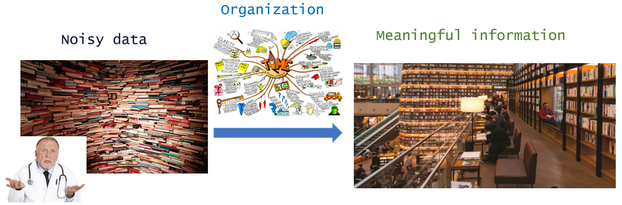
HierCon: Hierarchical Organization of Technical Documents Based on Concepts

Keqian Li, Shiyang Li, Semih Yavuz, Hanwen Zha, Yu Su, and Xifeng Yan
University of California, Santa Barbara & Google Brain & The Ohio State University

Introduction

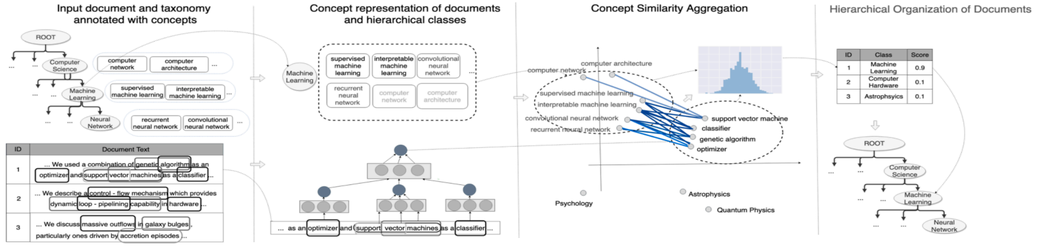
- Knowledge is being produced at an unprecedented level*
  - about 3 million scholarly journal articles each year, with an annual growth rate of 5%
- How to for better understanding and organizing the scientific literature ?
  - According to cognitive and social science[234], a key management strategy for such information is to organize them into a hierarchy of categories



- We study the hierarchical problem under a weakly supervised setting:
  - **Input**: unlabeled document set D, tree structured label set T a set of labeled training data, a set of l ≪ D document-label pairs
  - **Goal**: associate each document $d \in D$ with one or more relevant labels $L(d) \subseteq T$
- Challenges:
  - **Domain closeness**: Corpus in more technical fields are closed domain and not covered by existing knowledge bases
  - **Scarcity of labels**: Labels are expensive to obtain due to high expertise requirement and dynamic evolving nature of science
  - **Large scale label hierarchy:** categorizations needs to be stable and handle large number of hierarchical categories
  - **Collective signals**: Documents' main topic should be determined based on entire content instead of single keywords

[1] R. Johnson, A. Watkinson, and M. Mabe, "The STM report." 2018
[2] K. Lamberts, *Knowledge Concepts and Categories*. Psychology Press, 2013.
[3] J. S. Wilkins, "What is systematics and what is taxonomy," *Google Scholar*, 2011.
[4] B. S. Wynar, A. G. Taylor, and J. Osborn, *Introduction to cataloging and classification*. Libraries Unlimited Englewood, CO, 1992.

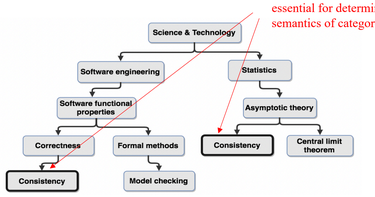
Approach overview



- **We follow a concept-representation based approach:**
  - We propose to represent the categories as *distributions* over concepts, which allows for more flexible combinations of the semantics of neighboring nodes in the hierarchy.
  - We propose a novel, adaptive concept level document representation model based on the hierarchical neural attention mechanism, which models the validity and importance of the concept recognition as a natural hierarchical process
  - We propose a principled approach for aggregating all possible concept interactions between the documents and each of the possible categories, to comparatively obtain document-category relevance and perform categorization.

Concept representation for taxonomy nodes

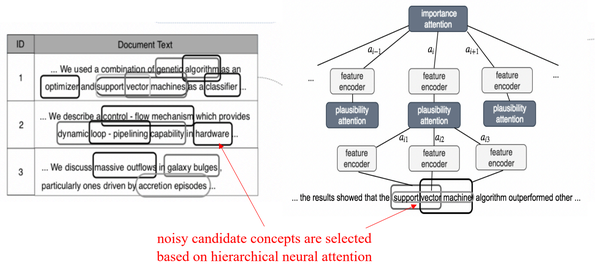
- Taxonomy nodes are represented by a weight distribution over concepts mined from corpus*
- Hierarchical structure can be encoded by enriching the concept representation of each nodes with aggregated semantics based from its descendants and path semantics based on all its ancestors
- Naturally enables assignment to intermediate node and top K category prediction



1. K. Li, H. Zha, Y. Su, and X. Yan, "Concept mining via embedding," in 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018, pp. 267–276.

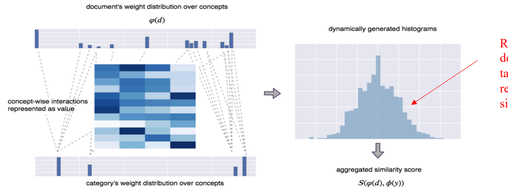
Concept representation for documents

- Concepts can be mined comprehensively using state-of-the-art chunking and text mining approaches
- The task now becomes, to select concepts that are 1) most plausible among different candidates, and 2) most important to the document's main theme
- We propose a model hierarchical neural attention mechanism to capture the plausibility attention and importance attention in an end-to-end fashion

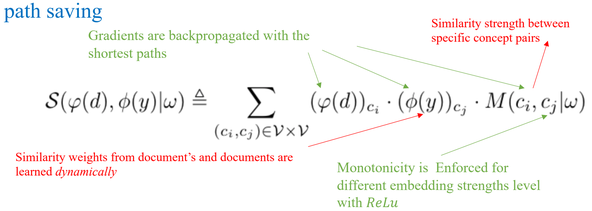


Associating concept in sematic embedding space

- Assigning documents to the correct taxonomy nodes based on similarity aggregation over concept representations



- Mapping from similarities to relevance score is learned end-to-end using *dynamic bin-pooling* with monotonicity enforcement and gradient path saving



$$\mathcal{S}(\varphi(d), \phi(y)|\omega) \triangleq \sum_{(c_i,c_j)\in\mathcal{V}\times\mathcal{V}} (\varphi(d))_{c_i} \cdot (\phi(y))_{c_j} \cdot M(c_i, c_j|\omega)$$

Evaluation

- **We extensively evaluate our approach for Computer Science + Physics & Math + Medicine with > 60 hierarchical categories and a maximum height of 5**
- **Our approach significantly outperform the state of baseline methods including WeSHClass, Pre-trained Bert, UNEC, Dataless**
- **Document's relevance to taxonomy nodes can also be visualized as a combination of the individual concepts' relevance weighted by attention**