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

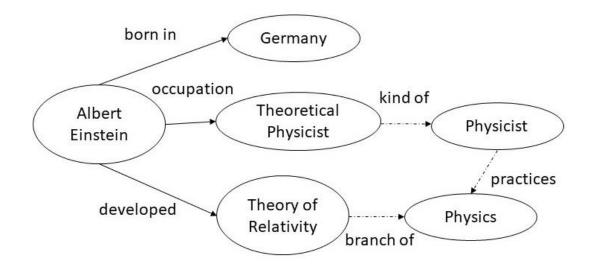
Overview

- 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

Natural Language Processing

Entity Extraction

Albert Einstein was a Germanborn theoretical physicist who developed the theory of relativity. Relation Extraction



Question Answering Common Sense Reasoning

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

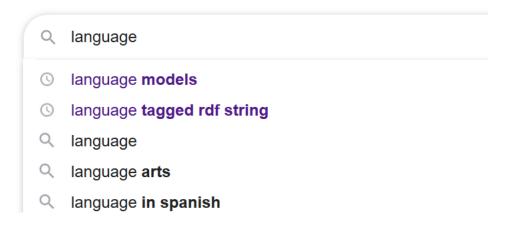
- Key tasks in Information Extraction
 - Entity extraction
 - People, Companies, Places, etc.
 - Relation extraction
 - works_for, has_location, has_address
 - Entity resolution
 - "John Smith", "He", "The company president"

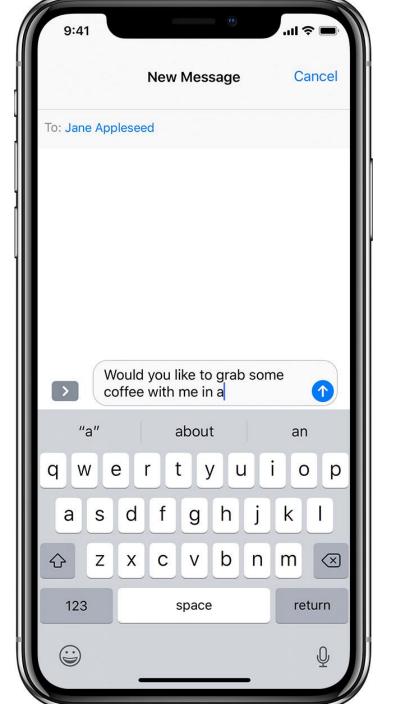
- A language model predicts what word comes in the text next
 - Given: "students opened their"
 - Predict the next word: books, laptops, exams, etc.

- A language model predicts what word comes in the text next
 - Given: A set of words x₁,...,x_{n-1}
 - Predict: $P(x_n | x_1,...,x_{n-1})$

- Practical Applications
 - Autocompleting search queries
 - Auto completion while typing on a phone







- 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.

[PER Cecilia Love], 52, a retired police investigator who lives in [LOC New Jersey], said she paid around [MONEY \$370] a ticket with tax for nonstop [ORG United Airlines] flight to [LOC Sacramento] from [LOC Boston] for her niece's high school graduation in [TIME June, 2020].

- 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

- We view entity extraction as a sequence labeling problem
 - For each word in the input, assign a label from [B, E, I, O, S]

Cecilia	В	Love	E	,	O	52	O	,	O
a	Ο	retired	O	police	O	investigato	r O	who	O
lives	O	in	О	Massachuset	ts S	,	O	said	O
she	O	paid	O	around	O	\$370	S	a	O
ticket	O	with	Ο	tax	O	for	O	nonstop	O
United	В	Airlines	Е	flights	O	to	O	Sacramento	S
from	О	Boston	S	for	O	her	O	niece's	O
high	O	school	O	graduation	O	in	O	June	В
,	I	2020	E						

- Three broad categories of approaches
 - Sequence Labeling
 - Neural Models
 - Rule-based

- 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

- 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

- 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

- 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 Vouitton 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

- Examples
- Approaches to Relation Extraction
- Challenges

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 <u>lives in</u> Massachusetts
 - United Airline <u>flies from</u> Boston
 - United Airlines <u>flies to</u> Sacramento

- Example
 - Extracting information from Wikipedia Infoboxes

Larry King



King in March 2017

Born Lawrence Harvey Zeiger^[1]

November 19, 1933 (age 86)

Brooklyn, New York, U.S.

Education Lafayette High School

Occupation Radio and television personality

Years active 1957-present

Spouse(s) Freda Miller

(m. 1952; ann. 1953)

Annette Kaye

(m. 1961; div. 1961)

Alene Akins

(m. 1961; div. 1963, m. 1967; div.

1972)

Mickey Sutphin (m. 1963; div. 1967)

Sharon Lepore (m. 1976; div. 1983)

Julie Alexander (<u>m.</u> 1989; <u>div.</u> 1992)

Shawn Southwick (m. 1997; sep. 2019)

- Domain-specific relation extraction
 - Unified medical language system
 - causes, treats, disrupts

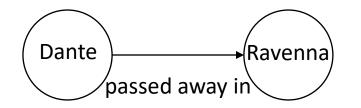
Approaches to Relation Extraction

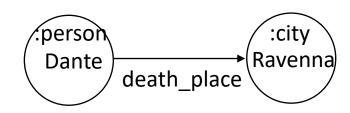
- Syntactic patterns (or rule-based)
- Supervised learning
- Open information extraction

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

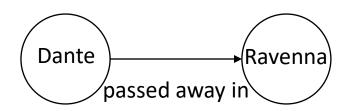


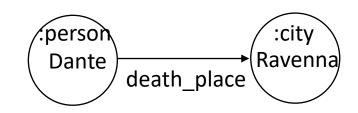


Approaches to Relation Extraction

- 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

Pattern Name	Example
such as	works by authors such as Herric, Goldsmith, and Shakespear
or other	Bruises, wounds, broken bones, or other injuries
and other	temples, treasuries, and other Civic Buildings,
including	All common law countries including Canada and England
especially	Most European countries especially France, England, and Spain,

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

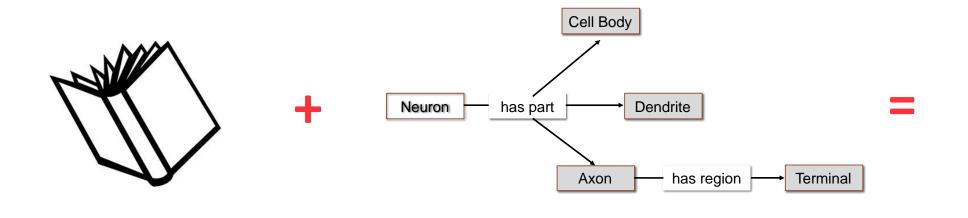
Create a Knowledge Graph from Text

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!"

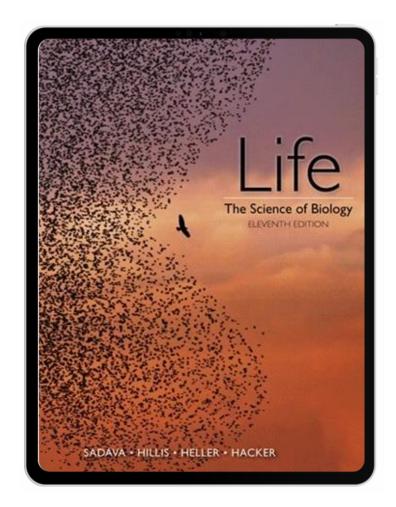
6.5 Large Molecules Enter and Leave a Cell through Vesicles

- , 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



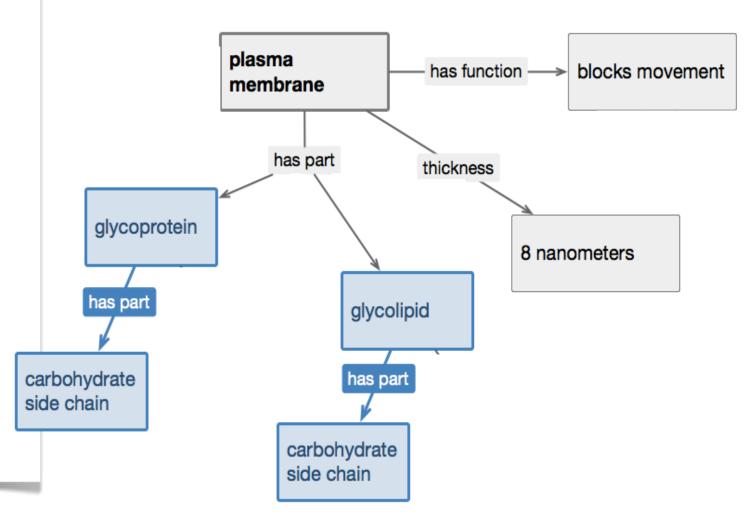
What is an Intelligent Textbook

- Demonstration
 - https://youtu.be/2MxZQOUKIdE

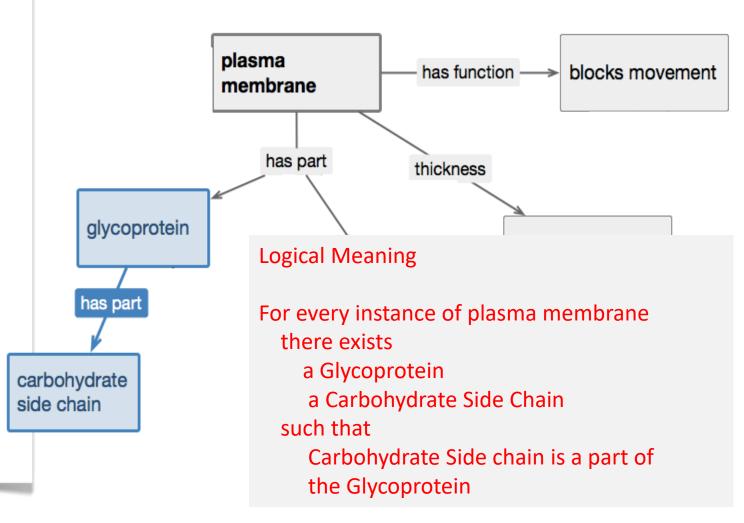
proteins, are in contact with the aqueous solution.

On the outer surface of the plasma membrane, carbohydrate side chains are found attached to proteins and lipids.

The hydrophobic parts, including phospholipid



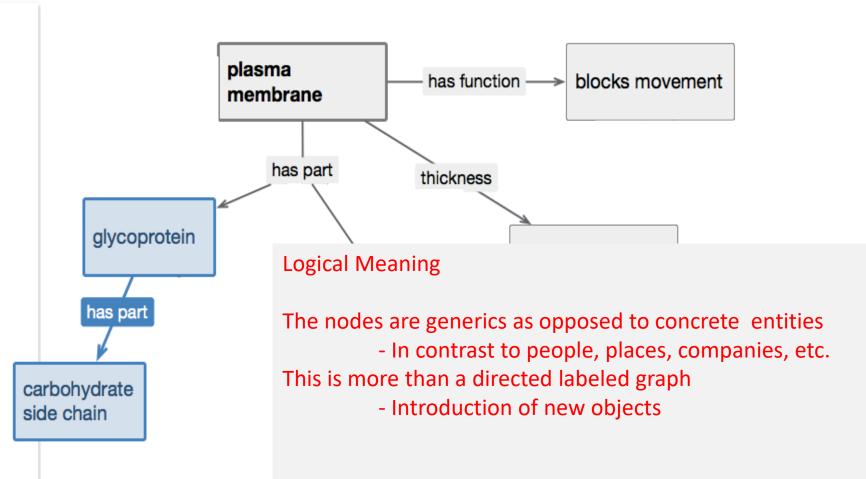
proteins, are in contact with the aqueous solution. On the outer surface of the plasma membrane, carbohydrate side chains are found attached to proteins and lipids. The hydrophobic parts, including phospholipid



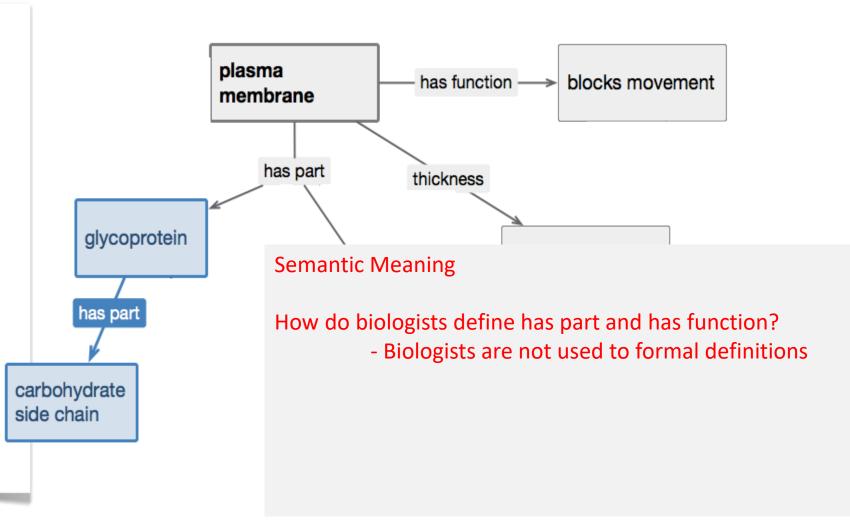
proteins, are in contact with the aqueous solution.
On the outer surface of

the plasma membrane, carbohydrate side chains are found attached to proteins and lipids.

The hydrophobic parts, including phospholipid

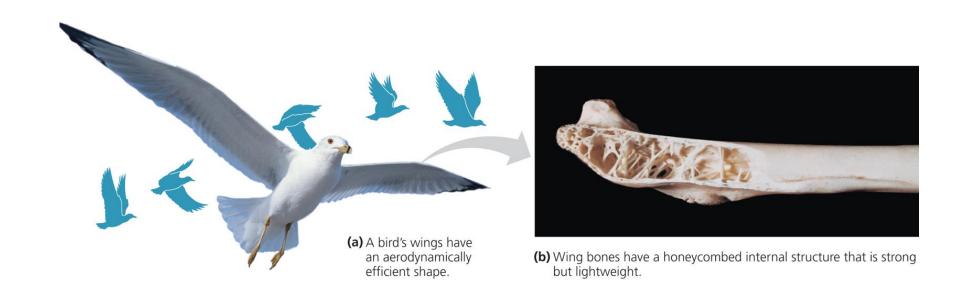


proteins, are in contact with the aqueous solution. On the outer surface of the plasma membrane, carbohydrate side chains are found attached to proteins and lipids. The hydrophobic parts, including phospholipid



Meaning of Structure & Function

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



FIGURES FROM BIOLOGY (9^{TH} EDITION) BY NEIL A. CAMPBELL AND JANE B. REECE. COPYRIGHT © 2011 BY PEARSON EDUCATION, INC. USED BY PERMISSION OF PEARSON EDUCATION, INC.

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

Meronymic	Spatial	Properties
has-part	is-at	length
has-region	is-inside	diameter
material	is-outside	height
possesses	abuts	area
element	is-between	depth
	is-along	volume

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

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

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

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

Plasma membrane
Carbohydrate side chain
Protein
Lipid

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

Repeat for each sentence to extract all terms

Term Extraction Training Data

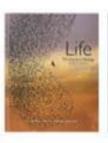
Term extraction training data created by tagging open source textbook sentences using hand-built glossaries.













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.

Repeat for each sentence to extract all terms

Term Extraction Training Data

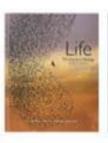
Term extraction training data created by tagging open source textbook sentences using hand-built glossaries.













Model Input: Textbook Sentence

All cells have cell membranes, but only some have cell walls.







Model Output: Textbook Sentence with Tagged Terms

All cells have cell membranes, but only some have cell walls.

Textbook	# Sentences	# Terms
OpenStax Anatomy & Physiology	21706	3196
OpenStax Astronomy	18844	810
OpenStax Biology 2e	24544	2757
OpenStax Chemistry 2e	13799	954
Life Biology	16673	0
OpenStax Microbiology	16190	4149
OpenStax Psychology	9967	1086
OpenStax Physics Volume I	15005	462
OpenStax Physics Volume II	11779	466
OpenStax Physics Volume III	9250	580

Split	Sources	Sentences	Terms
Train	All textbooks in table 1 except OpenStax Biology and Life Biology	57634	7167
Dev	OpenStax Biology Ch. 4 Sect. 2, Ch. 10 Sect. 2 & 4	206	254
Test	Life Biology Ch. 39	608	369

Repeat for each sentence to extract all terms

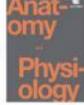
Term Extraction Training Data

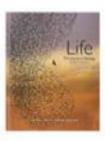
Term extraction training data created by tagging open source textbook sentences using hand-built glossaries.













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.

	Precision	Recall
Term Extraction	0.67	0.51

Repeat for each sentence to extract all terms

Term Extraction Training Data

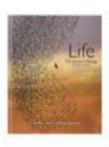
Term extraction training data created by tagging open source textbook sentences using hand-built glossaries.













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.

What are the challenges?

Multiple ways to refer to the same term

- DNA vs Deoxyribose Nucleic Acid
- Membrane vs cell membrane
- Mitochondrion vs mitochondria

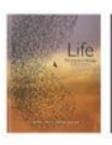
Repeat for each sentence to extract all terms

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. 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

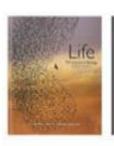
Repeat for each sentence to extract all terms

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. 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

Relation 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

```
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

```
Plasma membrane
 Carbohydrate side chain
 Protein
 Lipid
Relations
 Plasma membrane
   has region
      outer surface
        abuts, is-outside
            carbohydrate side chain
             protein
```

lipid

Entities

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

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

Where do we get the training data?

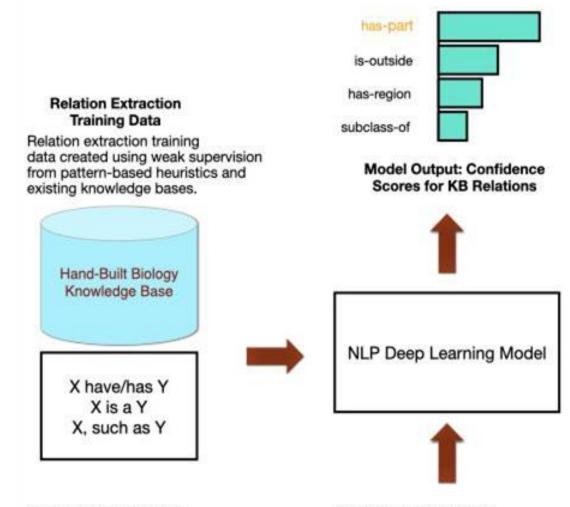
Use pre-existing KB

Use distant supervision

Use weak supervision

- 1.Define a set of label functions:
 - 1. f(sentence, term pair) -> 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:
 - Hard Labels: Majority vote to get the most voted relation as the label
 - Soft Labels: Use snorkel's label model to combine label functions based on estimated reliability to get a probability distribution across relations

Repeat for each term pair to extract all triples



Enumerate Term Pairs

(cell, cell wall) (cell, cell membrane) (cell wall, cell membrane)



Model Input: Textbook Sentence with Tagged Term Pair

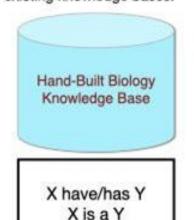
All cells have cell membranes, but only some have cell walls.

	Precision	Recall
Relation Extraction	0.65	0.54

Repeat for each term pair to extract all triples

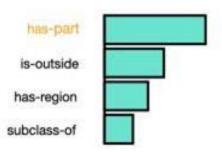
Relation Extraction Training Data

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



X, such as Y





Model Output: Confidence Scores for KB Relations



NLP Deep Learning Model



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.

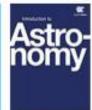
Term Extraction

Repeat for each sentence to extract all terms

Term Extraction Training Data

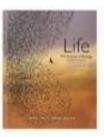
Term extraction training data created by tagging open source textbook sentences using hand-built glossaries.













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

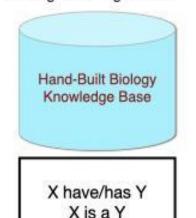
All cells have cell membranes, but only some have cell walls.



Repeat for each term pair to extract all triples

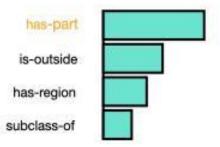
Relation Extraction Training Data

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



X, such as Y





Model Output: Confidence Scores for KB Relations



NLP Deep Learning Model



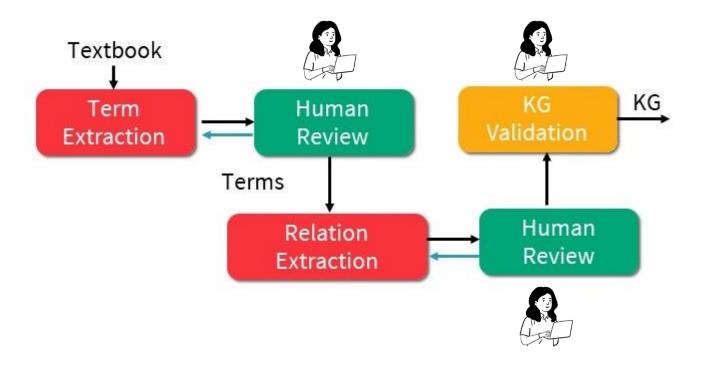
Enumerate Term Pairs

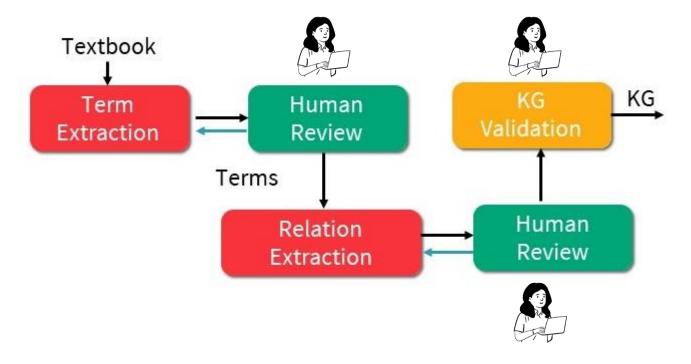
(cell, cell wall) (cell, cell membrane) (cell wall, cell membrane)



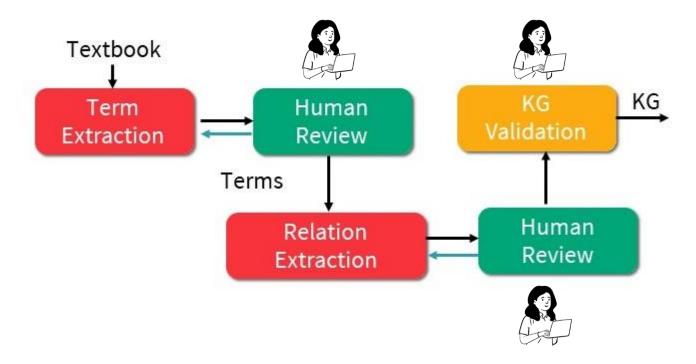
Model Input: Textbook Sentence with Tagged Term Pair

All cells have cell membranes, but only some have cell walls.



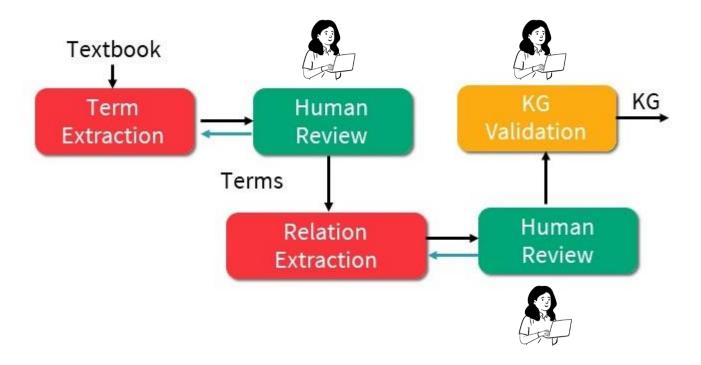


Human review to be done by the textbook author
- An integral step in the authoring process



Human review to be done by the textbook author

- An integral step in the authoring process
 - Glossary editor
 - Diagram editor



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

Ranjay Krishna



Creating Causal Knowledge Graphs for Language Understanding



Scene graphs for image understanding