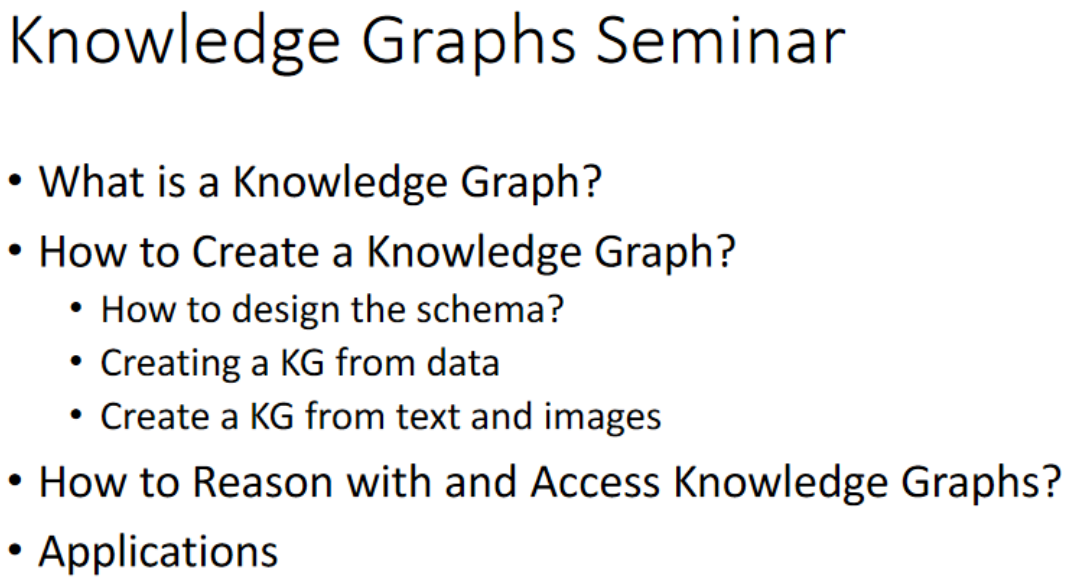
W5. Lecture Notes — By Junyi



Overview of the course

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

How to create a Knowledge Graph from Text?

▼ •Part I: Methods

▼ •Part II: Application

I. Methods

Overview

- lot of valuable information is available in text
- information extraction with NLP (this module treat NLP as a black box)
  - Key Tasks
    - **entity extraction**
    - **relation extraction** → language models
    - entity resolution

Language Models

- predict what word comes next

- created using deep learning models: (popular) RNN
- several variances of pre-trained language models are avaliable
  - training data
  - single direction/bi-direction
  - specific neural architecture used
- available off-the-shelf and can be adapted for task at hand

Entity Extraction

▼ 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].

▼ approaches

▼ sequence labeling

▼ 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

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

▼ Train a machine learning algorithm (e.g., Conditional Random Fields) using features (Significant Feature Engineering is Required)

▼ adapt a language model, neural models

▼ task-dependent training

Train the model on the **domain of interest**

▼ task-independent training

Introduce special tags in the input

Language model now predicts the occurrence of a distinguished token

▼ rule-based

- express the extraction rules in a formal rule language
- rules can be based on
  - regular expressions
  - references to dictionary
  - invoke custom extractors

▼ challenges

- ambiguity
- usually small and incomplete training data
- domain-specific varaincetions
- different forms of an entity → lexicon

Relation Extraction

▼ Example

- unified medical language system: causes, treats, disrupts

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

▼ Approaches

▼ syntactic patterns (/ rule-based)

▼ Example

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, ...

▼ To discover pattern for a new relation, collect several examples of that relation

- Has been difficult to find patterns for some relations, e.g., has part
- Limited success in automatically learning the patterns

▼ supervised learning

▼ requires a huge amount of training data →

▼ can use syntactic patterns to generate training data

▼ can write approximate labeling functions

▼ open information extraction

- does not rely on a designed set of relations
- can be difficult to use/understand the relations

▼ Challenges

- training data
- human verification
- specialized extraction for events and temporal information

II. Application

▼ What is an Intelligent Textbook

▼ What Knowledge Graph is required

- •Quest for meaning
  - logic meaning
  - semantic meaning
  - meaning of structure & function
  - computational meaning
- identify requirements in terms of a set of questions
  - diagnostic questions
    - what's the structure/function of x?
  - educationally useful questions
    - requirements
      - be of interest to teachers and students
      - Google hard
      - should not require solving an open-ended research problem
    - Examples
      - relate structures to functions

        What structure of Biomembrane facilitates a function of biomembrane, namely phagocytosis?

        - 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

- 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?

▼ •Entity Extraction

▼ Where do we get the training data

▼ challenges

▼ Multiple ways to refer to the same term → A good lexicon is essential for Term Extraction

▼ What exactly is a term? → Existing term extraction has a narrow scope

▼ •Relation Extraction

•Automated relation extraction

▼ Where do we get the training data

▼

▼ •Way forward

▼ •Knowledge Graph Authoring: Human review to be done by the textbook author