Abstract

In the digital age, the rapid growth of online learning platforms and the widespread availability of educational resources have transformed the landscape of education. However, this abundance of content presents a new challenge: learners often struggle to identify the most relevant materials and comprehend them effectively without guidance. A critical aspect of personalized and adaptive learning is the identification of core educational concepts and the understanding of prerequisite relationships among them. Towards this direction, the thesis makes three contributions.

First, we propose a probabilistic framework that identifies key educational concepts from unstructured text. Candidate concepts are first generated using Wikipedia anchor texts, leveraging their natural linkage to important entities. These candidates are then refined using a graph-based probabilistic model that incorporates contextual signals from the document itself. Our approach eliminates the need for extensive manual annotations or domain-specific heuristics, making it highly adaptable across disciplines. We evaluate the proposed method on two curated datasets—a Physics school textbook and a set of Physics research articles—annotated manually for this research. Experimental results demonstrate that our method consistently outperforms both supervised and unsupervised baselines, establishing its effectiveness in real-world educational scenarios.

Second, we address the problem of prerequisite relation extraction, identifying the dependencies between educational concepts that define the optimal sequence in which they should be learned. Traditionally, such relationships are manually curated by domain experts. To automate this process, we propose a graph neural network (GNN)-based model that operates on a heterogeneous graph structure comprising both concepts and associated resources. Our model leverages node-level attention mechanisms to capture fine-grained dependencies and contextual interactions within the graph. Experiments on multiple benchmark datasets show that our method significantly outperforms existing models, consistently achieving new state-of-the-art performance across evaluation metrics.

The third part of the thesis focuses on the efficiency of prerequisite relation extraction. As the number of concepts and their interconnections grows, existing graph-based models become increasingly resource-intensive, limiting their practicality in real-time learning systems. To overcome this challenge, we introduce an approach incorporating extreme value theory to selectively sample the most informative neighbours for each node in the resource-concept graph. This selective attention mechanism drastically reduces the size and complexity of the graph without compromising the model's performance. Our experiments demonstrate that the proposed method achieves competitive accuracy compared to state-of-the-art models while reducing GPU memory consumption and execution time by a large margin.

In summary, this thesis makes three contributions: (1) a probabilistic concept extraction method tailored for unstructured educational texts; (2) an attention-based graph neural network for accurate and automated prerequisite relation learning; and (3) a resource-efficient modeling strategy that maintains performance while enabling scalability.

Keywords: concept identification, prerequisite relation identification, pedagogical concepts, graph attention network, probabilistic model.