepts, we define a new task - CPL for unseen concepts, i.e. predicting prerequisite relationships for concepts that never appear in the training set.
Most existing graph-based CPL approaches cannot be simply used for unseen concepts, because the randomly initialized concept representations do not get updated for unseen concepts. A simple solution would be initializing the concept embeddings based on content-based models (Li et al., 2019; Jia et al., 2021; Zhang et al., 2022; Sun et al., 2022). To better take advantage of content information, we propose a novel CPL model that consists of two components: one that solely focuses on textual content associated with each concept and another one that focuses on the concept graph structure. To train our model, we propose an iterative knowledge distillation approach by alternating between these two components as “teacher” and “student”. Our main contributions include:
1. Exploring a new task to predict prerequisite relationships for unseen concepts.
2. Introducing a simple yet effective retrieval-augmented content-based approach for CPL.
3. Proposing an alternating knowledge distillation procedure that benefits both content- and graph-based models for CPL.
4. Advancing state-of-the-art on three public benchmarks. Extensive experiments shed light on the empirical contributions of each proposed
2 Methodology

This section introduces a CPL model that uses two complementary components. The first component models prerequisite relations conditioned on textual content retrieved for each concept (i.e., a retrieval-augmented model), and the second component casts the problem as a link prediction task and models prerequisite relations conditioned on the graph structure. The proposed method uses a knowledge distillation approach that alters between these two components as the teacher and student models, iteratively. An overview of the proposed solution is presented in Figure 1.

Notation and Problem Statement: Let $G = (V, E)$ be a directed concept graph whose vertices are associated with concepts and edges represent prerequisite relations. $V_h$ be a set of unseen concepts to dig out. The training set for concept prerequisite learning is equal to $D_{train} = \{(c_i, c_j, r_{ij}) : c_i, c_j \in V, r_{ij} \}$, where $r_{ij} = 1$ if $c_i \rightarrow c_j \in E$ and $r_{ij} = 0$ otherwise. The test set is $D_{test} = \{(c_i, c_j, r_{ij}) : c_i \in V_h \setminus c_j \in V_h, r_{ij} \}$.

2.1 Alternating Knowledge Distillation for Concept Prerequisite Learning

There are two formulations of the concept prerequisite learning problem, as follows:

A content-based formulation for CPL: In this formulation, the aim is to develop a model that predicts prerequisite relations based on textual information associated with the concepts, as follows:

$$\arg \min_{\theta_C} \sum_{(c_i, c_j, r_{ij}) \in D} L_C(p(c_i \rightarrow c_j | \phi, \theta_C), r_{ij}),$$

where $L_C$ and $\theta_C$ denote the loss function and the content-based model parameters, respectively. $\phi(\cdot)$ is a function that takes a concept and provide a textual description of the concept.

A graph-based formulation for CPL: An alternative formulation of the CPL problem is to predict prerequisite relationships based on the prerequisite graph structure, as follows:

$$\arg \min_{\theta_G} \sum_{(c_i, c_j, r_{ij}) \in D_{train}} L_G(p(c_i \rightarrow c_j | G, \theta_G), r_{ij}),$$

where $L_G$ and $\theta_G$ denote the loss function and the graph-based model parameters, respectively.

Alternating knowledge distillation for training: $\theta_C$ and $\theta_G$ can be trained independently or jointly on the training set $D_{train}$. We introduce a more effective alternative optimization called alternating knowledge distillation (AKD), in which the roles of teacher and student models alternate between $\theta_C$ and $\theta_G$ repeatedly. Our motivation is to improve generalization in both of these models that are complementary. Therefore, we first train our content-based CPL model $\theta_C$ (see Section 2.2) using the ground-truth training data (i.e., $D_{train}$). Then, we consider $\theta_C$ as the teacher model and produce a pseudo-labeled training set $\hat{D}$ based on $p(c_i \rightarrow c_j | \phi, \theta_C)$ as follows: For every concept $c_i$, we consider $k$ positive pseudo labels using topk($\{p(c_i \rightarrow c_j | \phi, \theta_C) : \forall c_j \neq c_i\}$). Note that we exclude the concepts with less than 0.5 prerequisite probability, if any. We also take $k'$ negative instances. These negative instances can be selected randomly or from the ones with the lowest probability. We then train the student model $\theta_G$ (see Section 2.3) on $D_{train} \cup \hat{D}$, use $\theta_C$ as the teacher, produce the pseudo-labeled training set and train the student model $\theta_C$. We repeat this teacher-student alternation process for $N$ steps. Empirically $N$ is set to 4.

Since $\theta_C$ and $\theta_G$ are complementary, we interpolate their scores linearly to acquire the final probability:

$$\alpha p(c_i \rightarrow c_j | \phi, \theta_C) + (1 - \alpha) p(c_i \rightarrow c_j | G, \theta_G),$$

where $\alpha \in [0, 1]$ is a hyper-parameter. In the following two subsections, we describe how we model $\theta_C$ and $\theta_G$, respectively.

2.2 Retrieval-Augmented Concept Prerequisite Learning

A simple approach for modeling $p(c_i \rightarrow c_j | \phi, \theta_C)$ is to use pre-trained language models (e.g.,
Table 1: CPL prediction results obtained by the proposed model and the baselines for unseen concepts. The highest number in each column is bold-faced.

<table>
<thead>
<tr>
<th></th>
<th>University Course</th>
<th>LectureBank</th>
<th>MOOC ML</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1 Score</td>
</tr>
<tr>
<td>VGAE (Li et al., 2019)</td>
<td>0.524</td>
<td>0.523</td>
<td>0.520</td>
</tr>
<tr>
<td>CPRL</td>
<td>0.554</td>
<td>0.566</td>
<td>0.540</td>
</tr>
<tr>
<td>BERT</td>
<td>0.771</td>
<td>0.771</td>
<td>0.767</td>
</tr>
<tr>
<td>NCF with BERT emds</td>
<td>0.668</td>
<td>0.676</td>
<td>0.663</td>
</tr>
<tr>
<td>Ours</td>
<td>0.844</td>
<td>0.841</td>
<td>0.842</td>
</tr>
</tbody>
</table>

Table 2: Data statistics. p+ and p- represent positive pairs and sampled negative pairs, respectively. Unused concepts are excluded.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#concepts</th>
<th>#p+</th>
<th>#p-</th>
</tr>
</thead>
<tbody>
<tr>
<td>University Course</td>
<td>407</td>
<td>1005</td>
<td>996</td>
</tr>
<tr>
<td>LectureBank</td>
<td>205</td>
<td>904</td>
<td>967</td>
</tr>
<tr>
<td>MOOC ML</td>
<td>244</td>
<td>1737</td>
<td>4975</td>
</tr>
</tbody>
</table>

BERT (Devlin et al., 2018)) to represent the concept names. In our initial experiments, we observed that concept names are not sufficient for prerequisite prediction and more descriptive content should be produced by $\phi$. Therefore, with the aim of taking advantage of a massive unstructured corpus from textual world knowledge, we augment the training data with passages retrieved from Wikipedia. To be concise, each concept name is regarded as a query in order to retrieve 100-token passages (similar to the Wikipedia DPR collection (Karpukhin et al., 2020)) using BM25. We use the first ranked passage for augmentation and compute $p(c_i \rightarrow c_j | \phi, \theta_C)$ by feeding “[CLS] n_i p_i [SEP] n_j p_j [SEP]” to BERT and using a fully-connected layer and sigmoid on top of the [CLS] representation. Note that, $n_i$ and $p_i$ are the concept name and the first retrieved passage for concept $c_i$, respectively. For the loss function, we use binary cross entropy.

2.3 Graph-based Concept Prerequisite Learning

For modeling $\theta_G$, we aim at predicting missing links in a concept graph. Various approaches based on matrix factorization, geometry, and graph neural networks have been developed for the link prediction problem. In this work, we use neural collaborative filtering (NCF) (He et al., 2017) to obtain node representations and corresponding link existence likelihood. NCF is efficient, less prone to overfit, and can be used for directed graphs. It has demonstrated successful results in a number of recommendation problems. We use NCF to learn a representation for every concept. The representations are initialized randomly and we train the model using a binary cross entropy loss function.

3 Experiments

3.1 Data

We evaluate the effectiveness of our approach on the following three manually annotated benchmarks.

University Course (Liang et al., 2017): This dataset includes concepts from computer science course descriptions provided by 11 universities in the United States. The concepts were extracted using Wikipedia Miner (Milne and Witten, 2013).

LectureBank (Li et al., 2019): This dataset was constructed by collecting online lecture files from 60 courses covering NLP and related topics.

MOOC ML (Pan et al., 2017): This dataset contains concept prerequisite relations extracted from video subtitles of Coursera’s machine learning courses using the approach presented by Parameswaran et al. (2010).

The statistics of these three benchmarks are presented in Table 2.

3.2 Experimental Setup

To evaluate the ability of predicting prerequisites of undiscovered concepts for our method, we randomly split the concept set of each dataset with a proportion of 9:1 as $V$ and $V_h$, then reconstruct the training and test set by dropping any pair with an unseen concept into test set. Thus, the representation of implicit concepts will never be updated during training. The experiments are repeated for three times with different random splits and the average is reported. We use Precision, Recall, and F1 Score (macro averaged) as evaluation metrics.

We set the NCF’s concept embedding dimensionality to 32, and the learning rates for $\theta_C$ and $\theta_G$ to
### University Course, LectureBank, MOOC ML

<table>
<thead>
<tr>
<th></th>
<th>University Course</th>
<th>LectureBank</th>
<th>MOOC ML</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1 Score</td>
</tr>
<tr>
<td>Ours</td>
<td>0.844</td>
<td>0.841</td>
<td>0.842</td>
</tr>
<tr>
<td>Ours w/o AKD</td>
<td>0.797</td>
<td>0.798</td>
<td>0.796</td>
</tr>
<tr>
<td>Ours w/o Retr-Aug BERT</td>
<td>0.739</td>
<td>0.738</td>
<td>0.738</td>
</tr>
<tr>
<td>Ours w/o NCF</td>
<td>0.809</td>
<td>0.811</td>
<td>0.808</td>
</tr>
</tbody>
</table>

Table 3: The ablation study results. w/o AKD means interpolation only. w/o model means this model is only used for distillation but not interpolation. The highest number in each column is bold-faced.

<table>
<thead>
<tr>
<th></th>
<th>University Course</th>
<th>LectureBank</th>
<th>MOOC ML</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1 Score</td>
</tr>
<tr>
<td>NCF $D_{train}$</td>
<td>0.674</td>
<td>0.730</td>
<td>0.652</td>
</tr>
<tr>
<td>NCF $D_{train}$, $\hat{D}$</td>
<td>0.734</td>
<td>0.733</td>
<td>0.734</td>
</tr>
<tr>
<td>NCF $\hat{D}$</td>
<td>0.712</td>
<td>0.740</td>
<td>0.705</td>
</tr>
</tbody>
</table>

Table 4: NCF results with ground truth data $D_{train}$, first iteration distilled data $\hat{D}$, and both.

5e-5 and 1e-3, respectively. We use BERT-base in all experiments. For the AKD process, $k$ is proportioned to the size of positive training instances for each concept (i.e., 10%). We also set $k' = k$. The hyper-parameter $\alpha$ was selected using grid search.

### 3.3 Baselines

We use the following competitive baselines: **VGAE** (Li et al., 2019) uses variational graph autoencoder to encode edges in the training set with a 2-layer Graph Convolutional Network (GCN), then adopts inner product to reconstruct the graph. **CPRL** (Jia et al., 2021) is the most recent CPL approach that produces state-of-the-art results by creating a heterogeneous graph for representing concepts in addition to learning objects. It uses a Relational Graph Convolutional Network (R-GCN) to encode nodes and a Siamese network to identify prerequisites. Note that CPRL uses an external resource. To provide a fairer comparison, we implemented a simplified version of CPRL that excludes the learning objects. We call this method **CPRL--**. **BERT** (Devlin et al., 2018) takes a pair of concept names and is fine-tuned to classify prerequisite relations using the binary cross entropy loss. **NCF with BERT emb**s is implemented to compare different combination methods of content and graph based models. We follow the strategies in (Li et al., 2019; Jia et al., 2021) that initialize NCF node representations with fine-tuned BERT embeddings.

### 3.4 Results

**Comparison with the Baselines:** According to Table 1, graph-based baselines perform poorly when dealing with unseen concepts. Unsurprisingly, carrying information from the pre-training step helps the BERT model produce the best results on both University Course and MOOC ML datasets. Implicitly using BERT representations is helpful, but the prediction ability of graph-based model is limited. Our method outperforms the baselines on all three datasets in terms of all metrics. The improvements come from the augmentation using passages retrieved from Wikipedia, the alternating knowledge distillation approach, and the explicit combination of complementary models.

**Ablation Study:** In our ablation study, we answer the following empirical research questions:

**Q1:** Does retrieval augmentation improve the generalizability of the content-based model? The F1 scores for the proposed retrieval-augmented BERT on University Course, LectureBank and MOOC ML are 0.796, 0.747, 0.800 respectively. Comparing them to the BERT’s performance reported in Table 1 demonstrates the generalizability of retrieval augmentation for this task.

**Q2:** How is the effect of different components in our AKD approach? In Table 3 we eliminate each of components and demonstrate substantial drop in nearly all cases. The large performance drops by removing the retrieval-augmented BERT model is due to its role of capturing content information. This experiment demonstrates that all the components used in developing our approach contributes to the final performance.

**Q3:** How does distilled data contribute to NCF training? In Table 4, $\hat{D}$ acts as a better training set than $D_{train}$, indicating that even weakly anno-
tated unseen concept pairs can play an important role in guiding graph based models. A combination of ground truth $D_{train}$ and $\hat{D}$ leads to strong improvement.  

**Learning Curve:** We plot the learning curves of our model for all three datasets in Figure 2. The model’s effectiveness is substantially improved by increasing the training data size. Given the slope of the learning curves, the proposed model is likely to achieve significantly higher F1 scores by increasing the training data size. This is an encouraging observation, especially given that the developed model is already very effective and obtains F1 scores of higher than 0.86 on all datasets.

**Results for Seen Concepts:** Even though this paper focuses on unseen concepts, we also compare our methods against the baselines for predicting prerequisite relations for seen concepts. Following previous work (Li et al., 2019; Jia et al., 2021), we split the concept pairs in each dataset into training and test sets with a proportion of 9:1 for LectureBank and 7:3 for others. Results in Table 6 show that our method outperforms baselines on all three datasets in terms of all metrics, indicating that our approach is equipped with the ability to deal with seen concepts.

**4 Conclusions and Future Work**

This paper explored the challenge of predicting prerequisites for unseen concepts in CPL. It proposed an alternating knowledge distillation approach that enables us to train more effective content-based and graph-based models, as well demonstrated that content-based CPL models can benefit from retrieval augmentation. In the future, we intend to extend the proposed solution to an online setting, where concept prerequisites can be extracted for every learning-oriented query in a search engine.
Limitations

One of the limitations of this work is that this work overlook the concept detection or extraction in the model design. Even though previous work also made similar assumptions (Liang et al., 2017; Li et al., 2019; Pan et al., 2017), we believe that this is an important aspect that should be considered in the future. Extracting concepts from unstructured data accurately can be challenging. Another limitation is related to the number of concepts in the datasets. In some real-world scenarios, the number of concepts would be significantly higher than those represented by the existing benchmarks. Increasing the number of concepts is likely to negatively impact the model’s effectiveness or raise efficiency concerns.

Mistakes made by the CPL models, including ours, if they are used in learning-oriented search engines, are likely to negatively impact the learning outcome. For instance, missing a prerequisite relation during a learning session may lead to some misunderstanding about the concepts being learned. Therefore, we suggest raising awareness of such mistakes to the users so they can make wise decisions while learning online.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #2106282. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References


