A Transfer Learning Pipeline for Educational Resource Discovery with Applications in Prerequisite Chain Learning and Survey Generation

Yale

Irene Li, Thomas George, Alex Fabbri, Tammy Liao, Benjamin Chen, Rina Kawamura, Richard Zhou, Vanessa Yan, Swapnil Hingmire, Dragomir Radev

LILY Lab

Introduction

Effective human learning depends on a wide selection of educational materials (such as textbooks, lecture slides, blog posts) that align with the person's current understanding of the topic. While the internet has revolutionized learning, a substantial barrier still exists. Providing online readers with high quality secondary literature resources on a given domain (e.g., natural language processing) is more accessible for beginners. In this paper, we propose a pipeline for building such an educational resource discovery system for new domains. The pipeline consists of three main steps: resource searching, feature extraction, and classification. The pipeline achieves F1 scores of 0.94 and 0.82 when evaluated on two similar domains respectively. Finally, we demonstrate how this pipeline can benefit two applications: prerequisite chain learning and survey generation. Additionally, we release a corpus of 39,728 manually labeled web resources and 659 queries from NLP, Computer Vision (CV) and Statistics (STATS).

Data Collection & Annotation

Data collection: we collected URLs based on selected queries and downloaded from search engines. We list the detailed data statistics in Table 5. Comparison with other similar datasets are listed in Table 2. We provide the total numbers in both file type and domain dimension. Among all three domains, we are able to collect 39,728 valid URLs using 659 different queries. For annotating, we have a group of 7 students who have a good background of NLP, CV and STATS.

Classification

We conducted tree-based methods, comparing with random forest and decision tree. We applied different types of features. Figure 1 shows pretrained QD-BERT model (query-document), and a list of BERT models are shown in Table 1. Table 4 is the comparison of the different feature groups.

Figure 1. QD-BERT MLM model pretraining.

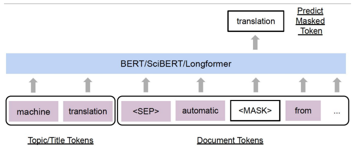

Table 1. MLM BERT models pre-trained.

arXiv_bert	BERT pre-trained on arXiv
arXiv_scibert	SciBERT pre-trained on arXiv
arXiv_lf	Longformer pre-trained on arXiv
bert_base	BERT base model
scibert_base	SciBERT base model
longformer_base	Longformer base model
AAN_lf	BERT pre-trained on AAN
AAN_bert	SciBERT pre-trained on AAN
AAN_scibert	Longformer pre-trained on AAN

Table 4. Classification performances.

Features	NLP→CV F1	NLP→CV Precision	NLP→CV Recall	NLP→STATS F1	NLP→STATS Precision	NLP→STATS Recall
Group 1	0.7238	0.5802	0.9617	0.6508	0.5405	0.8177
Group 1 + 2	0.8579	0.7772	0.9571	0.7990	0.8141	0.7845
Group 3, BERT Only*	0.7764	0.7522	0.8497	0.7923	0.7903	0.7944
Group 1 + 2 + 3	**0.9402**	0.9849	0.8994	**0.8225**	0.9965	0.7002

	NLP	CV	STATS
Token Number/per sentence			
Mean	18.28	26.37	23.28
Median	12	19	18
Max	2,302	458,363	20,066
Sentence Number			
Mean	161.60	122.49	107.32
Median	55	46	52
Max	5,929	21,301	52,793

Table 5. Detailed statistics.

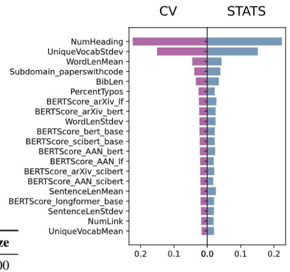

Figure 2. Feature Importance.

Name	Resource Type (with texts)	Domain Number	Annotation	Size
TutorialBank	Lecture sides, papers, blog posts	NLP only	Manually	6,300
LectureBank	Lecture sides	NLP only	Manually	1,717
MOOCcube	Papers	Multiple	Scrape from third-party	679,790
Our Pipeline	Lecture sides, papers, blog posts	3	Manually	39,728

Table 2. Comparison with similar datasets.

	NLP	CV	STATS	Total
Query Num	322	200	137	659
PPTX	1,216	733	1,463	3,412
PDF	4,961	3,782	1,449	10,192
HTML	9,368	9,302	7,454	26,124
Total	15,545	13,817	10,366	**39,728**
Pos.Num	9,589	11,101	6,742	27,432
Pos.Rate	0.6169	0.8034	0.6501	0.6905

Table 3. A detailed statistics of the datasets.

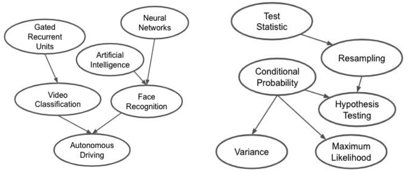

Figure 3. Reconstructed concept graph (part) for CV (left) and STATS (right).

Evaluation

We show the transfer learning results in Table 4. As we can see, adding group 2 features is better with group 1 only. Group 2 provides smaller granularity of the features, for example, number of tokens.
QD-BERT model is able to outperform Group 1 features in the F1 score. In general, we see that when combining all features, we would achieve the best score F1 scores. Besides, we see that CV has a higher score than STATS, this is because that CV and NLP share more common features than CV and STATS.
Figure 2 shows the importance scores.

Applications

Prerequisite chain learning is extremely helpful in the scenario of student learning. Knowing the prerequisite concepts of a target concept is beneficial when a learner wants to study new knowledge. We first train concept embeddings, then compare with there methods: logistic regression, one-layer neural network and a variational graph autoencoder model. For each method, we compare our pipeline and a basic pre-trained BERT model. We find that NN performs the best. Figure 3 shows the examples of both CV and STATS: reconstructed concept graph.
We also conducted survey generation using the data classified by our pipeline, due to space limitation, we eliminated examples here.

Conclusion

In this paper, we proposed a pipeline for building a knowledge resource system in an unfamiliar subject area. We tested on in-domain and out-of-domain applications and achieved promising results. We also released our dataset and annotations.

Acknowledgement

Thanks to my advisor Drago and my collaborators of this project. My sincere gratitude also goes to Puff, a cat influencer on Weibo, for her cute videos and photos posted online. Special thanks to Zijin Chen, for his excellent works I read this semester: *Low-intelligence Crime*, *Tracker* and *Forbidden Land*. Those are amazing masterpieces!