Recommended projects:

0. Distantly Supervised Named Entity Recognition w. Contextualized Representation

Abstract:
As most popular Named Entity Recognition (NER) systems require hand-crafted features and human annotations, recent progress on neural network and knowledge bases allows us to capture textual feature & conduct training in a data-driven manner. In this project, we would focus on integrating contextualized representation models with distant training. Specifically, the constructed model would first construct a powerful text representation model with massive text, then proceed to downstream tasks (i.e., NER) with distant supervision (instead of human annotations).

Requirements:
Motivated towards quality research/publication; proficient in python; familiar with pytorch; eager to learn and play with neural network models.

Contact: Lucas Liu (lychinalz@gmail.com)

1. Attribute-Aware Conditional Network Structure Learning

Abstract:
Network structure learning studies the problems of modeling and generating networks, which is fundamental for the understanding of network functional structures and mimicking real-world networks with special properties. Nowadays, attributes are becoming ubiquitous in networks, which may often reveal node properties leading to particular link structures. In this work, we focus on the learning of network structures under certain conditions indicated by associative attributes in the networks. For example, in a biomedical network, certain genes and proteins may interact in related but different ways for individuals in different age groups or with different diseases. To this end, we will construct generative graphical models with manipulatable latent factors, probably by combining recurrent neural networks with variational inference. We will train the models on several preprocessed attributed networks, and comprehensively evaluate them with experiments.

Requirements:
Motivated towards quality research/publication; proficient in python; familiar with pytorch/tensorflow and embedding; eager to learn and play with neural network models.

Contact: Carl Yang (jiyang3@illinois.edu)

2. Text-Rich Network Embedding with Smart Sampling

Abstract:
While many popular network embedding algorithms leverage the skip-gram model from word embedding, the study of proper ways to sample the network has shown to be important, which is a key step to construct the skip-gram objective. It is intuitive to ask questions like how long should the random walks be, and which neighbors should be more important. To answer these questions, in this work, we aim to leverage the widely available rich texts in networks. For example, nodes with rich texts might require less sampled structural contexts, and neighbors with different texts might require more samples. To this end, we will combine multi-armed bandit model, self-paced learning and attention structures to enable text-guided smart sampling, so that optimization and sampling can be mutually enhanced in a closed loop. We will train the models on several preprocessed text-rich networks, and comprehensively evaluate them with experiments.

Requirements:
Motivated towards quality research/publication; proficient in python; familiar with pytorch/tensorflow and embedding; eager to learn and play with neural network models.

Contact: Carl Yang (jiyang3@illinois.edu)
3. Heterogeneous Network Embedding Benchmarks

Abstract:

Network embedding has been attracting intensive research attention recently. However, unlike homogeneous networks, there is a lack of quality survey and standard benchmarks for heterogeneous network embedding. In this work, we will build an open-source system upon our existing rich surveys, datasets and implementations of heterogeneous network embedding. In the meantime, we will also systematically study the advanced network embedding techniques, understand their mathematical foundation and essential leverage, as well as analyze their power towards different real-world network mining tasks. This could be a very good start point for students who have a long-term interest in academia, especially research on network mining related topics.

Requirements:

Long-term interested towards network mining; proficient in python; eager to learn and program.

Contact: Carl Yang (jiyang3@illinois.edu)

4. Wide-Window Pattern Discovery for Open Information Extraction

Abstract:

Open information extraction aims at extracting all the relation tuples from the corpus without requiring pre-specified relation types with no or little supervision. The existing methods extract ill-structured or incomplete information and often fail on the corpus with long and complicated sentence structures. In this work, we will build a novel pattern-based information extraction method for the wide-window pattern discovery in text. The three major challenges are (1) the long sentences with long-distanced entity mentions, (2) the hierarchical or n-ary relations among long-distanced entities mentioned in one sentence, and (3) the completeness of extractions. We have some existing work in this direction, and we are looking for students interested in text mining and information extraction.

Requirements:

Motivated towards quality research/publication; proficient in python; eager to learn and program.

Contact: Xuan Wang (xwang174@illinois.edu)

5. Pattern-enhanced Named Entity Recognition with Distant Supervision (full)

Abstract:

State-of-the-art Named Entity Recognition (NER) systems apply supervised machine learning models, i.e., relying on human effort for training data annotation, which leads to highly specialized systems that are hard to be directly used to recognize new types of entities. Moreover, meta-patterns discovered from text carry the information of long-distance label dependency that is not captured by current models for label prediction. In this work, we will build an NER model that recognizes entities from text with distant supervision from open knowledge bases. The major challenges include (1) how to use the textual patterns to automatically generate a high-quality training corpus for NER, and (2) how to incorporate the long-distance label dependency into the neural model for label prediction. We have some existing work in this direction, and we are looking for students interested in text mining and information extraction.

Requirements:

Motivated towards quality research/publication; proficient in python; eager to learn and play with neural network models.

Contact: Xuan Wang (xwang174@illinois.edu)

6. Ontology-Guided Information Retrieval

Abstract:

Information retrieval (IR) is the age-long task of ranking and retrieving text based on user-generated queries. Ontologies, or pre-defined relational graphs for words, entities or concepts, serve as additional sources of information that are orthologous to (i.e. cannot be inferred from) the corpus. Previous work has shown that proper use of ontological information can improve the quality of retrieved documents. Unfortunately, when the ontology becomes deep and domain-specific, we observe that some state-of-the-art practices (e.g. hierarchy embedding) may fail. In this project, we will explore how to utilize deep and domain-specific ontologies to enhance information retrieval.

Requirements:

Motivated towards quality research/publication; proficient in python; eager to learn and program; experience in IR is preferred but not required.
7. Pattern-enhanced N-ary Relation Classification

Abstract:

Previous studies on relation extraction mainly focus on binary relations (i.e., relations between two entities). However, several high-valued domains (e.g., healthcare) often consider extracting facts/claims connecting multiple entities. For such complicated relations, it is expensive and time-consuming to annotate training samples. In this project, we study N-ary relation extraction under weak supervision. We aim to adopt a co-training framework with a pattern module (e.g., tree patterns, meta-patterns with skips) and a classification module (e.g., tree LSTMs, biDAG LSTMs), where the two modules can benefit each other. We will build the model and conduct experiments on benchmark datasets. Related papers can be found at https://docs.google.com/document/d/1T354rf5k19E4v9hBblp0c_j1YcDiKZ2XwDcaMI/XYSHU/edit?usp=sharing.

Requirements:

Motivated towards quality research/publication; familiar with pytorch; familiar with basic NLP concepts

Contact: Yu Zhang (yuz9@illinois.edu)