CS520: KNOWLEDGE GRAPHS
Data Models, Knowledge Acquisition, Inference, Applications

Lectures and Invited Guests

Spring 2021, Tu/Thu 4:30-5:50, cs520.Stanford.edu

Learn about the basic concepts, latest research & applications
Knowledge Graphs Seminar

- What is a Knowledge Graph?
- How to Create a Knowledge Graph?
- How to Reason with and Access Knowledge Graphs?
- Applications
Knowledge Graphs Seminar

• What is a Knowledge Graph?
• How to Create a Knowledge Graph?
  • How to design the schema?
  • Creating a KG from data
  • Create a KG from text and images
• How to Reason with and Access Knowledge Graphs?
• Applications
How to create a Knowledge Graph from Text?

• Part I: Methods
• Part II: Application
Knowledge Graphs

How to Create a Knowledge Graph from Text?
Part I: Methods
Outline

• Overview
• Language Models
• Entity Extraction
• Relation Extraction
• Summary
Overview

• Lot of valuable information is available in text
  • SEC Filings
  • Wall Street Journal
  • Financial News

• We can use natural language processing (NLP) for information extraction
  • This module is not in-depth discussion of NLP
  • Focus is to use NLP as a black box in service of KG construction
Albert Einstein was a German-born theoretical physicist who developed the theory of relativity.
Overview

• Key tasks in Information Extraction
  • Entity extraction
    • People, Companies, Places, etc.
  • Relation extraction
    • works_for, has_location, has_address
  • Entity resolution
    • “John Smith”, “He”, “The company president”
Overview

• Key tasks in Information Extraction
  • **Entity extraction**
    • People, Companies, Places, etc.
  • **Relation extraction**
    • works_for, has_location, has_address
  • Entity resolution
    • “John Smith”, “He”, “The company president”

Language Models
Language Models

• A language model predicts what word comes in the text next
  • Given: “students opened their .....”
  • Predict the next word: books, laptops, exams, etc.
Language Models

• A language model predicts what word comes in the text next
  • Given: A set of words $x_1,\ldots,x_{n-1}$
  • Predict: $P(x_n \mid x_1,\ldots,x_{n-1})$
Language Models

• Practical Applications
  • Autocompleting search queries
  • Auto completion while typing on a phone
Language Models

• Created using deep learning models
  • Recurrent Neural Networks is a popular approach

• Several variations of pre-trained language models are available
  • Data used for training
  • Single direction or Bi-direction
  • Specific neural architecture used

• Available off-the-shelf and can be adapted for task at hand

A popular language model: BERT (Bidirectional Encoder Representations from Transformers)
Entity Extraction

- Example
- Approaches to Entity Extraction
- Challenges
Example

Cecilia Love, 52, a retired police investigator who lives in Massachusetts, said she paid around $370 a ticket with tax for nonstop United Airlines flights to Sacramento from Boston for her niece's high school graduation in June, 2020.
Example

Cecilia Love, 52, a retired police investigator who lives in Massachusetts, said she paid around $370 a ticket with tax for nonstop United Airlines flights to Sacramento from Boston for her niece's high school graduation in June, 2020.

A named entity is anything that can be referred to using a proper name
• Places, Companies, People, etc.
• Extended to include dates, times, numerical expressions
Example

Cecilia Love, 52, a retired police investigator who lives in Massachusetts, said she paid around $370 a ticket with tax for nonstop United Airlines flights to Sacramento from Boston for her niece's high school graduation in June, 2020.

Cecilia Love, 52, a retired police investigator who lives in New Jersey, said she paid around $370 a ticket with tax for nonstop United Airlines flight to Sacramento from Boston for her niece's high school graduation in June, 2020.
Approach to Entity Extraction

• We view entity extraction as a sequence labeling problem
  • For each word in the input, assign a label from [B, E, I, O, S]
    • B – First word in the entity
    • E – Last word in the entity
    • I – Internal word in the entity
    • O – Word not in the entity
    • S – Single word entity
Approach to Entity Extraction

• We view entity extraction as a sequence labeling problem

  • For each word in the input, assign a label from [B, E, I, O, S]
Approach to Entity Extraction

• Three broad categories of approaches
  • Sequence Labeling
  • Neural Models
  • Rule-based
Approach to Entity Extraction

• Sequence labeling
  • Train a machine learning algorithm (e.g., Conditional Random Fields) using features such as:
    • Part of speech
    • Presence in a named entity list
    • Word embedding
    • Word prefix
    • Whether the word is in all CAPS

Significant Feature Engineering is Required
Approach to Entity Extraction

• Adapt a Language Model
  • Task-independent training
    • Train the model on the domain of interest
  • Task-dependent training
    • Introduce special tags in the input

[CLS] Cecilia Love [SEP], 52, a retired police investigator who lives in [CLS] New Jersey [SEP], said she ......
Approach to Entity Extraction

• Adapt a Language Model
  • Task-independent training
    • Train the model on the domain of interest
  • Task-dependent training
    • Introduce special tags in the input

[CLS] Cecilia Love [SEP], 52, a retired police investigator who lives in [CLS] New Jersey [SEP], said she ......

Language model now predicts the occurrence of a distinguished token
Approach to Entity Extraction

• Rule-based approach
  • Express the extraction rules in a formal rule language
  • The rules can be based on
    • Regular expressions
    • References to dictionary
    • Invoke custom extractors
Challenges in Entity Extraction

• Ambiguity
  • Louis Vuitton – Can be company, person, or product

• Training Data
  • Data is usually small and incomplete

• Domain-specific Variations
  • Duplication of a cell by fission
  • Attach

• Many different forms of an entity
  • Need to have a lexicon
Relation Extraction

• Examples
• Approaches to Relation Extraction
• Challenges
Relation Extraction

Cecilia Love, 52, a retired police investigator who lives in Massachusetts, said she paid around $370 a ticket with tax for nonstop United Airlines flights to Sacramento from Boston for her niece's high school graduation in June, 2020.

• Example
  • Cecilia Love *lives in* Massachusetts
  • United Airline *flies from* Boston
  • United Airlines *flies to* Sacramento
Relation Extraction

• Example
  • Extracting information from Wikipedia Infoboxes
Relation Extraction

• Domain-specific relation extraction
  • Unified medical language system
    • *causes, treats, disrupts*
Approaches to Relation Extraction

- Syntactic patterns (or rule-based)
- Supervised learning
- Open information extraction
Approaches to Relation Extraction

- Syntactic patterns (or rule-based)
- Supervised learning
- Open information extraction
  - Does not rely on a designed set of relations

Dante passed away in Ravenna
Approaches to Relation Extraction

- Syntactic patterns (or rule-based)
- Supervised learning
- Open information extraction
  - Does not rely on a designed set of relations
  - Can be difficult to use / understand the relations

Dante passed away in Ravenna
Approach to Relation Extraction

• Syntactic patterns (aka Hearst Patterns)

The bow lute, such as the Bambara ndang, is plucked and has an individual curved neck for each string.
Approach to Relation Extraction

• Syntactic Patterns (aka Hearst Patterns)

The bow lute, such as the Bambara ndang, is plucked and has an individual curved neck for each string.

Even though we have never heard of Bambara ndang, but we can extract that it is a kind of bow lute
## Approach to Relation Extraction

- **Syntactic Patterns**

<table>
<thead>
<tr>
<th>Pattern Name</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>such as</em></td>
<td>... works by authors <em>such as</em> Herric, Goldsmith, and Shakespear ...</td>
</tr>
<tr>
<td><em>or other</em></td>
<td>Bruises, wounds, broken bones, <em>or other</em> injuries ...</td>
</tr>
<tr>
<td><em>and other</em></td>
<td>... temples, treasuries, <em>and other</em> Civic Buildings, ...</td>
</tr>
<tr>
<td><em>including</em></td>
<td>All common law countries <em>including</em> Canada and England ...</td>
</tr>
<tr>
<td><em>especially</em></td>
<td>Most European countries <em>especially</em> France, England, and Spain, ...</td>
</tr>
</tbody>
</table>
Approach to Information Extraction

• **Syntactic Pattern**
  • To discover pattern for a new relation, collect several examples of that relation
  • Look for generalities to discover new patterns
    • Has been difficult to find patterns for some relations, e.g., has part
    • Limited success in automatically learning the patterns
Approach to Relation Extraction

• Supervised learning
  • Requires a huge amount of training data
  • We can use syntactic patterns to generate training data
  • We can also write approximate labeling functions
    • An approximate labeling function for has_part is to produce a dependency parse of a sentence, and look for nodes directly connected by “has” or “have”
    • An approximate labeling function for subclass_of: If two entities end with the same base word but one has an extra modifier (e.g., cell and eukaryotic cell)
Approach to Relation Extraction

• Adapting Language Model

[TERM1-START] Cecilia Love [TERM1-END], 52, a retired police investigator who lives in [TERM2-START] Massachusetts [TERM2-END]

For training data such as the sentences above, we provide the relational label as the output

The model then learns to predict the labels for input sentences as augmented above
Challenges to Relation Extraction

• Training Data
• Human Verification
• Specialized extraction for
  • Events
  • Temporal information
Summary

• Entity extraction and relation extraction are fundamental problems to creating knowledge graphs from text

• Use of rule-based methods for training data generation that can be fed into pre-trained language models is becoming an increasingly popular paradigm

• Entity linking and resolution will eventually play an important role
How to create a Knowledge Graph from Text?

• Part I: Methods
• Part II: Application
Intelligent Textbooks

Create a Knowledge Graph from Text

Part II: Methods
Outline

• What is an Intelligent Textbook
• What Knowledge Graph is required?
  • Quest for meaning
• Entity Extraction
• Relation Extraction
  • Automated relation extraction
• Way forward
  • Knowledge Graph Authoring
What is an Intelligent Textbook?

Traditional Textbook + Knowledge Graph = Intelligent Textbook
Intelligent Textbook

How might we make it easy for students to learn complex new concepts?
Intelligent Textbook

Aniea
1st Year Biology Student

“Biology is complex, book has a huge amount of new words, and I am lost!”
6.5 Large Molecules Enter and Leave a Cell through Vesicles

- Macromolecules and particles enter the cell by endocytosis
- Receptor-mediated endocytosis is highly specific
- Exocytosis moves materials out of the cell

Macromolecules such as proteins, polysaccharides, and nucleic acids are simply too large and too charged or polar to pass through biological membranes. This is actually fortunate—think of the consequences if such molecules diffused out of cells: A red blood cell would not retain its hemoglobin! As you saw in Chapter 5, the development of a selectively permeable membrane was essential for the functioning of the first cells when life on Earth began. The interior of a cell can be maintained as a separate compartment with a different composition from that of the exterior environment, which is subject to abrupt changes. However, cells must sometimes take up or secrete (release to the external environment) intact large molecules. In Key Concept 5.3 we described phagocytosis, the mechanism by which solid particles can be brought into the cell by means of vesicles that pinch off from the cell membrane. The general terms for
Intelligent Textbook

• Classroom Trials
  • Improve student learning by full letter grade
  • Help under-performing students
What is an Intelligent Textbook

• Demonstration
  • https://youtu.be/2MxZQOUKIdE
proteins, are in contact with the aqueous solution. On the outer surface of the plasma membrane, carbohydrate side chains are found attached to proteins and lipids. The hydrophobic parts, including phospholipid
What knowledge graph is required?

proteins, are in contact with the aqueous solution. On the outer surface of the **plasma membrane**, **carbohydrate side chains** are found **attached to proteins** and **lipids**. The hydrophobic parts, including phospholipid

---

**Logical Meaning**

For every instance of plasma membrane there exists
- a Glycoprotein
- a Carbohydrate Side Chain
such that Carbohydrate Side chain is a part of the Glycoprotein
What knowledge graph is required?

proteins, are in contact with the aqueous solution. On the outer surface of the plasma membrane, carbohydrate side chains are found attached to proteins and lipids. The hydrophobic parts, including phospholipid...

Logical Meaning

- The nodes are generics as opposed to concrete entities
- In contrast to people, places, companies, etc.
- This is more than a directed labeled graph
- Introduction of new objects
What knowledge graph is required?

proteins, are in contact with the aqueous solution. On the outer surface of the **plasma membrane**, **carbohydrate side chains** are found **attached to proteins and lipids**. The hydrophobic parts, including phospholipid

Semantic Meaning

How do biologists define has part and has function?
- Biologists are not used to formal definitions
Meaning of Structure & Function

Structure and function are correlated at all levels of biological organization: *The form fits the function*
Computational Meaning

• Identify the requirements in terms of a set of questions
  • Diagnostic questions
    • Help assess the basics of KR&R
  • Educationally useful questions
    • The question must be of interest to teachers and students
    • The question must be "Google hard"
    • The question should not require solving an open-ended research problem
Diagnostic Questions

• What is the structure of X?
• What is the function of X?
Educationally Useful Questions

• Relate Structures to Functions
  • What structure of Biomembrane facilitates a function of biomembrane, namely phagocytosis?

• Qualitative Comparisons
  • If the Loop of Henle gets longer, how will its function be impacted?

• Detailed Comparisons
  • What is the functional similarity between prions and viroids?

• Similarity Reasoning
  • Glucose is to Glycogen as ATP is to what?

• Negatively Modified Structures Impacting Functions
  • If hydrogen is removed from a saturated fatty acid, then how is its function impacted?
Defining Structure

- Structure of an entity represents its parts, their spatial arrangements and sizes

<table>
<thead>
<tr>
<th>Meronymic</th>
<th>Spatial</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>has-part</td>
<td>is-at</td>
<td>length</td>
</tr>
<tr>
<td>has-region</td>
<td>is-inside</td>
<td>diameter</td>
</tr>
<tr>
<td>material</td>
<td>is-outside</td>
<td>height</td>
</tr>
<tr>
<td>possesses</td>
<td>abuts</td>
<td>area</td>
</tr>
<tr>
<td>element</td>
<td>is-between</td>
<td>depth</td>
</tr>
<tr>
<td></td>
<td>is-along</td>
<td>volume</td>
</tr>
</tbody>
</table>
Defining structural relations

• It must make sense to say ‘X has Y’ in English
• X has-region Y if
  • Y is a region of space defined in relation to X
  • It does not make sense to associate Y with properties such as mass or density, but can be associated with measures such as length, area, or volume
• X has material Y only if
  • Y is tangible and pervasive in X
• X has element Y if
  • X is a set of entities of the same type (or sibling types) that Y is an instance of
• X possesses Y only if
  • Y is Energy, bond or gradient
• Otherwise X has part Y
Outline

• What is an Intelligent Textbook
• What Knowledge Graph is required?
  • Quest for meaning
• Entity Extraction
• Relation Extraction
  • Automated relation extraction
• Way forward
  • Knowledge Graph Authoring
proteins, are in contact with the aqueous solution. On the outer surface of the plasma membrane, carbohydrate side chains are found attached to proteins and lipids. The hydrophobic parts, including phospholipid.

Extract:
- Plasma membrane
- Carbohydrate side chain
- Protein
- Lipid
proteins, are in contact with the aqueous solution. On the outer surface of the **plasma membrane**, **carbohydrate side chains** are found **attached to proteins** and **lipids**. The hydrophobic parts, including **phospholipid**

**Extract**
- Plasma membrane
- Carbohydrate side chain
- Protein
- Lipid

**Where do we get the training data?**
**Why not use the glossary of the textbook?**
Entity Extraction

proteins, are in contact with the aqueous solution. On the outer surface of the plasma membrane, carbohydrate side chains are found attached to proteins and lipids. The hydrophobic parts, including phospholipid

Extract
Plasma membrane
Carbohydrate side chain
Protein
Lipid

Where do we get the training data? Why not use the glossary of the textbook?
Term Extraction

Repeat for each sentence to extract all terms

Term Extraction Training Data
Term extraction training data created by tagging open source textbook sentences using hand-built glossaries.

Model Input: Textbook Sentence
All cells have cell membranes, but only some have cell walls.

NLP Deep Learning Model

Model Output: Textbook Sentence with Tagged Terms
All cells have cell membranes, but only some have cell walls.
Term Extraction
Repeat for each sentence to extract all terms

Term Extraction
Training Data
Term extraction training data created by tagging open source textbook sentences using hand-built glossaries.

Model Input: Textbook Sentence
All cells have cell membranes, but only some have cell walls.

NLP Deep Learning Model

Model Output: Textbook Sentence with Tagged Terms
All cells have cell membranes, but only some have cell walls.

<table>
<thead>
<tr>
<th>Textbook</th>
<th># Sentences</th>
<th># Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenStax Anatomy &amp; Physiology</td>
<td>21706</td>
<td>3196</td>
</tr>
<tr>
<td>OpenStax Astronomy</td>
<td>18844</td>
<td>810</td>
</tr>
<tr>
<td>OpenStax Biology 2e</td>
<td>24544</td>
<td>2757</td>
</tr>
<tr>
<td>OpenStax Chemistry 2e</td>
<td>13799</td>
<td>954</td>
</tr>
<tr>
<td>Life Biology</td>
<td>16673</td>
<td>0</td>
</tr>
<tr>
<td>OpenStax Microbiology</td>
<td>16190</td>
<td>4149</td>
</tr>
<tr>
<td>OpenStax Psychology</td>
<td>9967</td>
<td>1086</td>
</tr>
<tr>
<td>OpenStax Physics Volume I</td>
<td>15005</td>
<td>462</td>
</tr>
<tr>
<td>OpenStax Physics Volume II</td>
<td>11779</td>
<td>466</td>
</tr>
<tr>
<td>OpenStax Physics Volume III</td>
<td>9250</td>
<td>580</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Split</th>
<th>Sources</th>
<th>Sentences</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>All textbooks in table 1 except OpenStax Biology and Life Biology</td>
<td>57634</td>
<td>7167</td>
</tr>
<tr>
<td>Dev</td>
<td>OpenStax Biology Ch. 4 Sect. 2, Ch. 10 Sect. 2 &amp; 4</td>
<td>206</td>
<td>254</td>
</tr>
<tr>
<td>Test</td>
<td>Life Biology Ch. 39</td>
<td>608</td>
<td>369</td>
</tr>
</tbody>
</table>
Term Extraction
Repeat for each sentence to extract all terms

Term Extraction
Training Data
Term extraction training data created by tagging open source textbook sentences using hand-built glossaries.

Model Input: Textbook Sentence
All cells have cell membranes, but only some have cell walls.

NLP Deep Learning Model

Model Output: Textbook Sentence with Tagged Terms
All cells have cell membranes, but only some have cell walls.

<table>
<thead>
<tr>
<th>Term Extraction</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.67</td>
<td>0.51</td>
</tr>
</tbody>
</table>
What are the challenges?

Multiple ways to refer to the same term
- DNA vs Deoxyribose Nucleic Acid
- Membrane vs cell membrane
- Mitochondrion vs mitochondria
What are the challenges?

Multiple ways to refer to the same term
- DNA vs Deoxyribose Nucleic Acid
- Membrane vs cell membrane
- Mitochondrion vs mitochondria

A good lexicon is essential for Term Extraction
What are the challenges?

What exactly is a term?
- Faulty tumor suppressor gene
- Control of blood flow to skin
- Attach, Synthesis

Existing term extraction has a narrow scope
Outline

• What is an Intelligent Textbook
• What Knowledge Graph is required?
  • Quest for meaning
• Entity Extraction
• Relation Extraction
  • Automated relation extraction
• Way forward
  • Knowledge Graph Authoring
proteins, are in contact with the aqueous solution.

On the outer surface of the plasma membrane, carbohydrate side chains are found attached to proteins and lipids.
The hydrophobic parts, including phospholipid

Entities
Plasma membrane
Carbohydrate side chain
Protein
Lipid

Relations
Plasma membrane has part
Carbohydrate side chain abuts
Protein
Lipid
proteins, are in contact with the aqueous solution.
On the outer surface of the plasma membrane, carbohydrate side chains are found attached to proteins and lipids.
The hydrophobic parts, including phospholipid

Entities
- Plasma membrane
- Carbohydrate side chain
- Protein
- Lipid

Relations
- Plasma membrane has region
- outer surface
- abuts, is-outside
- carbohydrate side chain
- protein
- lipid
Relation Extraction

Where do we get the training data?

- Use pre-existing KB
- Use distant supervision
- Use weak supervision

proteins, are in contact with the aqueous solution. On the outer surface of the plasma membrane, carbohydrate side chains are found attached to proteins and lipids. The hydrophobic parts, including phospholipid...
Relation Extraction

Where do we get the training data?
Use pre-existing KB
Use distant supervision
Use weak supervision

1. Define a set of label functions:
   1. \( f(\text{sentence, term pair}) \rightarrow \text{relation or ABSTAIN} \)
2. Apply each of these label functions to every training instance
3. Aggregate these sets of labels into a single label for each instance:
   1. Hard Labels: Majority vote to get the most voted relation as the label
   2. Soft Labels: Use snorkel's label model to combine label functions based on estimated reliability to get a probability distribution across relations
Relation Extraction
Repeat for each term pair to extract all triples

Relation Extraction Training Data
Relation extraction training data created using weak supervision from pattern-based heuristics and existing knowledge bases.

Hand-Built Biology Knowledge Base

Model Output: Confidence Scores for KB Relations

Model Input: Textbook Sentence with Tagged Term Pair
All cells have cell membranes, but only some have cell walls.

Enumerate Term Pairs
(cell, cell wall) (cell, cell membrane) (cell wall, cell membrane)

NLP Deep Learning Model
<table>
<thead>
<tr>
<th>Relation Extraction</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.65</td>
<td>0.54</td>
</tr>
</tbody>
</table>

**Relation Extraction**

Repeat for each term pair to extract all triples

**Relation Extraction Training Data**
Relation extraction training data created using weak supervision from pattern-based heuristics and existing knowledge bases.

**Model Output: Confidence Scores for KB Relations**

**Hand-Built Biology Knowledge Base**

- X have/has Y
- X is a Y
- X, such as Y

**Enumerate Term Pairs**
- (cell, cell wall)
- (cell, cell membrane)
- (cell wall, cell membrane)

**Model Input: Textbook Sentence with Tagged Term Pair**
All cells have cell membranes, but only some have cell walls.
**Term Extraction**
Repeat for each sentence to extract all terms

**Relation Extraction**
Repeat for each term pair to extract all triples

**Term Extraction Training Data**
Term extraction training data created by tagging open source textbook sentences using hand-built glossaries.

**Model Input: Textbook Sentence**
All cells have cell membranes, but only some have cell walls.

**NLP Deep Learning Model**

**Model Output: Textbook Sentence with Tagged Terms**
All cells have cell membranes, but only some have cell walls.

**Relation Extraction Training Data**
Relation extraction training data created using weak supervision from pattern-based heuristics and existing knowledge bases.

**Hand-Built Biology Knowledge Base**

**X have/has Y**

**X is a Y**

**X, such as Y**

**Model Input: Textbook Sentence with Tagged Term Pair**
All cells have cell membranes, but only some have cell walls.

**Enumerate Term Pairs**
- (cell, cell wall)
- (cell, cell membrane)
- (cell wall, cell membrane)

**Model Output: Confidence Scores for KB Relations**
- has-part
- is-outside
- has-region
- subclass-of
Way Forward
Way Forward

Human review to be done by the textbook author
- An integral step in the authoring process
Way Forward

Human review to be done by the textbook author
- An integral step in the authoring process
- Glossary editor
- Diagram editor
Way Forward

Human review to be done by the textbook author
- An integral step in the authoring process
  - Glossary editor
  - Diagram editor

A new kind of professional
Summary

• Entity extraction and relation extraction are fundamental problems to creating knowledge graphs from text

• Use of rule-based methods for training data generation that can be fed into pre-trained language models is becoming an increasingly popular paradigm
  • Human oversight and participation is essential to the process

• Entity linking and resolution will eventually play an important role
April 29, 2021

Aditya Kalyanpur

Creating Causal Knowledge Graphs for Language Understanding

Ranjay Krishna

Scene graphs for image understanding