From Graph to Knowledge Graph: Algorithms and Applications

Module 4: Knowledge Graph Fundamentals and Construction

Outline

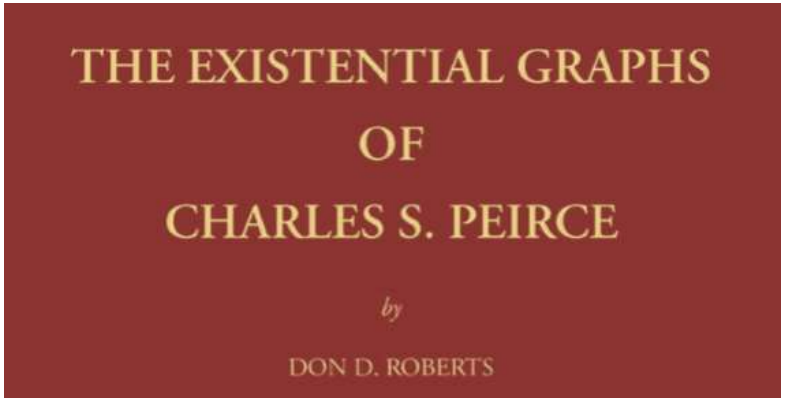
- Knowledge graph fundamentals
  - A brief history of knowledge graph
  - Knowledge graph representation
- Knowledge graph construction
  - How to identify / recognize entities (nodes)
    - Named Entity Recognition
    - Entity Linking
  - How to obtain relationships (edges)
    - Relation extraction

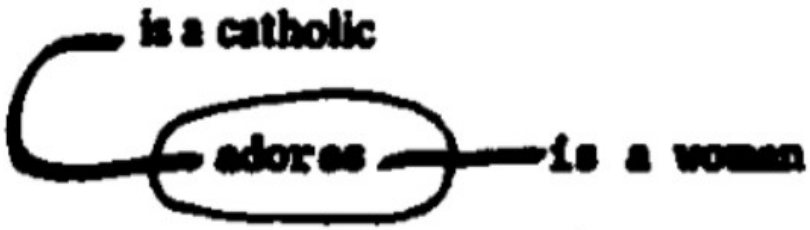
Outline

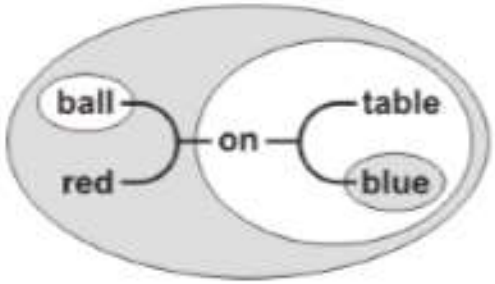
- Knowledge graph fundamentals
  - ***A brief history of knowledge graph***
  - Knowledge graph representation
- Knowledge graph construction

A Brief History of Knowledge Graph

- Charles Peirce's existential graphs (1882 - 1914)



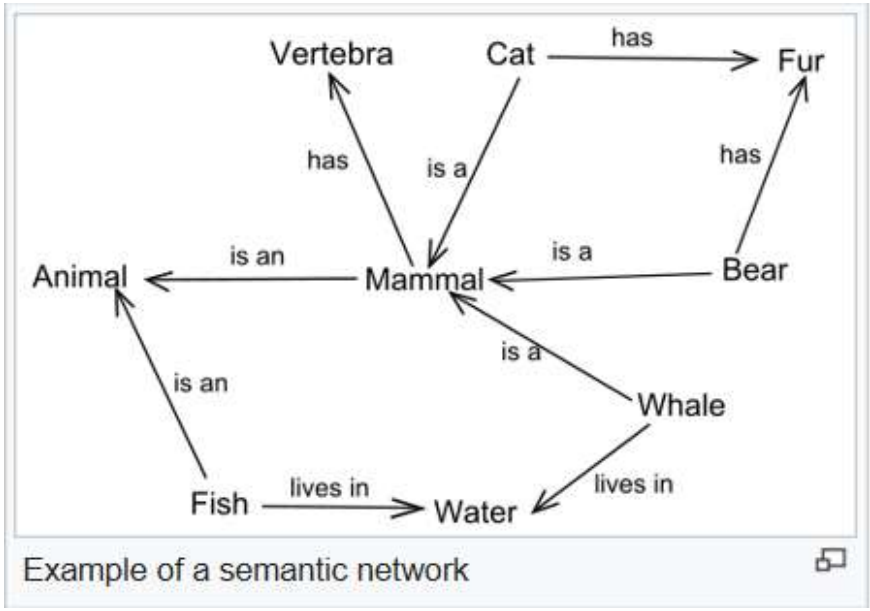




https://en.wikipedia.org/wiki/Existential_graph

A Brief History of Knowledge Graph

- Semantic network (name coined at 1956 by Richard Richens)



Example of a semantic network

Image credit: https://en.wikipedia.org/wiki/Semantic_network

A Brief History of Knowledge Graph

- Linked Data (name coined at 2006 by Tim Berners-Lee)

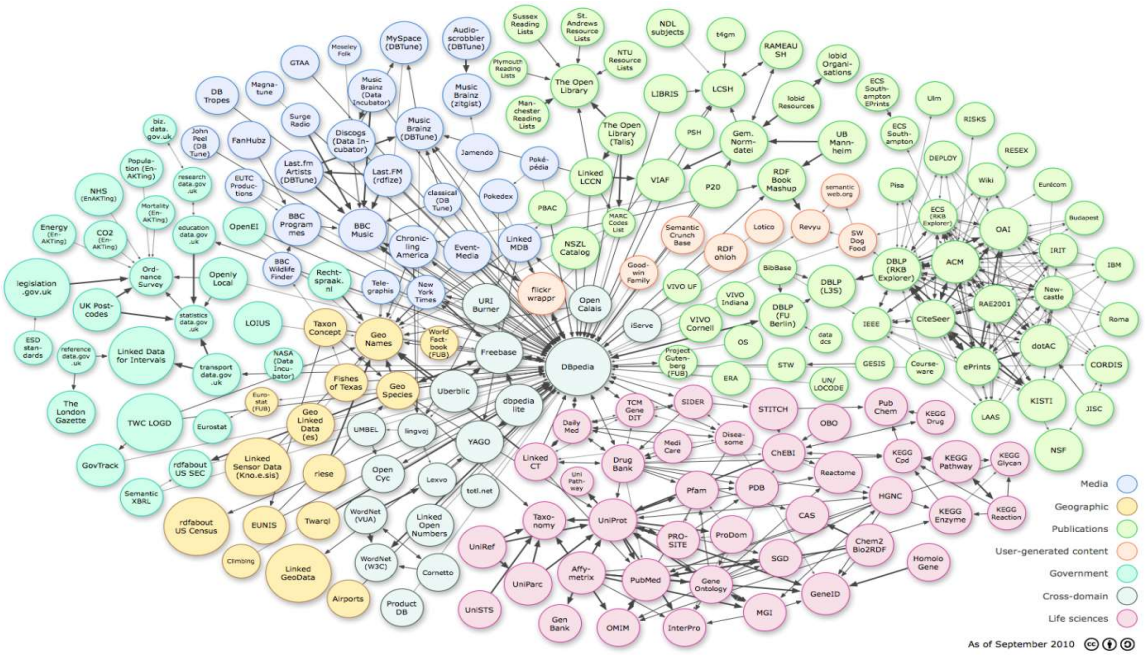


Image credit: https://en.wikipedia.org/wiki/Linked_data

Semantic Web project

World Wide Web Consortium (W3C)

A Brief History of Knowledge Graph

- Expert systems (1970s)

Knowledge Base

Inference Engine

Knowledge base

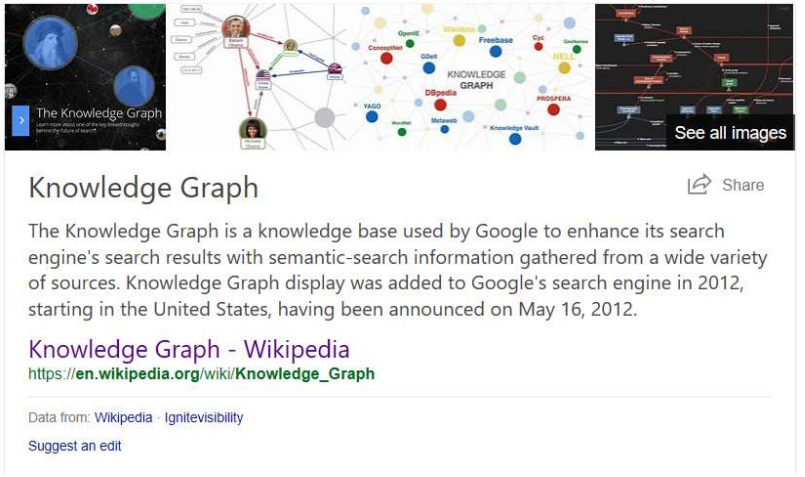
Share

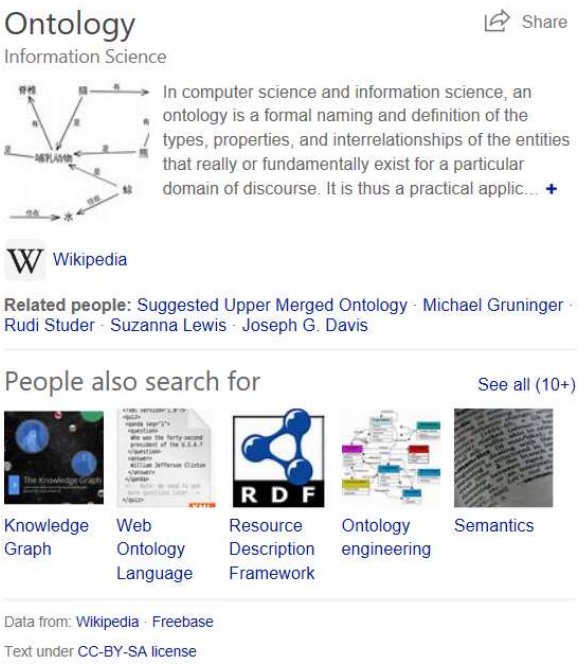
A knowledge base (KB) is a technology used to store complex structured and unstructured information used by a computer system. The initial use of the term was in connection with expert systems which were the first knowledge-based systems.

Knowledge base - Wikipedia
https://en.wikipedia.org/wiki/**Knowledge_base**

A Brief History of Knowledge Graph

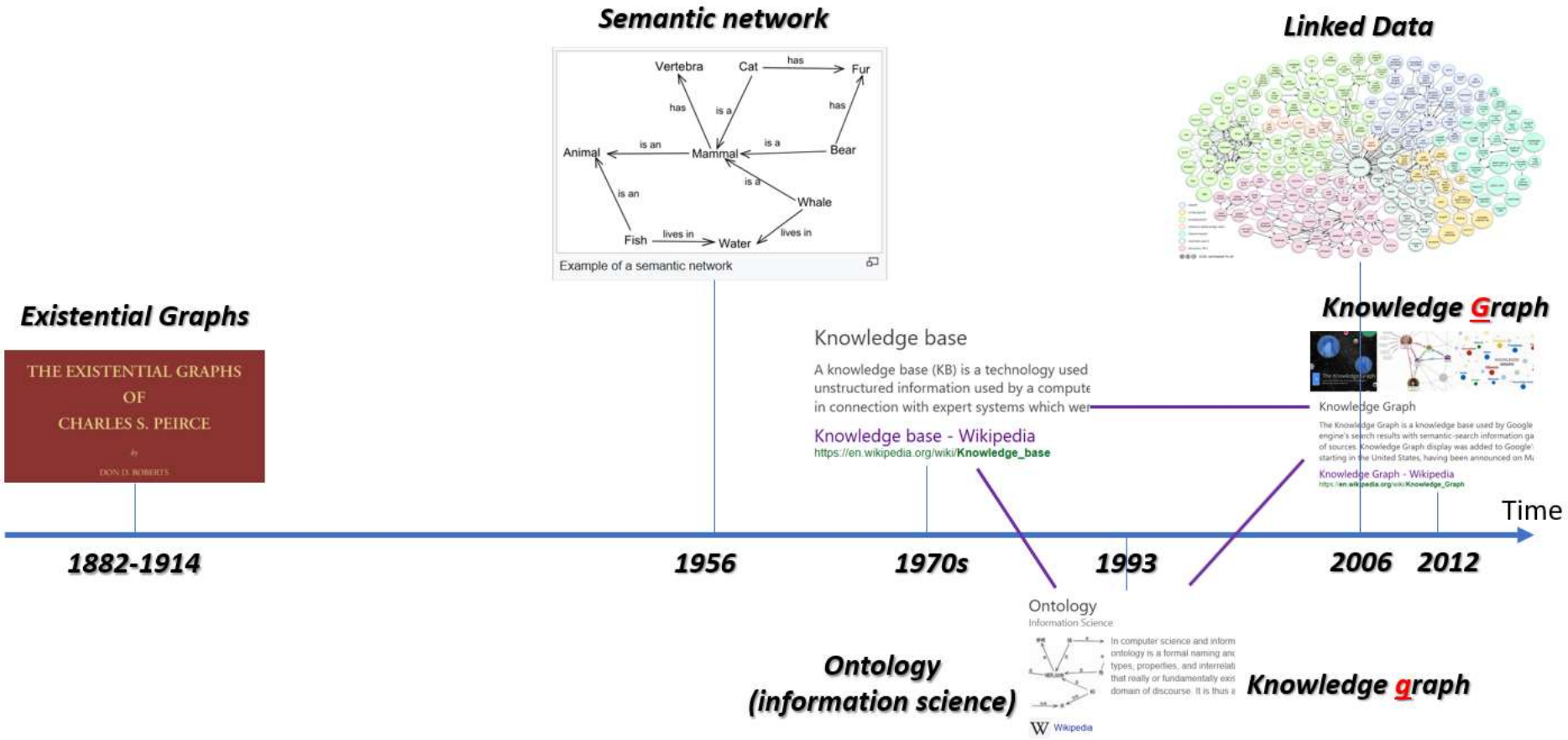
- Knowledge ***G***raph (by Google in 2012)
  - Google's knowledge base





- Knowledge ***g***raph
  - Ontology (information science)

A Brief History of Knowledge Graph



Outline

- Knowledge graph fundamentals
  - A brief history of knowledge graph
  - ***Knowledge graph representation***
- Knowledge graph construction

Knowledge Graph Representation

- Graph
  - Adjacency matrix
- Database
  - Table – Schema (SQL) vs. Schema-less (Non-SQL)
- Knowledge graph
  - (***S***ubject, ***P***redicate, ***O***bject) [“schema” + “data”]
    [Resource Description Framework (RDF)]

Knowledge Graph Representation

- Graph
  - Node
    - Attribute1
    - Attribute2
    - Attribute3
    - …
  - Edge

    (Node1, Node2, weights)

Homogeneous	vs	Heterogeneous
Node: 1 table Edge: 1 table		Node: multiple tables Edge: multiple tables

- (***S***ubject, ***P***redicate, ***O***bject)
  - Each ***node*** has an universal id (***S***)
  - It's ***attribute*** is represented as:

    (***S***, attributeName (***P***), attributeValue(***O***))
  - An ***edge*** connected two nodes (e.g. ***S1***, ***S2***)

    (***S1***, relationName (***P***), ***S2***(***O***))

Homogeneous	vs	Heterogeneous
Node + Edge: SINGLE table		

Knowledge Graph Representation : Example

Traditional database representation

Paper table

PaperId	Title	Year	VenueId
100001	Deep learning	2015	300001
100002	Mastering the game of Go without human knowledge	2017	300001

Journal table

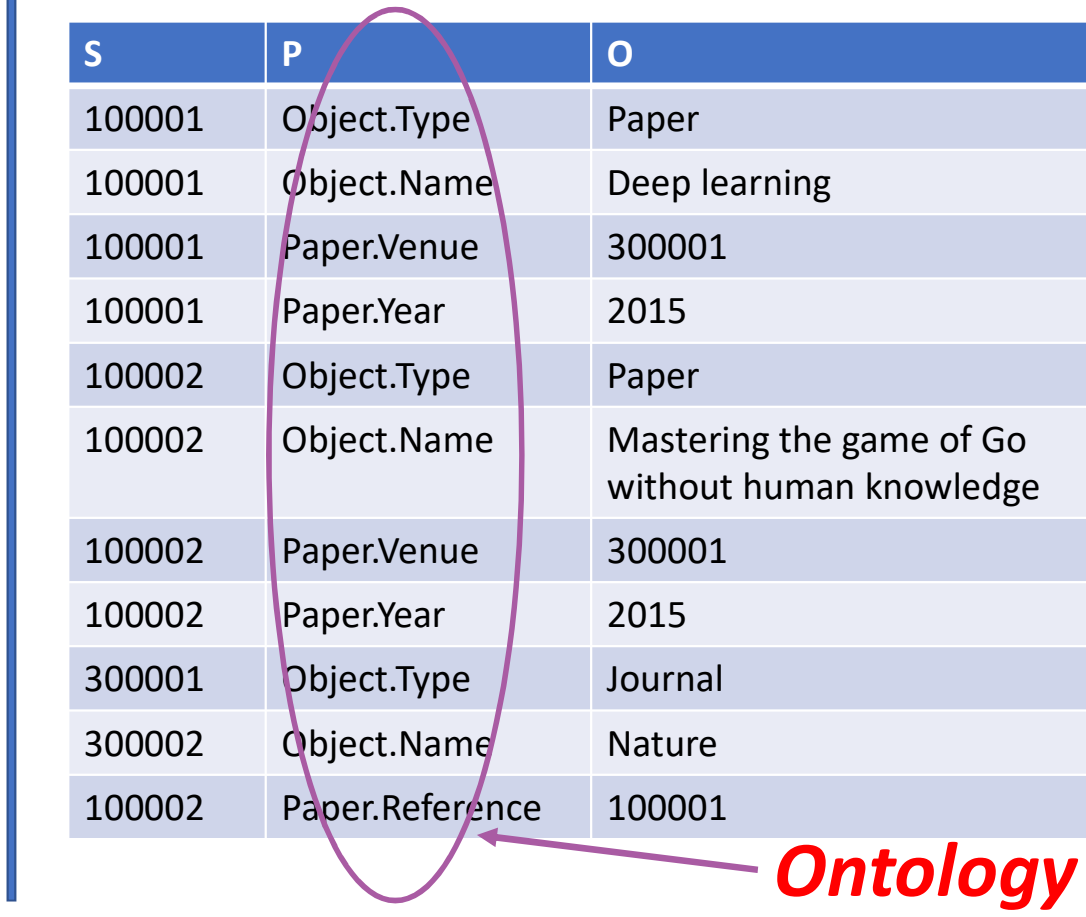
JournalId1	Name
300001	Nature

Citation table

PaperId1	PaperId2
100002	100001

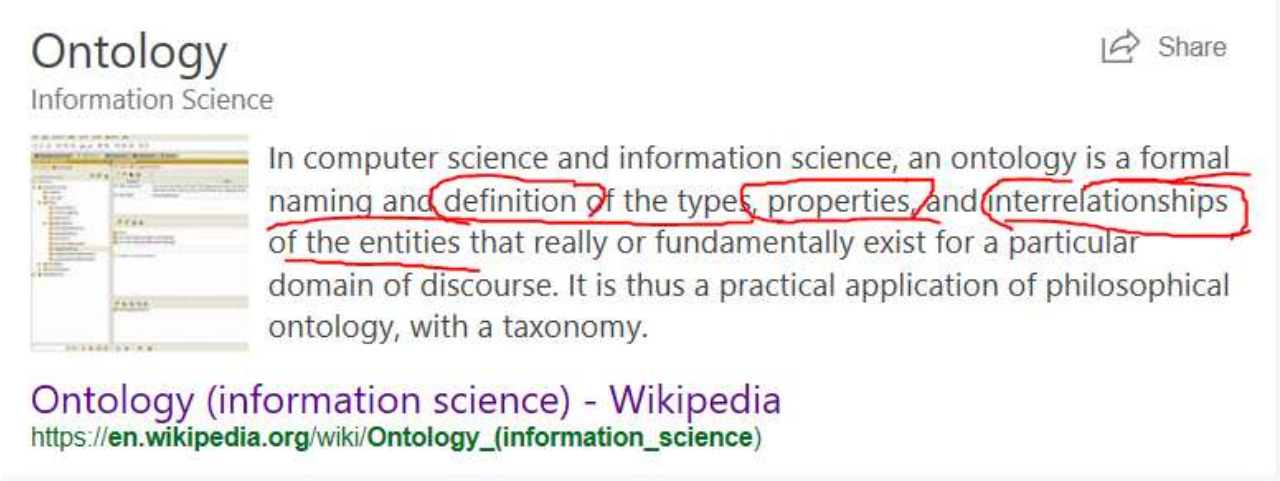
Knowledge graph (S, P, O) representation

S	P	O
100001	Object.Type	Paper
100001	Object.Name	Deep learning
100001	Paper.Venue	300001
100001	Paper.Year	2015
100002	Object.Type	Paper
100002	Object.Name	Mastering the game of Go without human knowledge
100002	Paper.Venue	300001
100002	Paper.Year	2015
300001	Object.Type	Journal
300002	Object.Name	Nature
100002	Paper.Reference	100001



Knowledge Graph Representation

- (***S***ubject, ***P***redicate, ***O***bject)
  - Pros:
    - Universal
    - Simple
    - Schema-less on the form
    (“Schema” defined in Ontology)
  - Cons:
    - Could be very complex for “simple / direct” relationship



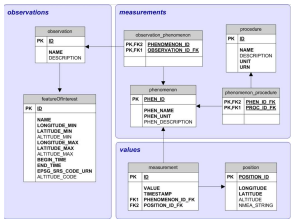
Outline

- Knowledge graph fundamentals
- Knowledge graph construction
  - ***KG construction overview***
    - NLP overview and basics
    - How to identify / recognize entities (nodes)
    - How to obtain relationships (edges)

Knowledge Graph Construction - Overview

Unstructured Documents

Natural language processing (NLP)

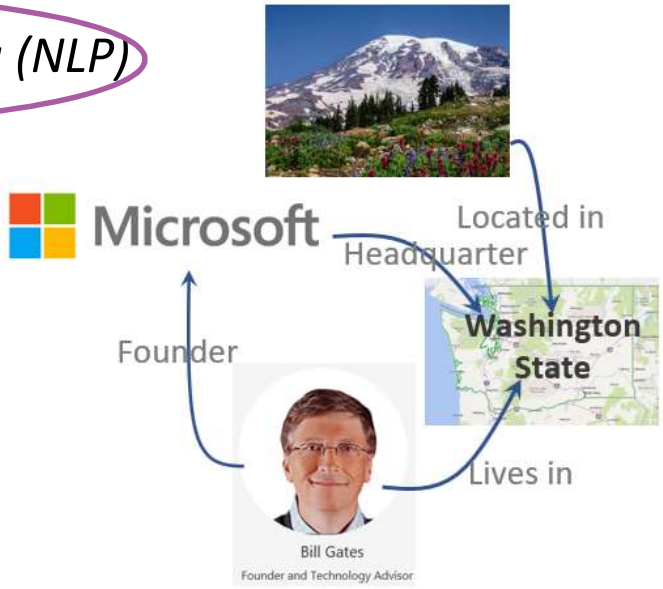


Existing Relational Databases

Data pipeline processing

Human Common Sense

Manual efforts

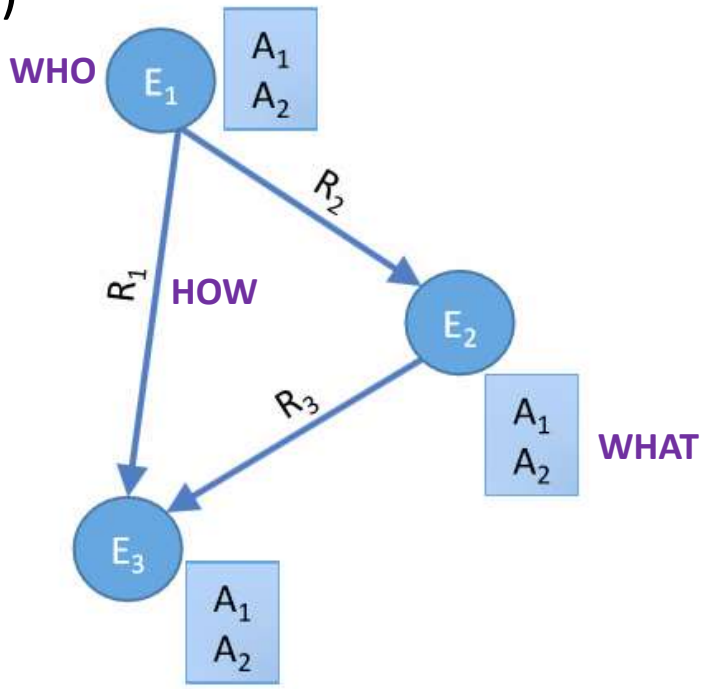


Knowledge in the ***graph*** form

Knowledge Graph Construction

- NLP Fundamentals

- What is Natural Language Processing (NLP)
- Named Entity Recognition (**WHO**)
- Entity Linking (**WHO / WHAT**)
- Relation Extraction (**HOW**)



Knowledge Graph Construction

Type: Mountain
Name: Mount Rainier
Elevation: 14,411 feed
First ascent: Aug 17, 1870
Last eruption: 1894

Microsoft

Located in

Headquarter

Founder

Lives in

WHO : entity / *node*
NER / Entity Linking

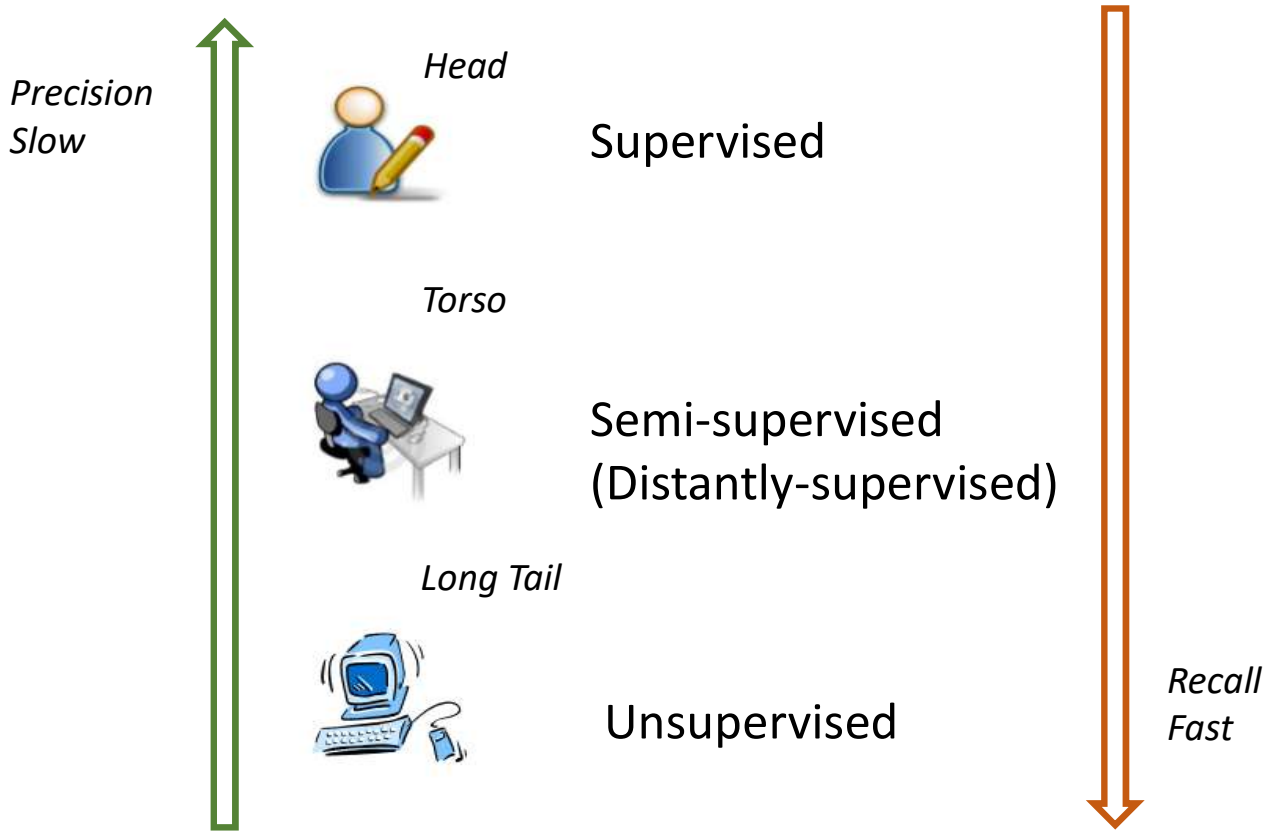
WHAT : attribute, type / *label*
Entity Linking

HOW : relation / *edge*
Relation Extraction

Knowledge in the *Graph* Form

Knowledge Graph Construction

- Challenges:
  - Incomplete
  - Inconsistent
  - Ambiguous

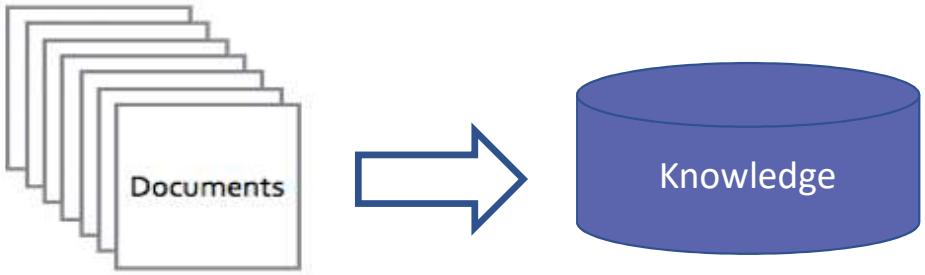


Outline

- Knowledge graph fundamentals
- Knowledge graph construction
  - KG construction overview
  - ***NLP overview and basics***
  - How to identify / recognize entities (nodes)
  - How to obtain relationships (edges)

What is Natural Language Processing (NLP)

- An area concerned with the understanding of natural language



- Unstructured
- Ambiguous / noisy
- Huge volume
- Growing

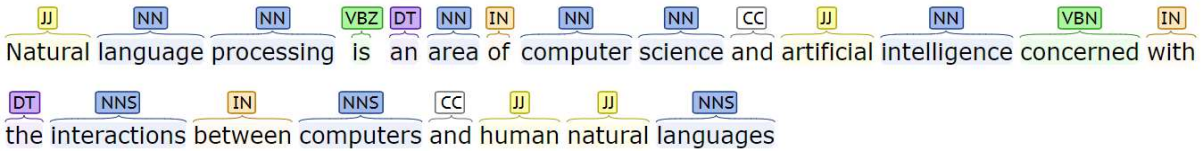
- Structured
- Precise / clean
- Indexable
- Actionable

What is Natural Language Processing (NLP)

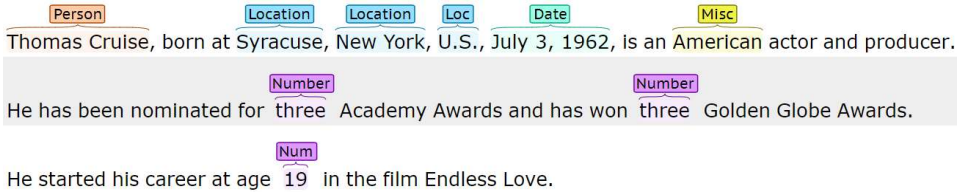
- Sentence level
- Document level (across sentences)
- Information extraction

Typical NLP problems - Sentence level

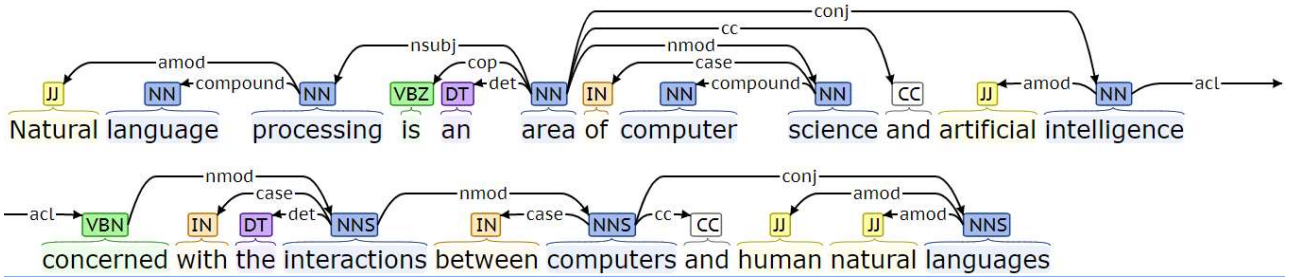
Part-of-speech tagging (POS tagging)



Named entity recognition

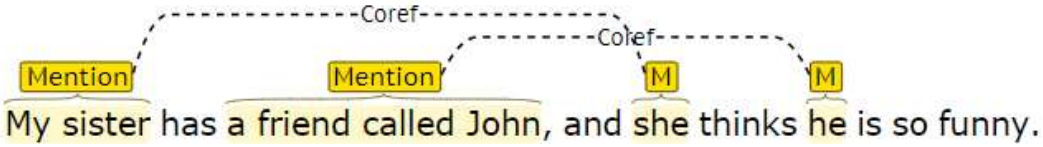


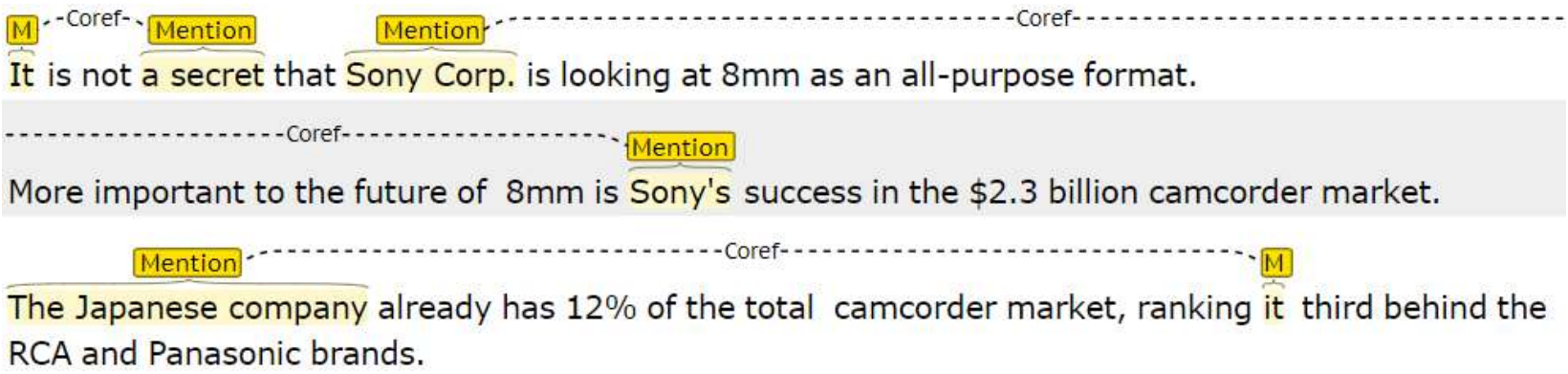
Dependency Parsing



Typical NLP Problems – Document level

- Coreference resolution





Typical NLP Problems – Information Extraction

- Entity Resolution
- Entity Linking
- Relation Extraction

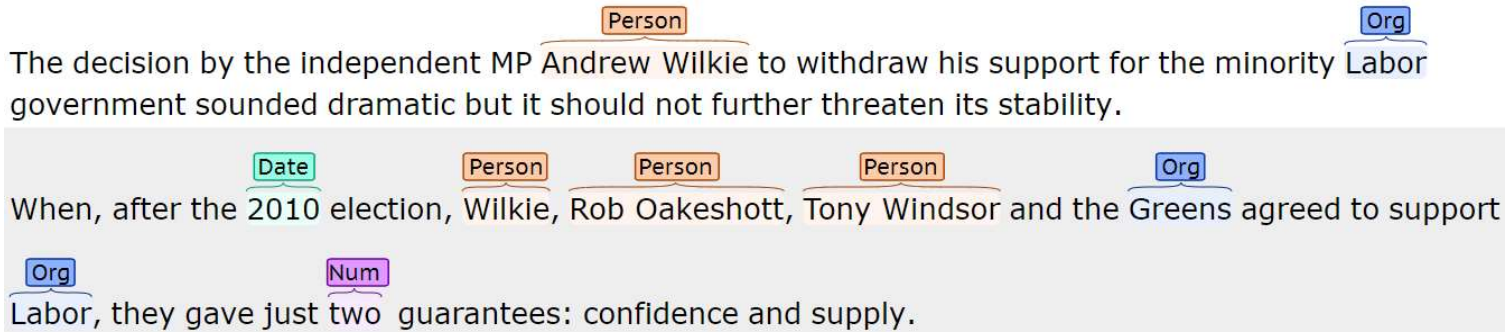
Outline

- Knowledge graph fundamentals

- Knowledge graph construction
  - KG construction overview
  - NLP overview and basics
  - How to identify / recognize entities (nodes)
    - ***Named Entity Recognition***
      - Entity Linking
  - How to obtain relationships (edges)

Named Entity Recognition (NER)

- Identify entity mentions in text, and then classify them into predefined set of types of interest



Named Entity Recognition (NER)

- Super important for digital assistant!

Named Entity Recognition (NER)

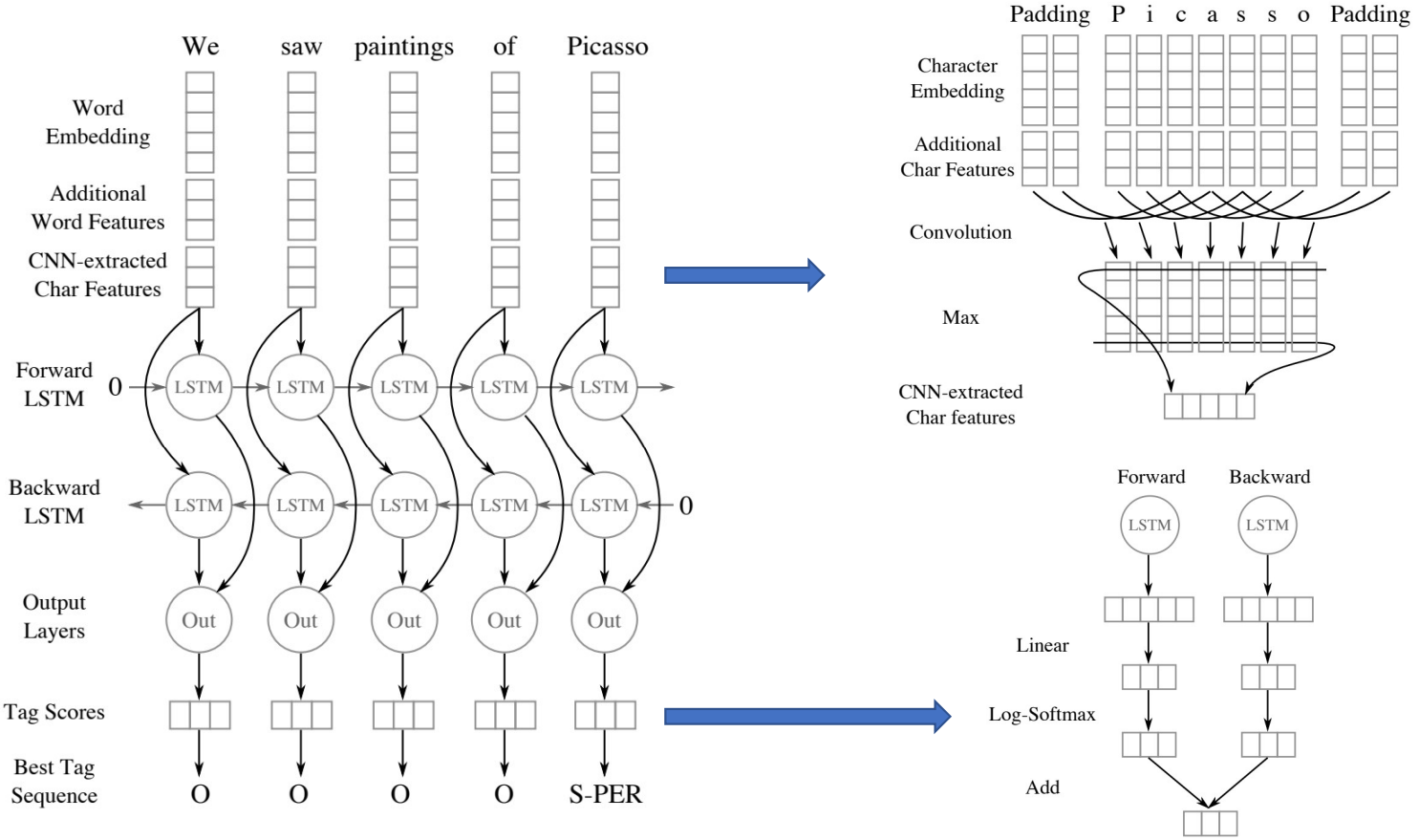
- Problem settings

	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
‘s	O	O
new	O	O
painting	O	O

Named Entity Recognition (NER)

- Traditional approach
  - Extract features
    - Words Feature: Current Word (essentially like a learned dictionary), Previous/next word (context)
    - Other kind of inferred linguistic feature: Part-of-speech tags
    - Label context: Previous (and perhaps next) label
    - Word Shapes: map “mRNA” to “xXXX”, map “CPA1” to “XXXd”, etc.
  - Algorithms
    - Naïve Bayes (NB)
    - Hidden Markov Model (HMM)
    - Conditional Random Field (CRF)

NER - A deep learning approach



Chiu et al., TACL'16. Named entity recognition with bidirectional LSTM-CNNs.

Outline

- Knowledge graph fundamentals

- Knowledge graph construction
  - KG construction overview
  - NLP overview and basics
  - How to identify / recognize entities (nodes)
    - Named Entity Recognition

    - ***Entity Linking***

  - How to obtain relationships (edges)

Why is Entity Linking important?

- Enable Semantic Search experience
- Used for Knowledge Graph population
- Used as feature for improving:
  - Classification
  - Retrieval
  - Question and answering
  - Semantic similarity

Entity Linking – Problem Definition

- Linking free text to entities
  - Any piece of text
    - News document
    - Blog posts
    - Tweets
    - Queries
- Entities taken from a knowledge base
  - Freebase
  - Wikipedia

Entity Linking – Common Steps

- Determine “linkable” phrases
  - Mention detection
- Select candidate entity links
  - Link generation
  - May include NILs
    (null values, i.e., no target in KB)
- Use “context” to disambiguate/filter/improve

Entity Linking – An Example

Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (**DFS**) is an algorithm for traversing or searching a tree tree structure or graph. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before backtracking.

Formally, DFS is an uninformed search that progresses by expanding the first child node of the search tree that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search backtracks, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly expanded nodes are added to a LIFO stack for exploration.

sense	commonness	relatedness
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
Tree (data structure)	**2.57%**	**63.26%**
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.0%
Binary tree	0.04%	62.43%
Family tree	0.04%	16.31%
...		

Milne, et al., CIKM'08, Learning to Link with Wikipedia

Entity Linking – An Example

- Commonness

$$\frac{|L_{w,c}|}{\sum_{c'} |L_{w,c'}|}$$

Number of links
with target c' and anchor text w

Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (**DFS**) is an algorithm for traversing or searching a tree tree structure or graph. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before backtracking.

Formally, DFS is an uninformed search that progresses by expanding the first child node of the search tree that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search backtracks, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly expanded nodes are added to a LIFO stack for exploration.

sense	commonness	relatedness
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
Tree (data structure)	**2.57%**	**63.26%**
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.0%
Binary tree	0.04%	62.43%
Family tree	0.04%	16.31%
...		

Milne, et al., CIKM'08, Learning to Link with Wikipedia

Entity Linking – An Example

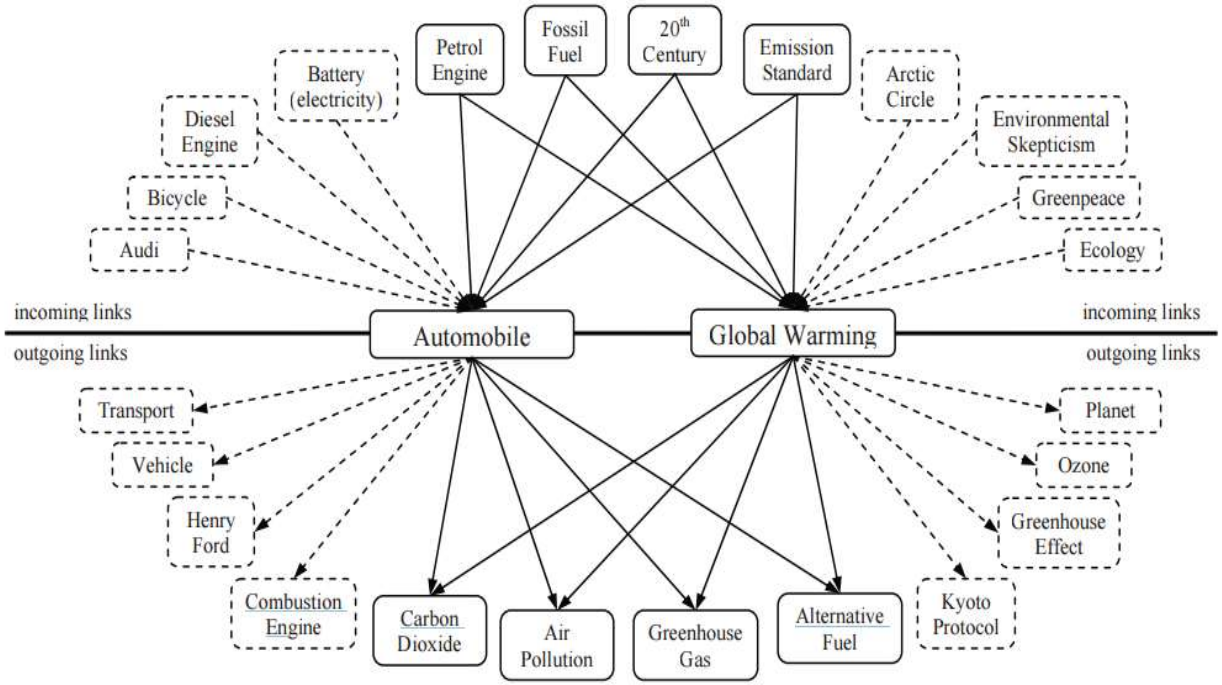
- Relatedness

Number of links
with target c

Intersection of inlinks
with target c and c'

$$\frac{\log(\max(|L_c|, |L_{c'}|)) - \log(|L_c \cap L_{c'}|)}{\log(|WP|) - \log(\min(|L_c|, |L_{c'}|))}$$

Total number of Wikipedia articles



Milne, et al., AAAI'08. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links

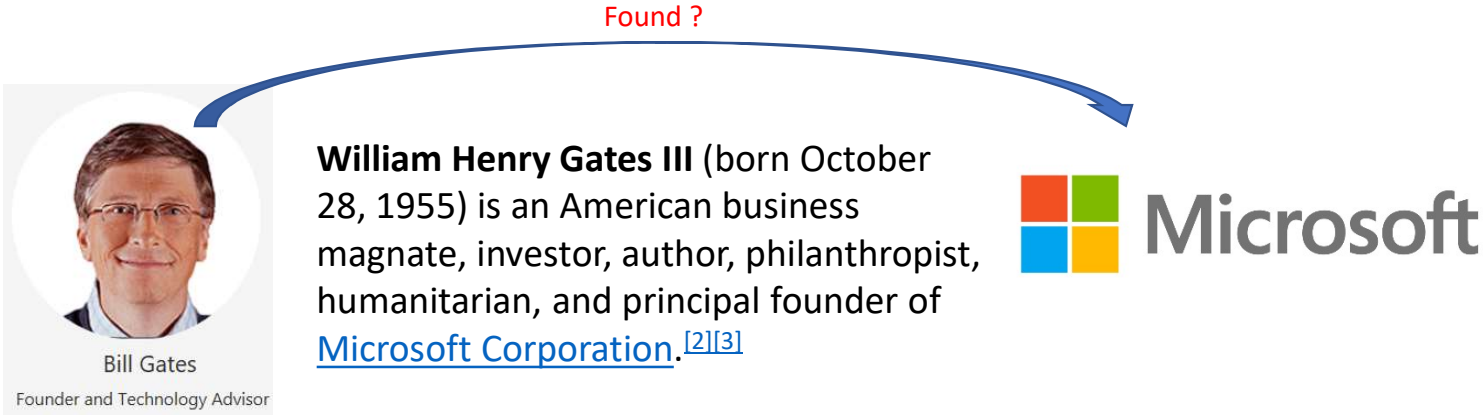
Outline

- Knowledge graph fundamentals

- Knowledge graph construction
  - KG construction overview
  - NLP overview and basics
  - How to identify / recognize entities (nodes)
    - Named Entity Recognition
    - Entity Linking
  - How to obtain relationships (edges)

  - ***Relation extraction***

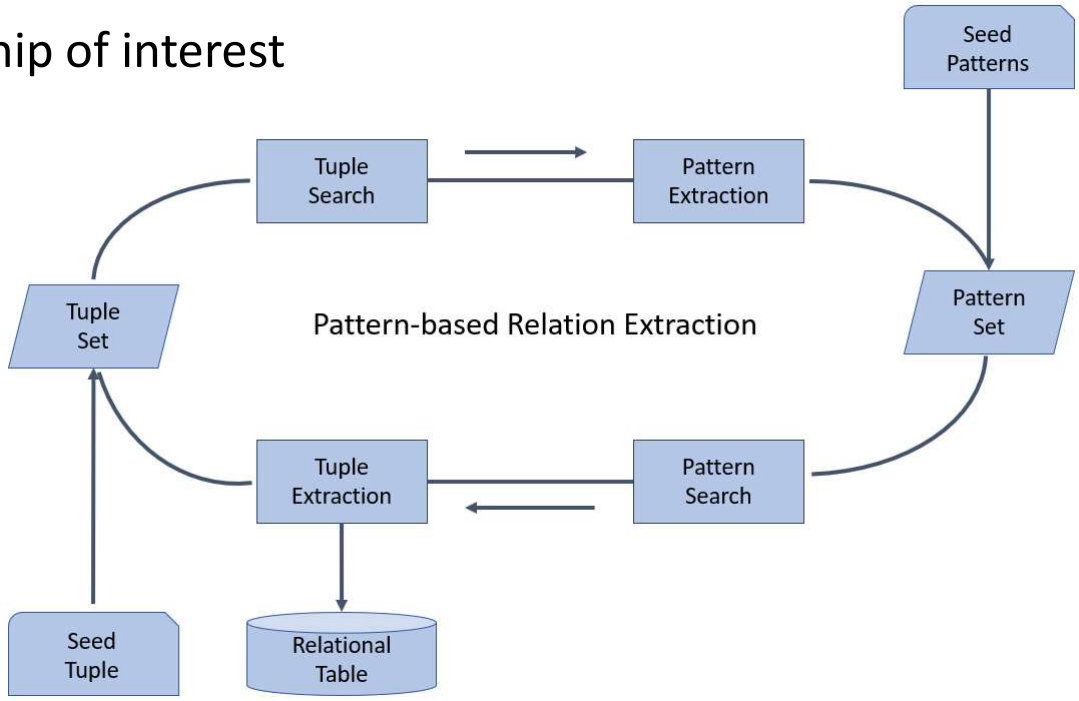
Relation Extraction



- Extract semantic relationships between entities
  - Undefined vs pre-defined set of relations
  - Binary vs multiple relations
  - Supervised vs unsupervised vs distant-supervision

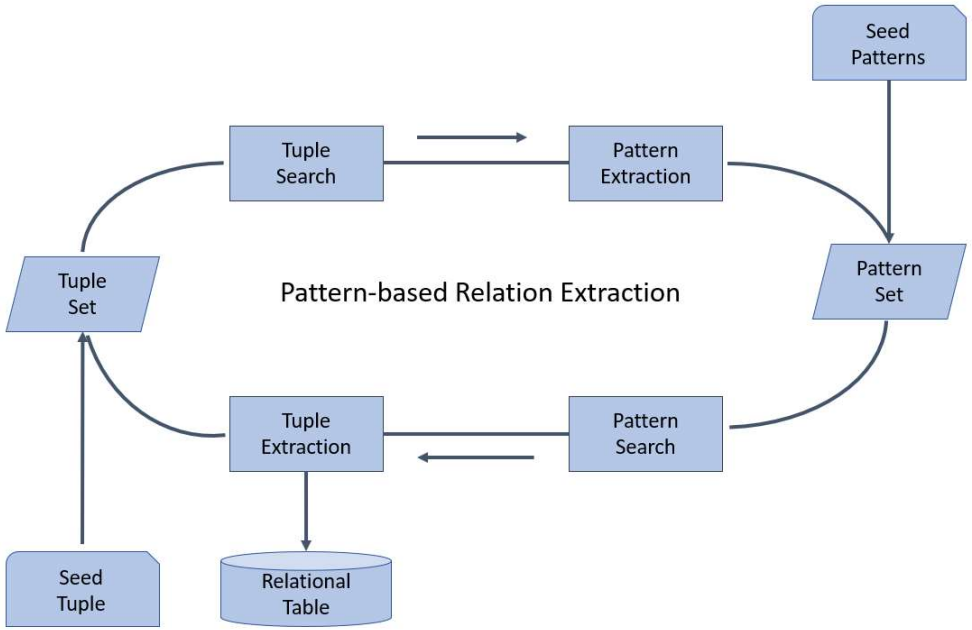
Bootstrapping Method

- Requires:
  - Seed instances of the relationship of interest
  - Unannotated text or document
- A semi-supervised approach



Reproducing the flowchart from: https://web.stanford.edu/class/cs224u/materials/cs224u-2016-relation-extraction.pdf

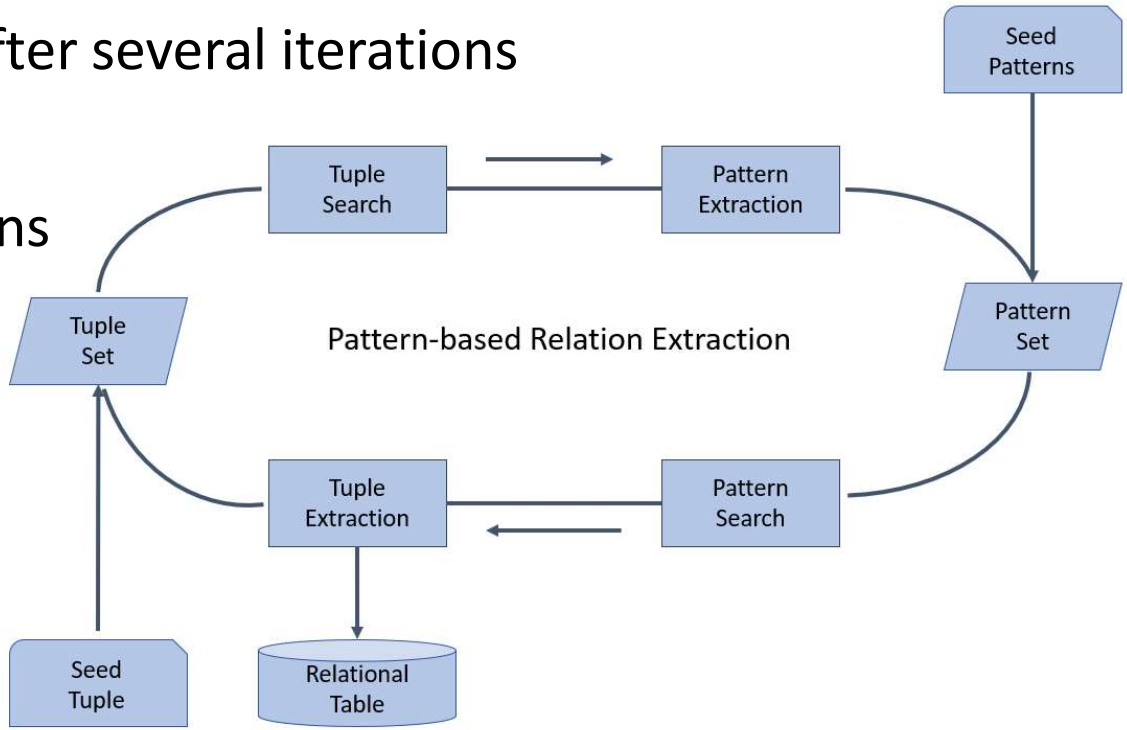
Bootstrapping Method



- Targeted relations: Acquisition
- Seed relation: (Microsoft, LinkedIn)
- Search "Microsoft" and "LinkedIn" using search engines:
  - "Microsoft has acquired LinkedIn" ➔ X has acquired Y
  - "Microsoft purchase of LinkedIn" ➔ X's purchase of Y
  - "Microsoft buys LinkedIn" ➔ X buys Y
- Use these new patterns to find new tuples

Bootstrapping Limitations

- Need to use seeds for each relation
- Semantics might be drifted after several iterations
- Hard to control precisions
- No probabilistic interpretations



Supervised Relation Extraction

- Utilizing labels of relation mentions
  - **Elon Musk** is the founder, CEO, and lead designer of **SpaceX** ➔ (ElonMusk, CEO, SpaceX)
  - **Elon Musk** has stated the goals of **SpaceX**, Tesla and SolarCity. ➔ NIL
- Traditional relation extraction datasets
  - ACE 2004
  - MUC-7
  - Biomedical datasets
- Learn classifiers from those positive and negative examples

Supervised Relation Extraction – at a glance

Typical features

- Bags of words & bigrams between, before, and after the entities
- POS tags
- The types of the entities
- Dependency path between entities
- Distance between entities
- Tree distance between the entities
- NER tags

Classifiers

(any classifiers that supports multiclass prediction)

- SVM
- Multiclass logistic regression
- Naïve Bayes

Pros *versus* *Cons*

- Higher accuracy
- Explicit negative examples

- Very expensive to label data
- Doesn't generalize well to different relations

Distantly Supervised Relation Extraction

- Basic assumption:
  - Existing knowledge base has rich information
  - Existing knowledge base + unlabeled text
    - ➔ generate training examples
      - Locate pairs of related entities in text
      - Hypothesizes that the relation is expressed

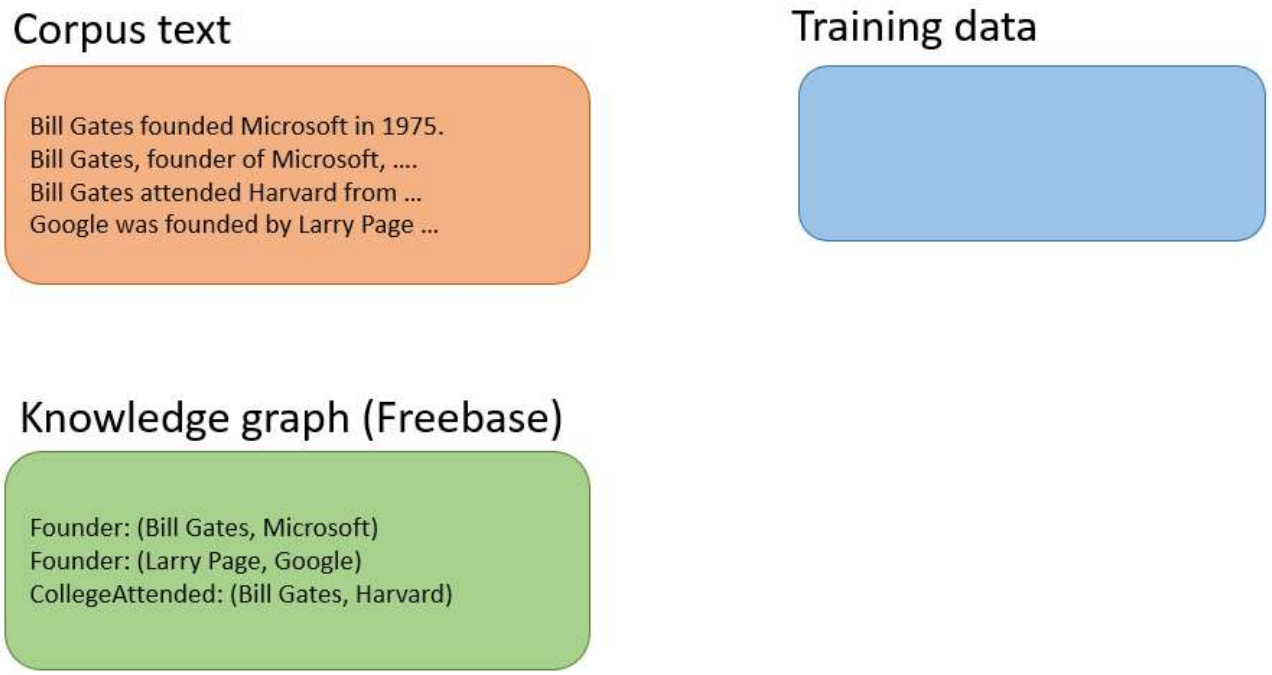
Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Table 2: The 23 largest Freebase relations we use, with their size and an instance of each relation.

Mintz, et al., ACL'09, Distant supervision for relation extraction without labeled data

Distantly Supervised Relation Extraction

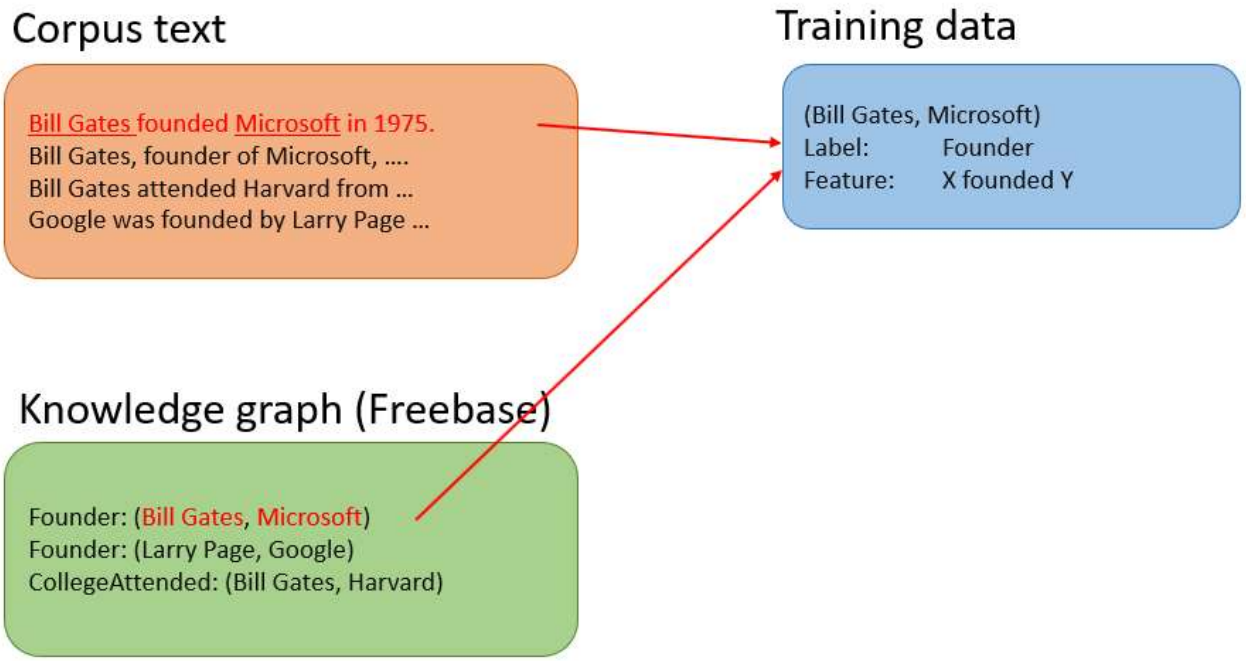
- Collection training data



Reproducing the example from: https://web.stanford.edu/class/cs224u/materials/cs224u-2016-relation-extraction.pdf

Distantly Supervised Relation Extraction

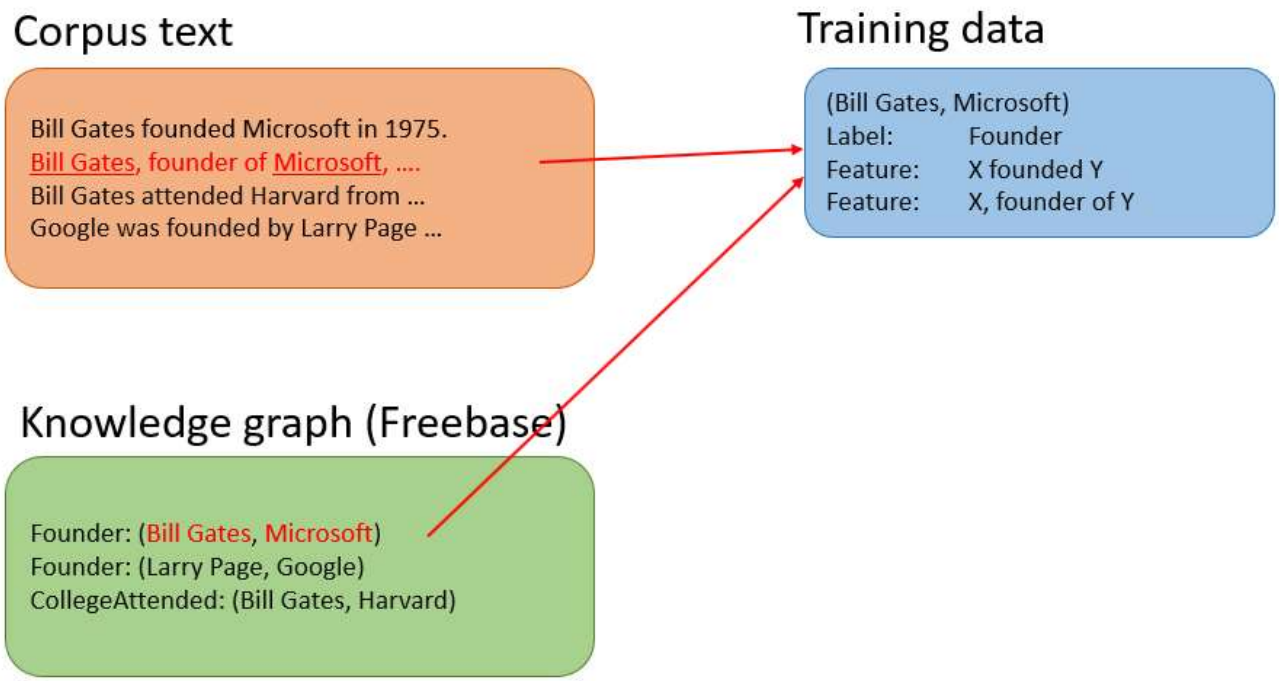
- Collection training data



Reproducing the example from: https://web.stanford.edu/class/cs224u/materials/cs224u-2016-relation-extraction.pdf

Distantly Supervised Relation Extraction

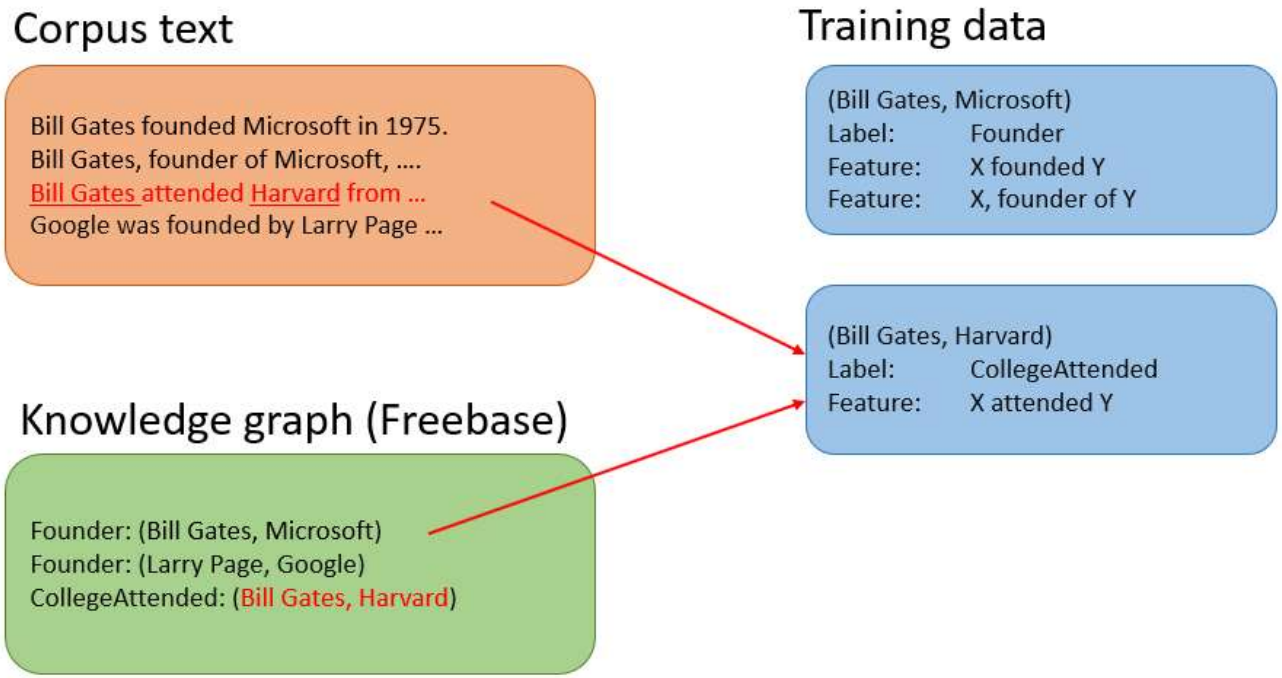
- Collection training data



Reproducing the example from: https://web.stanford.edu/class/cs224u/materials/cs224u-2016-relation-extraction.pdf

Distantly Supervised Relation Extraction

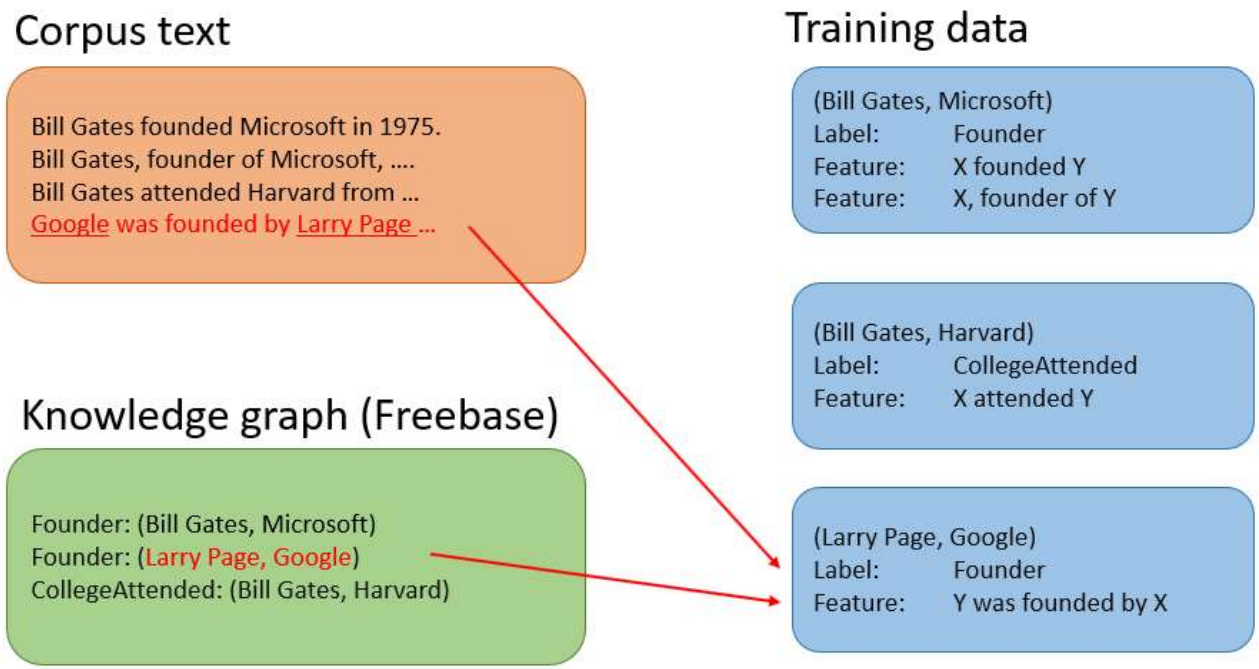
- Collection training data



Reproducing the example from: https://web.stanford.edu/class/cs224u/materials/cs224u-2016-relation-extraction.pdf

Distantly Supervised Relation Extraction

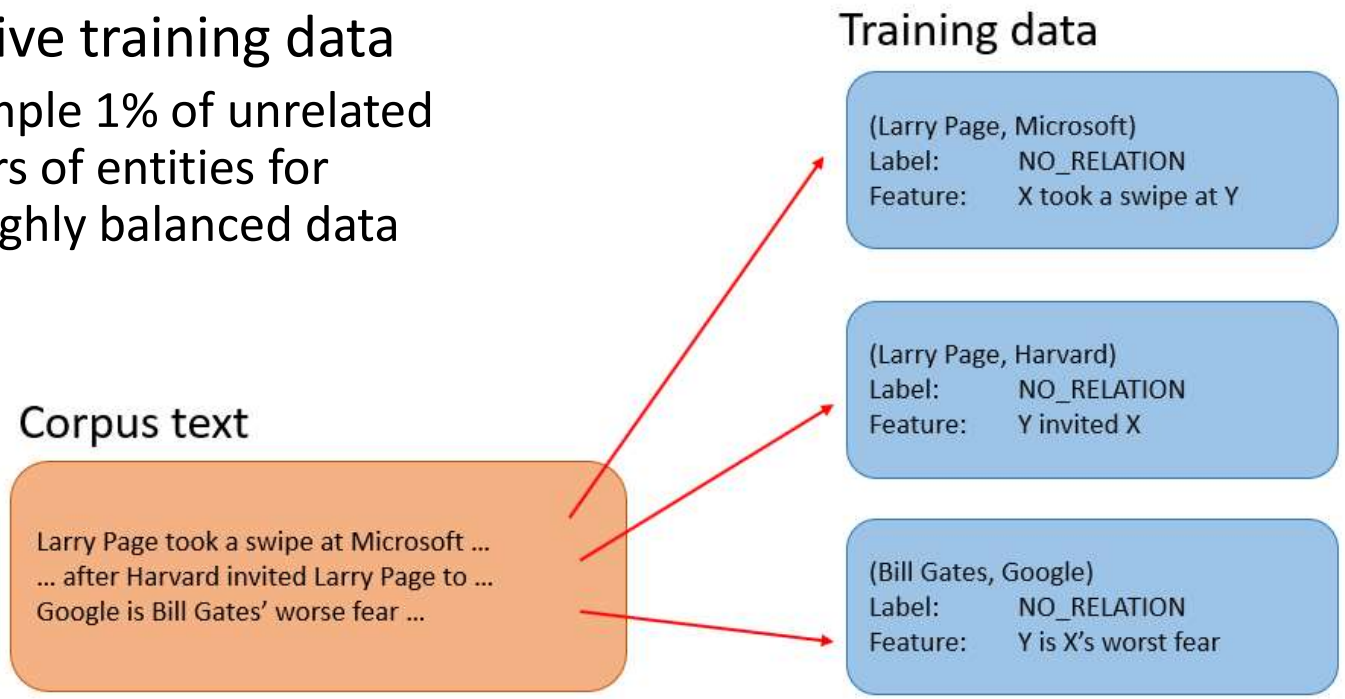
- Collection training data



Reproducing the example from: https://web.stanford.edu/class/cs224u/materials/cs224u-2016-relation-extraction.pdf

Distantly Supervised Relation Extraction

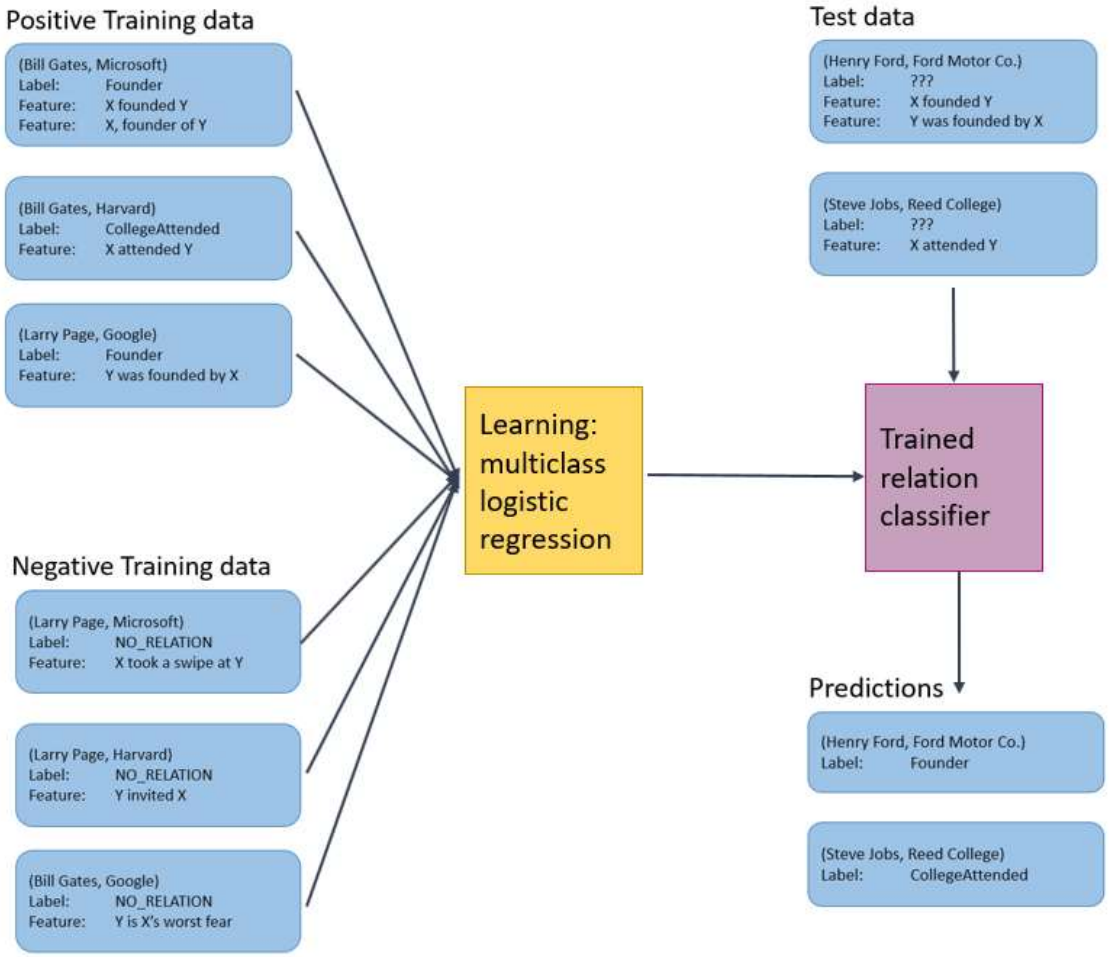
- Negative training data
  - Sample 1% of unrelated pairs of entities for roughly balanced data



Reproducing the example from: https://web.stanford.edu/class/cs224u/materials/cs224u-2016-relation-extraction.pdf

Distantly Supervised Relation Extraction

- Experiment



Reproducing the example from: https://web.stanford.edu/class/cs224u/materials/cs224u-2016-relation-extraction.pdf

Distantly Supervised Relation Extraction

Pros

- Can scale since no supervision required
- Leverage rich and reliable data from knowledge base
- Leverage unlimited amounts of text data
- Can generalize to different domains

versus

Cons

- Needs high quality entity matching

Summary

- Knowledge Graph Fundamental
  - A brief history
    - Existential graph -> Semantic network -> Linked data -> Knowledge graph
    - Expert system -> Knowledge base
  - Representation (***S***, ***P***, ***O***)

- Knowledge Graph Construction
  - NLP fundamental (why need NLP here?)
    - To understand text, extract information and build knowledge graph
  - Knowledge graph construction
    - Challenges
    - Key steps -> can be phrased into diff. **NLP** problems

Summary

- Knowledge graph construction -> Key **NLP** problems
  - Name Entity Recognition (NER)
  - Entity Linking
  - Relation Extraction