How should I learn: prerequisite chain learning and the introduction of LectureBank

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

Department of Computer Science, Yale University

LILY Lab

Introduction

As more and more open courses and materials are becoming reachable to everyone, attending real-word lectures are not the only way of acquiring knowledge of interests. However, it is expensive to search for a learning path without solid domain knowledge. Imagine this scenario, when a student wants to learn a specific concept, for example, "convolutional neural networks". In most of the cases, the student needs to understand the pre-requisite of the concept which would helpful for learning the new concept, here "back-propagation" maybe a pre-requisite. Although many search engines are able to provide relevant documents or learning resources, most of the results are based on relevancy or semantics, very few are able to give a reasonable list based on a learning path. Thus, for educational purpose, we are expecting to learn the dependency of each concept pair and eventually apply them on a search engine. In this paper, we introduce our works of pre-requisite chain learning on two manually created datasets. Our contributions are the following: 1) design a framework to first extract concepts using the embedding-based topic modelling way; 2) we then further learn a concept graph as the prerequisite chains using a dataset in the field of natural language processing which contains about 1000 concepts on 527 manually-collected course slides.

Materials

We learn prerequisite from two datasets. LectureBank: we collected 26 courses online in the field of natural language processing. The dataset is expected to focus on the courses from a various disciplines (NLP, ML, AI, IR, etc). We extracted about 1000 topics in total. TutorialBank: manually collected and categorized corpus of about 5,600 resources mainly on NLP, and these resources are excluding scientific papers. It contains a good taxonomy with a coverage about 305 topics.

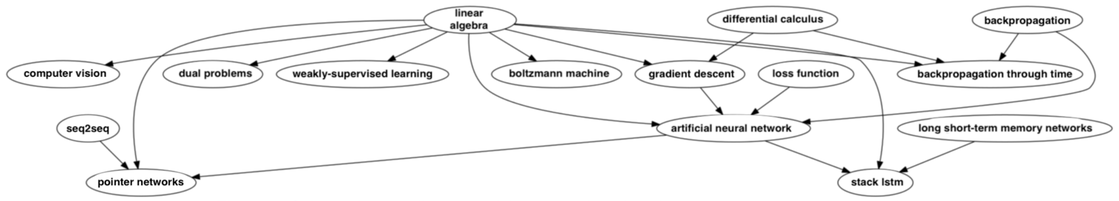

Figure 1. Subset of manually prerequisite annotations from inter-annotator agreement round.

Baseline Method

Reference Distance (RefD, Liang et al 2015): models the relationship of two concepts by considering how differently they refer to each other, such as asymmetry and irreflexivity.

$$RefD(A,B) = \frac{\sum_{i=1}^{k} r(c_i,B)\cdot w(c_i,A)}{\sum_{i=1}^{k} w(c_i,A)} - \frac{\sum_{i=1}^{k} r(c_i,A)\cdot w(c_i,B)}{\sum_{i=1}^{k} w(c_i,B)}$$

Where $w(c_i, A)$ weights the importance of c_i to A; $r(c_i, A)$ is an indicator showing whether c_i refers to A, which could be links in Wikipedia, mentions in books or citations.

Proposed Method

We follow a two-step approach to learn the prerequisite chains. The first step is to classify the documents or lecture pages into topics or concepts. The second step is to learn the prerequisite dependency of the topics. In the first step, we apply embedding-based methods. In stead of Wikipedia links or any third-party vocabulary, we train our distributed representation for the topics with our corpus. The method includes LDA topic modelling, LDA2vec (Moody, 2016) and Doc2vec.

For the second step, following by the method from (Liang et al 2018), after we have representations from the topics, we can train each pair with some classifiers as a simple binary classification task. Another method is to apply a variational graph autoencoder to model the topic relations. It is a combination of graph convolutional network encoder with an inner product decoder.

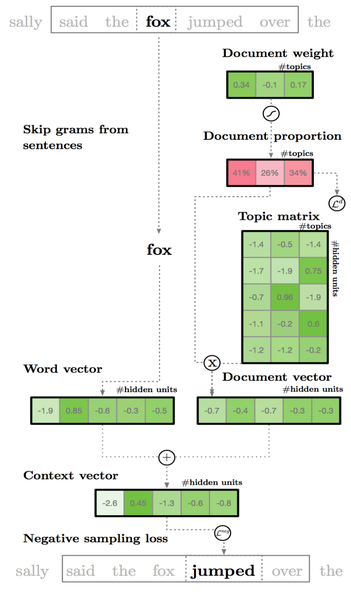

Figure 2. Lda2vec diagram.

Table 1. Doc2vec Baseline Scores.

Method	Precision	Recall	F1
Logistic Regression	**0.720**	**0.758**	**0.739**
Naive Bayesian	0.593	0.5634	0.613
SVM	0.624	0.569	0.596

Preliminary Results

We train our Doc2vec model using Gensim library on the TutorialBank dataset. A reason is the TutorialBank dataset has richer text contents, which we think is better to learn word or document embeddings. We set the dimension to be 300, and the sliding window size is 10. For TutorialBank dataset, we come up with 210 concepts, and we make annotations for each pair of the concepts (if concept A is a prerequisite of concept B). Eventually we achieved 569 positive samples, out of a total number of 27797. Then we take the labelled dataset as training samples, because the positive samples are sparse, we apply oversampling on the training samples. We use our pre-trained Gensim model to infer these concepts, and simply concatenate two concept vectors. Then the new combined vectors are treated as inputs to the following classifiers and we report Precision, Recall and F1 score: Logistic Regression, Naïve Bayesian and SVM. Logistic Regression has a higher score based on the same Gensim concept features.

Conclusion

Learning prerequisite chain is an interesting research topic as it will make a difference on the traditional learning process for the learners. While some of the related words focused on traditional feature extracting approach, we apply dense embedding-based method to tackle this problem. We are currently having some promising results and in the future we will expand the work to new approach like graph-based methods. It is also possible to expand the method to other educational fields and courses.

References

Liang, Chen, et al. "Measuring prerequisite relations among concepts." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.

Liang, Chen, et al. "Investigating Active Learning for Concept Prerequisite Learning." Proc. EAAI (2018).

Moody, Christopher E. "Mixing dirichlet topic models and word embeddings to make lda2vec." arXiv preprint arXiv:1605.02019 (2016).

Acknowledgement

Special thanks to Prof Dragomir Radev for his advising. Thanks to my colleagues Alex and Robert for the collaboration and support. Congratulations to Robert's graduation and new job in advance. I also want to extend my gratitude to Xinyang for his spiritual support.