

Introduction

Natural Language Processing (NLP) has been growing rapidly in recent years with the development of Deep Learning (DL) and its application to NLP tasks. As a result, the job of the student, engineer and research of keeping up to date with the growing body of knowledge in the field is becoming more difficult. To directly address this situation, we introduce TutorialBank, a new, publicly available dataset which aims to facilitate NLP education and research. We have manually collected and categorized over 5,600 resources on NLP as well as the related fields of Artificial Intelligence(AI), Machine Learning (ML) and Information Retrieval (IR). Our dataset is notably the largest manually-picked corpus of resources intended for NLP education which does not include only academic papers. Additionally, we have created both a search engine and a command-line tool for the resources and have annotated the corpus to include lists of research topics, relevant resources for each topic, prerequisite relations among topics, relevant sub-parts of individual resources, among other annotations.

Taxonomy Top-level Topics
1 - Introduction and Linguistics
2 - Language Modeling, Syntax, Parsing
3 - Semantics and Logic
4 - Pragmatics, Discourse, Dialogue, Applications
5 - Classification
6 - Information Retrieval and Topic Modeling
7 - Neural Networks and Deep Learning
8 - Artificial Intelligence
9 - Other Topics

Table 1: Top-level Taxonomy Topics.

Resource Category	Count
corpus	126
lecture	126
library	920
naelo	190
paper	1186
resource	797
survey	342
tutorial	1917

Table 3: Corpus count by pedagogical feature.

Annotation Process

We first collected resources and categorized them into a taxonomy of over 300 topics (Table 1). The taxonomy be as an attempt to cover the topics as described in course on NLP as well as AI, ML and IR but has since been expanded. As seen in Table 2, the most frequent resource topics pertain to NLP and recent advances in DL. Notably, rather than add all resources found on a topic, resources were filtered for quality by annotators. Resources consist of PowerPoint presentations as well as html pages. We converted PowerPoints to text files using PDFBox. While we cannot release the resources in their entirety due to copyright issues, we release the urls as well as our collected meta data and scripts for re-building the dataset.

One of the meta data labels present in our dataset refers to pedagogical function. Recent work has aimed to perform text categorization on the value that an online source provides to the user. Our corpus includes the eight pedagogical functions, as seen in Table 3.

Recent work on survey generation for scientific topics has focused on creating summaries from academic papers.

We frame the task of creating surveys of scientific topics as a document retrieval task. We first identified by hand 200 potential topics for survey generation in the fields of NLP, ML, AI and IR. Once the list of topics was compiled, annotators were assigned topics and asked to search that topic in the TutorialBank search engine and find relevant resources. In order to impose some uniformity on the dataset, we chose to only include resources which consisted of PowerPoint slides as well as HTML pages labeled as tutorials. We asked the annotators to choose five resources per topic and rank the resources in terms of relevance to the topic. Then, the resources were divided into content cards (by slide or HTML divider) and asked to determine whether each card is helpful for learning the given topic (on a -1,0,1 scale).

Topic Category	Count
Introduction to Neural Networks and Deep Learning	503
Tools for Deep Learning	424
Miscellaneous Deep Learning	283
Machine Learning	236
Python Basics	135
Recurrent Neural Networks	128
Word Embeddings	118
Reinforcement learning	99
Convolutional Neural Networks	97
Machine Learning Resources	75

Table 2: Corpus count by taxonomy topic for the most frequent topics (excluding topic “Other”).

By creating content for an information retrieval task we simultaneously created reading lists for each of these topics.

Even with a collection of resources and a list of topics, a student may not know where to begin studying a topic of interest. For this purpose, we annotated which topics are prerequisites of others for each of the topics from our topic list. We expanded our list of potential prerequisites to include eight additional topics which were too broad for survey generation (e.g., Linear Algebra) but which are fundamental prerequisites. We define a prerequisite in the following way: Topic Y is a prerequisite of X if understanding Topic Y help you to understand Topic X? The annotator can label the prerequisite relationship as not a prerequisite, somewhat or a prerequisite.

Dataset Statistics

We created reading lists for 182 of the 200 topics we identify in Section 4.2. Resources were not found for 18 topics due to the granularity of the topic (e.g., Radial Basis Function Networks) or due to the coverage of TutorialBank, something we plan . The average number of resources per reading list for the 182 topics is 3.94. As an extension to the reading lists we collected Wikipedia pages for 184 of the topics and present these urls as part of the dataset. For survey extraction, we automatically

split 313 resources into content cards which we annotated for usefulness in survey extraction. For prerequisite labels, our network consists of 794 unidirectional edges and 33 bidirectional edges. This shows the sparsity of prerequisite relations in the dataset. Bidirectional edges correspond to topics which go hand-in-hand such as Bleu and Rouge. Finally, we collected a total of 2,000 images and matched them with the taxonomy topic name of the resource it came from as well as the url of the resource.

Conclusion and Future Work

In this paper we introduce the TutorialBank Corpus, a collection of over 5,600 hand-collected resources about NLP and related fields. Our corpus is notably larger than similar datasets which deal with pedagogical resources and topic dependencies and unique in use as an educational tool. To this point, we believe that this dataset will be an invaluable tool to the students, educators and researchers of NLP and promote research on tasks not limited to pedagogical function classification, topic modeling and prerequisite relation labelling. We are planning additional layers of annotation 1) we are going to have multiple annotators annotate the prerequisite relations under less ambiguous conditions and take the majority vote as input into a prerequisite chain model 2) we also plan to add additional hand-written surveys and explore better parsers for HTML’s and PowerPoints 3) as TutorialBank grows, we will modify the taxonomy to reflect current research trends. Finally, we are constantly looking for ways to improve the AAN website and hope to add user input in future annotations and models.

Acknowledgement

Thank you to Professor Radev for advising this project and for providing the foundation of work in AAN to build upon. Thank you to all the co-authors as well as Jungo Kasai, Michihiro Yasunaga and Rui Zhang for their conversations and help with annotation.