Unsupervised Prerequisite Chain Learning using Relational Graph Autoencoders

Irene Li, Alexander R. Fabbri and Dragomir R. Radev

Yale School of Computer Science, New Haven, CT, USA

LILY Lab

Introduction

Recent years have seen the expansion of online educational resources, especially in the fields of Artificial Intelligence (AI), Machine Learning (ML) and Natural Language Processing (NLP). Online resources are often tagged with a topic, but the user does not necessarily know the order in which topics and associated materials should be learned. To address this situation, we frame learning prerequisite relationships as an unsupervised task with access only to resource-topic labels. We propose Relational-Graph AutoEncoder (R-GAE) model to predict concept relations within a concept and resource graph. Experimental results show that our unsupervised approach performs on par with relevant supervised methods, on an expanded, annotated dataset. Additionally, we examine the effect of topic-label quality, the type of input resources and sparsity on downstream prerequisite learning. We also provide a new collection of our previous work LectureBank.

Materials

We expanded the concept list from LectureBank1.0, where there are 208 concepts within same categories. We eliminated few concepts which are very specific or combined some concepts to make more fine-grained ones. For example, there are concepts BLUE, ROUGE, replaced them using “machine translation evaluation”. Besides, we included some new concepts with a better coverage of the five categories, thus eventually we have a total number of 322 concepts. Compared with the 208 topics provided by our previous work, they were proposed randomly, in our new concept list, we carefully checked each slide file in our corpus, and referred some concepts from taxonomy. Besides, we also tried to include some recent new concepts such as variations of word embeddings and GAN models.
The next step is to do annotation on the new proposed concept pairs. We will follow the similar approach as what we did for the previous two papers. However, we kept the old version from LectureBank1.0, and will only label the new ones. Finally we will calculate kappa score to show the level of agreement.

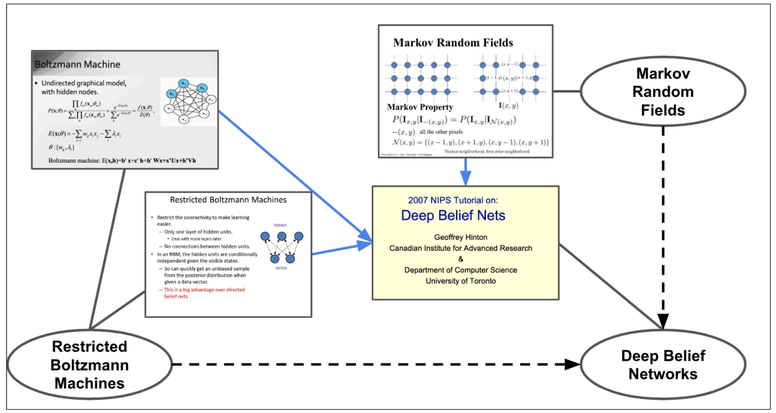

Figure 1. Illustration about concept and resource graph.

Method

Graph Structure
We propose Relational-Graph AutoEncoder (R-GAE) to unsupervised prerequisite chain learning. It contains a relational-graph convolution network:

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}\right)$$

Where we define three types of edges that calculate message flows using corresponding parameter matrix W. For each layer, there is a matrix W0 that computes self connected flow.
Then we introduce the variational Graph AutoEncoder, inference model is given as:

$$q(\mathbf{Z}\,|\,\mathbf{X},\mathbf{A}) = \prod_{i=1}^{N} q(\mathbf{z}_i\,|\,\mathbf{X},\mathbf{A})$$

Where we have $\mu = RGCN_\mu(\mathbf{X},\mathbf{A})$
The generative model is defined as a dot product:

$$p\left(\mathbf{A}\,|\,\mathbf{Z}\right) = \prod_{i=1}^{N}\prod_{j=1}^{N} p\left(A_{ij}\,|\,\mathbf{z}_i,\mathbf{z}_j\right)$$

Adjacency Matrix
There are 3 types of relationships in total:
a) Document-document relation: we define this in the next section.
b) Document-concept relation: 1 if the document is labeled as this concept, 0 otherwise;
c) Concept-concept relation: 1 if the concept is a prerequisite of the other concept, 0 otherwise.
We define the Adjacency matrix A due to the above criteria.

Document Relationships
It is possible to use Pointwise mutual information (PMI) or cross-entropy methods.

$$\mathrm{npmi}(x;y) = \frac{\mathrm{pmi}(x;y)}{h(x,y)}$$

Feature Matrix
We Initially we went through each file and concept list to get a vocabulary by filtering out stop words. Then for the concept nodes, we applied one-hot vector representations. For the file nodes, we performed TF-IDF features under our vocabulary. Then we have a complete node matrix X, where each row is a vector representing a node.

Table 1. LectureBank 2.0 Statistics.

Domain	#courses	#lectures	#slides	#tokens	#tokens/lecture	#tokens/slide
NLP	49	1,049	41,555	2,090,708	1993.05	50.31
ML	17	357	14,358	1,060,990	2971.96	73.90
AI	7	149	6,975	317,266	2129.30	45.49
DL	7	249	7,104	713,047	2863.64	100.37
IR	4	79	3,377	157,808	1997.57	46.73
Overall	84	1,883	73,369	4,339,819	2391.10	63.36

Baseline

Most of the related works evaluated using an information retrieval approach. We decide to keep accuracy and F1 scores as our evaluation matric. We will compare with the results from our previous paper, and report accuracy and F1 scores.

Conclusion and Next Steps

One of our contribution is to solve the problem of prerequisite chain learning in an unsupervised way. In our previous work, we showed that graph convolution network has the potential to tackle this problem. In this project, we will try to propose an unsupervised model based on relational graph convolution networks with heterogenous nodes and directed edges.
We release a new batch of LectureBank along with a larger annotation set for prerequisite relation. We also propose our relational graph autoencoder model to solve prerequisite chain learning in an unsupervised way. The future work will be test our model and report accuracy, F1 scores and compare with some supervised methods.

Acknowledgement

I want to thank my advisor Drago and my best lab mate Alex. My special thanks also goes to the awesome restaurant Bonchon, and my friends who spent a great time with me there: Tianwei She, Tao Yu, Suyi Li, Yuang Jiang and Yiliang Zhang.