TutorialBank: Using a Manually-Collected Corpus for Prerequisite Chains, Survey Extraction and Resource Recommendation

Alexander R. Fabbri, Irene Li, Prawat Trairatvorakul, Yijiao He,
Wei Tai Ting, Robert Tung, Caitlin Westerfield, Dragomir R. Radev PhD
Yale University

LILY Lab: lily.yale.edu

Introduction

- Natural Language Processing has been growing rapidly in recent years.
- We introduce TutorialBank, a new, publicly available dataset which aims to facilitate NLP education and research.
- We have manually collected and categorized about 8,000 resources on NLP as well as the related fields of Artificial Intelligence, Machine Learning and Information Retrieval.
- We have created both a search engine and a command-line tool for the resources and have annotated the corpus.

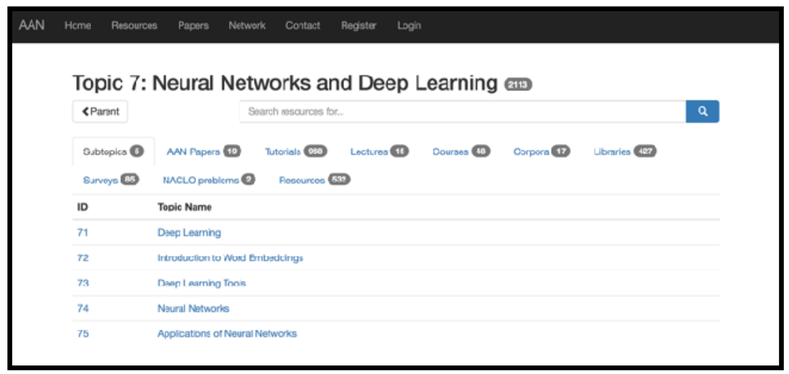
Taxonomy Top-level Topics
1 - Introduction and Linguistics
2 - Language Modeling, Syntax, Parsing
3 - Semantics and Logic
4 - Pragmatics, Discourse, Dialogue, Applications
5 - Classification
6 - Information Retrieval and Topic Modeling
7 - Neural Networks and Deep Learning
8 - Artificial Intelligence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

Table 1: Top-level Taxonomy Topics.

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").

Search Engine with 8,000 NLP Resources — **http://aan.how/**



Annotation Process

- We collected resources and categorized them into a taxonomy of over 300 topics.
- We identified 200 potential topics for survey generation, which we frame as document retrieval.
- We asked the annotators to choose five resources per topic and rank the resources in terms of relevance to the topic.
- Resources were divided into content cards (by slide or HTML divider) and annotators were asked to determine whether each card is helpful for learning the given topic (on a -1,0,1 scale).
- We annotated which topics are prerequisites of other topics for each topic of the 200-topic list.

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

Table 3: Corpus count by pedagogical feature.

Capsule Networks
Domain Adaptation
Document Representation
Matrix factorization
Natural language generation
Q Learning
Recursive Neural Networks
Shift-Reduce Parsing
Speech Recognition
Word2Vec

Table 4: Random sample of the list of 200 topics used for prerequisite chains, readling lists and survey extraction.

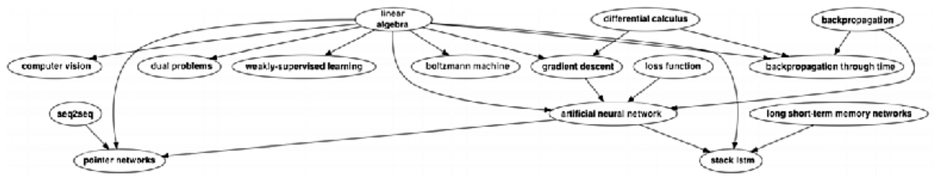


Figure 1: Subset of prerequisite annotations taken from inter-annotator agreement round.

Dataset Statistics

- We created reading lists for 182 topics.
- The average number of resources per reading list for the 182 topics is 3.94.
- We collected Wikipedia pages for 184 of the topics.
- We automatically split 313 resources into content cards for survey extraction.
- Our prerequisite network consists of 794 unidirectional edges and 33 bidirectional edges.
- We collected 2,000 images and matched them with taxonomy topics

Future Work

- Additional annotation:
  - We will to have multiple annotators annotate the prerequisite relations under less ambiguous conditions.
  - We plan to add additional hand-written surveys and explore better parsers for HTMLs and PowerPoints
  - As TutorialBank grows, we will modify the taxonomy to reflect current research trends.
- We are constantly looking for ways to improve the AAN website and hope to add user input in future annotations and models.

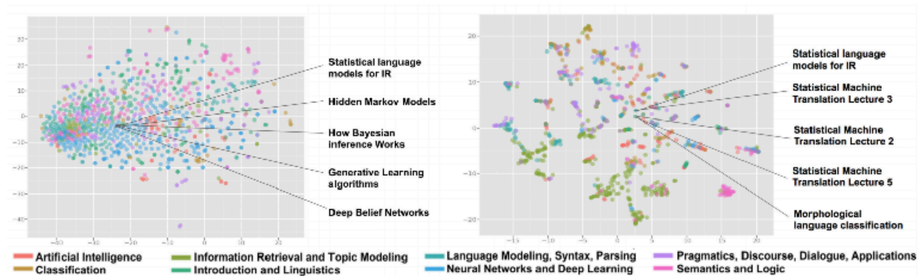


Figure 2: Plot showing a query document with title "Statistical language models for IR" and its neighbour document clusters as obtained through tSNE dimension reduction for Doc2Vec (left) and LDA topic modeling (right). Nearest neighbor documents titles are shown to the right of each plot.