Relation Extraction

Prof. Sameer Singh

CS 295: STATISTICAL NLP

WINTER 2017

With some slides from Luke Zettlemoyer

February 23, 2017

Outline

- Introduction to Relation Extraction
- Hand-written Patterns
- Supervised Machine Learning
- Semi and Unsupervised Learning

Outline

- Introduction to Relation Extraction
- Hand-written Patterns
- Supervised Machine Learning
- Semi and Unsupervised Learning

Goal: "machine reading"

- Acquire structured knowledge from unstructured text

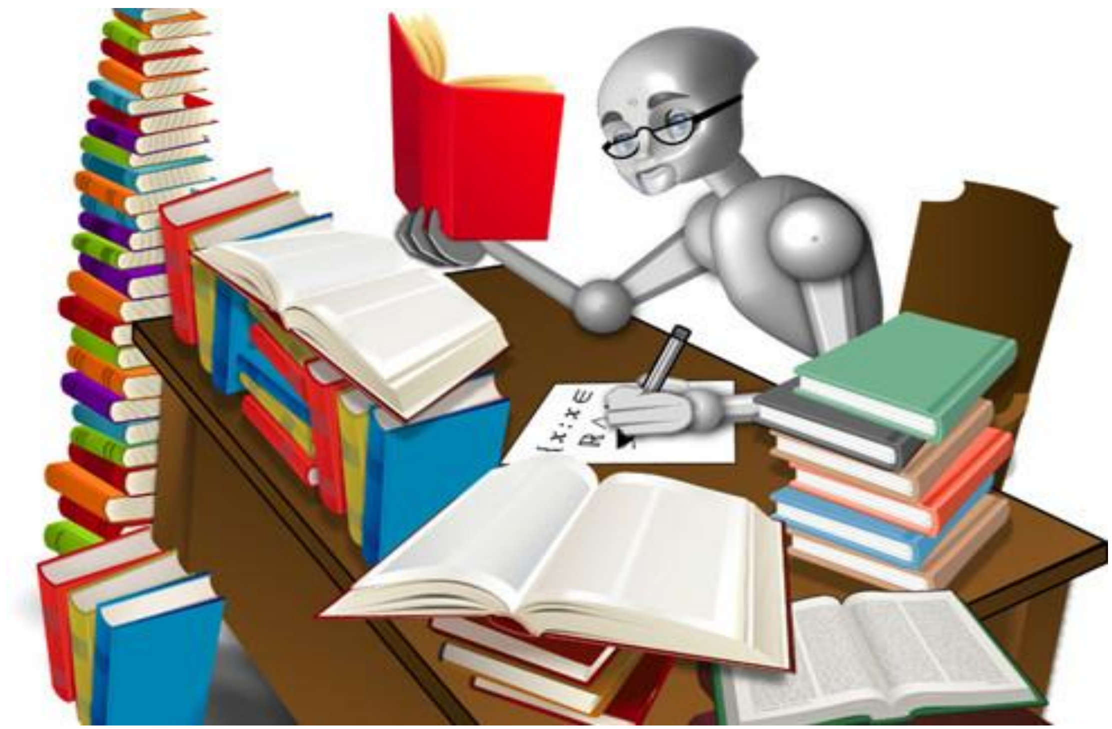


illustration from DARPA

Information extraction

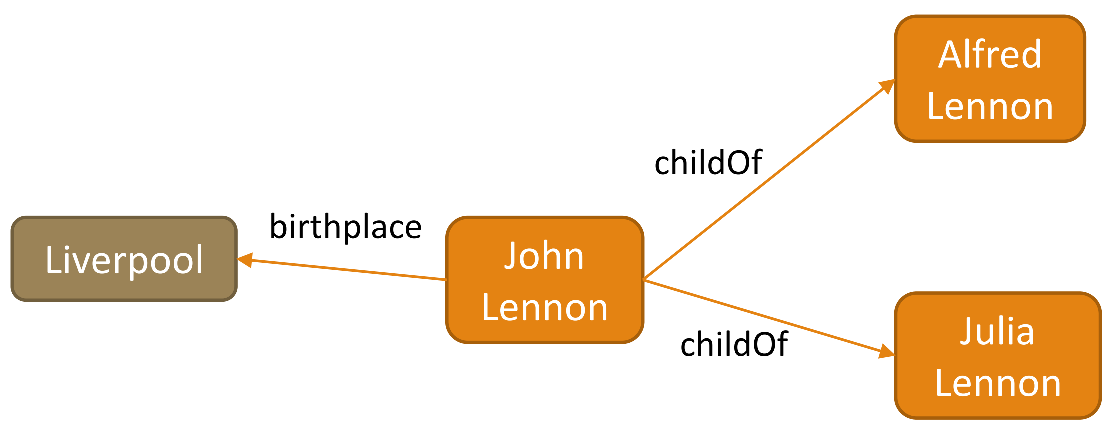
- IE = extracting information from text
- Sometimes called *text analytics* commercially
- Extract entities
  - People, organizations, locations, times, dates, prices, ...
  - Or sometimes: genes, proteins, diseases, medicines, ...
- Extract the relations between entities
  - Located in, employed by, part of, married to, ...
- Figure out the larger events that are taking place

Knowledge Extraction

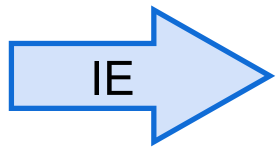
John was born in Liverpool, to Julia and Alfred Lennon.

Text

Literal Facts



Machine-readable summaries



Subject	Relation	Object
p53	**is_a**	protein
Bax	**is_a**	protein
p53	has_function	apoptosis
Bax	has_function	induction
apoptosis	involved_in	cell_death
Bax	is_in	mitochondrial outer membrane
Bax	is_in	cytoplasm
apoptosis	related_to	caspase activation
...	...	...

textual abstract:
summary for human

structured knowledge extraction:
summary for machine

More applications of IE

- Building & extending knowledge bases and ontologies
- Scholarly literature databases: Google Scholar, CiteSeerX
- People directories: Rapleaf, Spoke, Naymz
- Shopping engines & product search
- Bioinformatics: clinical outcomes, gene interactions, …
- Patent analysis
- Stock analysis: deals, acquisitions, earnings, hirings & firings
- SEC filings
- Intelligence analysis for business & government

Read **Biology Papers** → Extract **Genes binding to Proteins**

Biology

Read **Court Opinions** → Extract **Judge, case-type, holding, decision**

Law

Read **Web health Forums** → Extract **Drugs and reported side effects**

Health

Read **News** → Extract **country1, country2, diplomatic event, time**

Political Science

Read **CVEs** → Extract **vulnerable products, versions, exploit type**

Security

Read **Nanorobotic papers** → Extract **nano-rod size, release mechanism, release time**

Nanotech + Robotics

Relation Extraction

Company report: "International Business Machines Corporation (IBM or the company) was incorporated in the State of New York on June 16, 1911, as the Computing-Tabulating-Recording Co. (C-T-R)…"

Extracted Complex Relation:

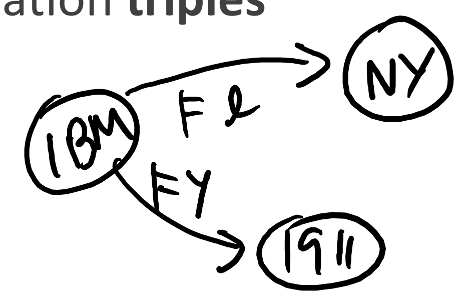
Company-Founding

Company	IBM
Location	New York
Date	June 16, 1911
Original-Name	Computing-Tabulating-Recording Co.

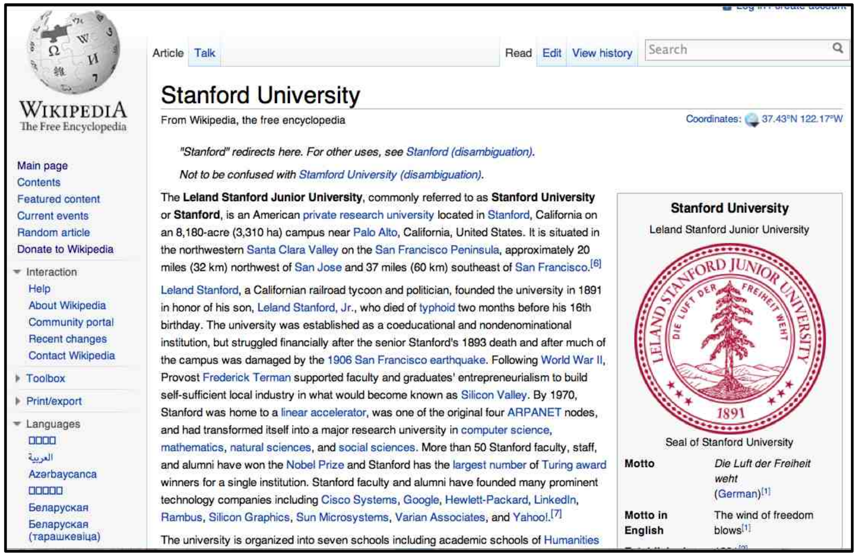
But we will focus on the simpler task of extracting relation **triples**

Founding-year(IBM,1911)

Founding-location(IBM,New York)



Extracting Relation Triples



Stanford University

From Wikipedia, the free encyclopedia

"Stanford" redirects here. For other uses, see Stanford (disambiguation).

Not to be confused with Stamford University (disambiguation).

The **Leland Stanford Junior University**, commonly referred to as **Stanford University** or **Stanford**, is an American private research university located in Stanford, California on an 8,180-acre (3,310 ha) campus near Palo Alto, California, United States. It is situated in the northwestern Santa Clara Valley on the San Francisco Peninsula, approximately 20 miles (32 km) northwest of San Jose and 37 miles (60 km) southeast of San Francisco.[6]

Leland Stanford, a Californian railroad tycoon and politician, founded the university in 1891 in honor of his son, Leland Stanford, Jr., who died of typhoid two months before his 16th birthday. The university was established as a coeducational and nondenominational institution, but struggled financially after the senior Stanford's 1893 death and after much of the campus was damaged by the 1906 San Francisco earthquake. Following World War II, Provost Frederick Terman supported faculty and graduates' entrepreneurialism to build self-sufficient local industry in what would become known as Silicon Valley. By 1970, Stanford was home to a linear accelerator, was one of the original four ARPANET nodes, and had transformed itself into a major research university in computer science, mathematics, natural sciences, and social sciences. More than 50 Stanford faculty, staff, and alumni have won the Nobel Prize and Stanford has the largest number of Turing award winners for a single institution. Stanford faculty and alumni have founded many prominent technology companies including Cisco Systems, Google, Hewlett-Packard, LinkedIn, Rambus, Silicon Graphics, Sun Microsystems, Varian Associates, and Yahoo!.[7]

The university is organized into seven schools including academic schools of Humanities

The Leland Stanford Junior University, commonly referred to as Stanford University or Stanford, is an American private research university located in Stanford, California … near Palo Alto, California… Leland Stanford…founded the university in 1891

Stanford EQ Leland Stanford Junior University
Stanford LOC-IN California
Stanford IS-A research university
Stanford LOC-NEAR Palo Alto
Stanford FOUNDED-IN 1891
Stanford FOUNDER Leland Stanford

News Domain

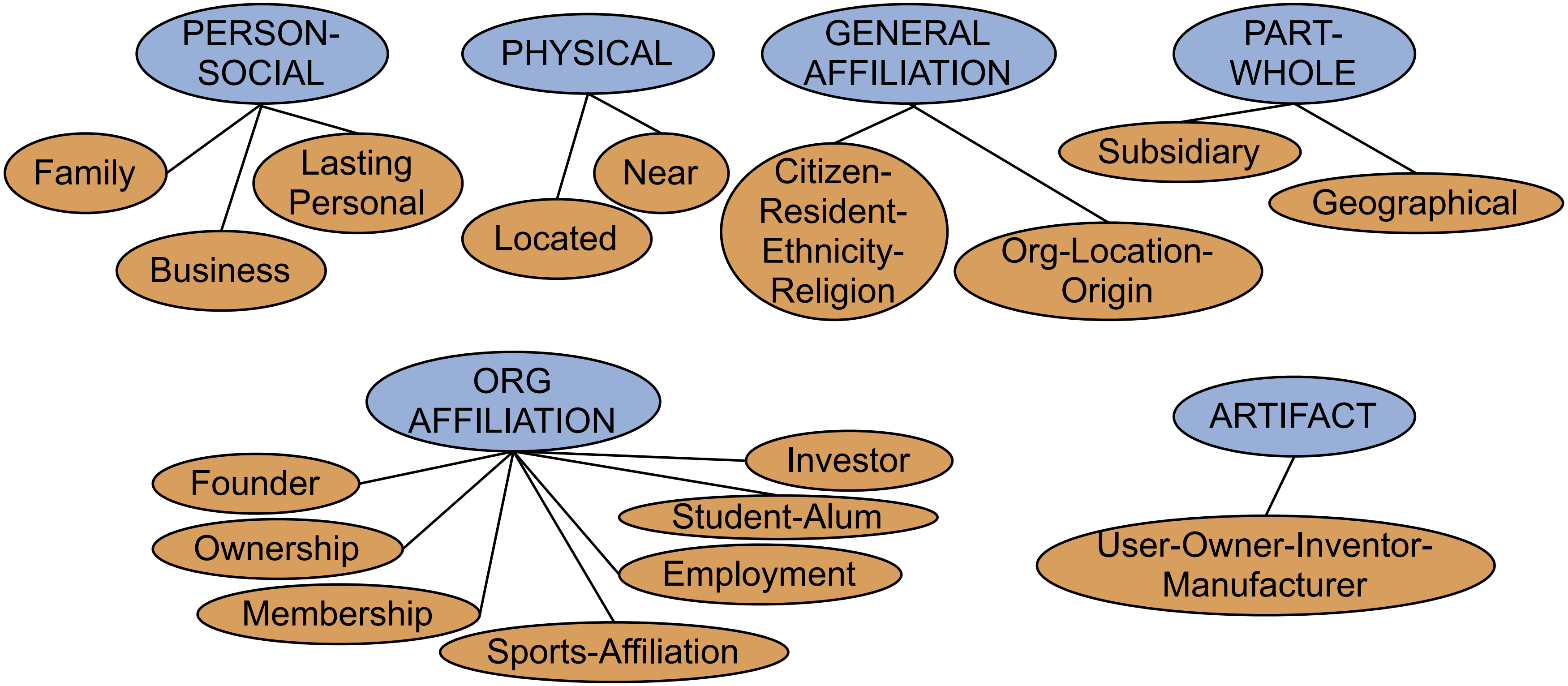
ROLE: relates a person to an organization or a geopolitical entity
- subtypes: member, owner, affiliate, client, citizen

PART: generalized containment
- subtypes: subsidiary, physical part-of, set membership

AT: permanent and transient locations
- subtypes: located, based-in, residence

SOCIAL: social relations among persons
- subtypes: parent, sibling, spouse, grandparent, associate

Automated Content Extraction



ACE Relations Examples

Physical-Located PER-GPE

`He was in Tennessee`

Part-Whole-Subsidiary ORG-ORG

`XYZ, the parent company of ABC`

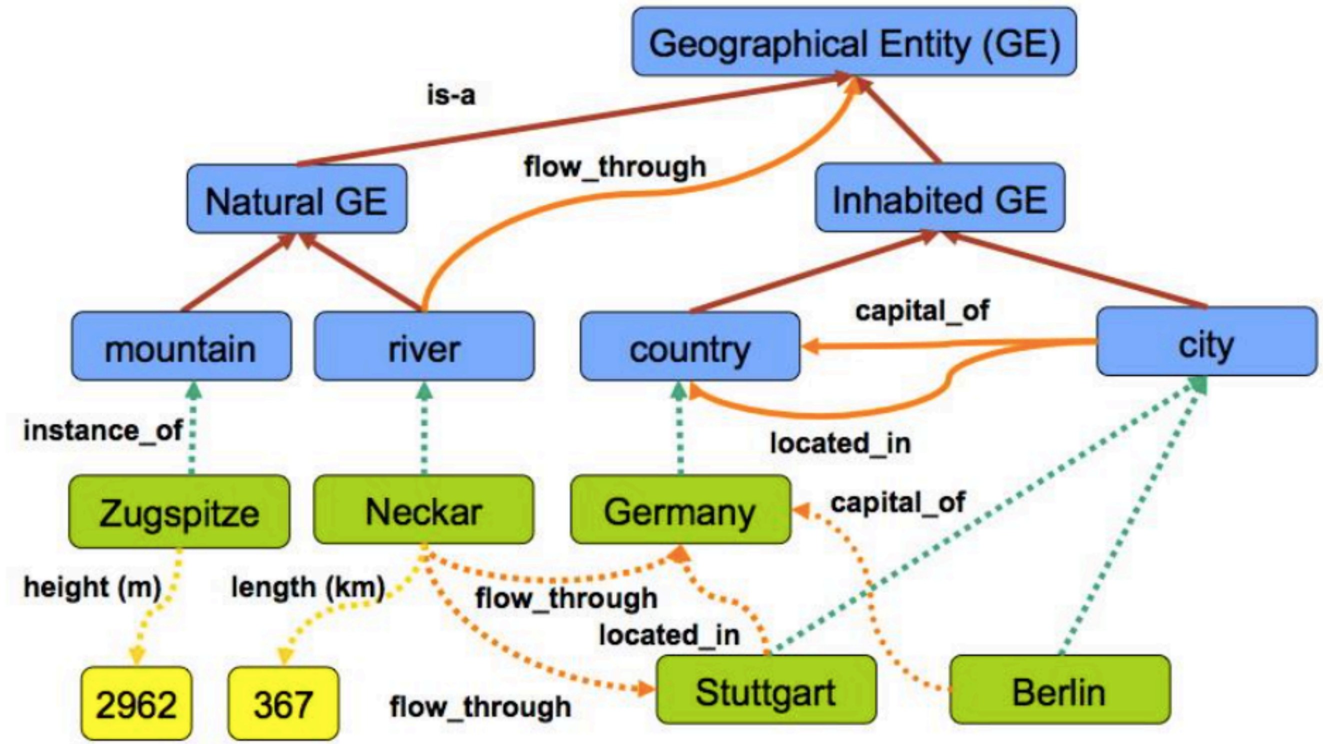
Person-Social-Family PER-PER

`John's wife Yoko`

Org-AFF-Founder PER-ORG

`Steve Jobs, co-founder of Apple…`

Geographical Relations



Medical Relations

UMLS Resource

Injury	disrupts	Physiological Function
Bodily Location	location-of	Biologic Function
Anatomical Structure	part-of	Organism
Pharmacologic Substance	causes	Pathological Function
Pharmacologic Substance	treats	Pathologic Function

Medical Relations

Doppler echocardiography can be used to diagnose left anterior descending artery stenosis in patients with type 2 diabetes

↓

Echocardiography, Doppler DIAGNOSES Acquired stenosis

Freebase Relations

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

Thousands of relations and millions of instances!

Manually created from multiple sources including Wikipedia InfoBoxes

Ontological Relations WordNet

IS-A (hypernym): subsumption between classes

- Giraffe IS-A ruminant IS-A ungulate IS-A mammal IS-A vertebrate IS-A animal…

Instance-of: relation between individual and class

- San Francisco instance-of city

Outline

Introduction to Relation Extraction

Hand-written Patterns

Supervised Machine Learning

Semi and Unsupervised Learning

Rules for IS-A Relation

Early intuition from Hearst (1992)

"Agar is a substance prepared from
a mixture of red algae, such as Gelidium,
for laboratory or industrial use"

What does Gelidium mean?

How do you know?

Hearst's Patterns for IS-A relations

Is-A

"Y such as X ((, X)* (, and|or) X)"

"such Y as X"

"X or other Y"

"X and other Y"

"Y including X"

"Y, especially X"

Hearst's Patterns for IS-A relations

Is-A

Is-A

Hearst pattern	Example occurrences
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	Bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
Such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y , especially X	European countries, especially France, England, and Spain...

Patterns for learning meronyms

- Berland & Charniak (1999) tried it
- Selected initial patterns by finding all sentences in a corpus containing *basement* and *building*

whole NN[-PL] ’s POS part NN[-PL]	... building’s basement ...
part NN[-PL] of PREP {the|a} DET mods [JJ|NN]* whole NN	... basement of a building ...
part NN in PREP {the|a} DET mods [JJ|NN]* whole NN	... basement in a building ...
parts NN-PL of PREP wholes NN-PL	... basements of buildings ...
parts NN-PL in PREP wholes NN-PL	... basements in buildings ...

- Then, for each pattern:
  1. found occurrences of the pattern
  2. filtered those ending with *-ing*, *-ness*, *-ity*
  3. applied a likelihood metric — poorly explained
- Only the first two patterns gave decent (though not great!) results

Extracting Richer Relations

Intuition:

Relations often hold between specific types of entities

- located-in (ORGANIZATION, LOCATION)
- founded (PERSON, ORGANIZATION)
- cures (DRUG, DISEASE)

Start with Named Entity tags to extract relation!

Entity Types aren’t enough

Which relations hold between 2 entities?

Drug

Cure?

Prevent?

Cause?

Disease

Which relations hold between two entities?

PERSON

Founder?

Investor?

Member?

Employee?

President?

ORGANIZATION

Extracting Richer Relations Using Rules and Named Entities

Who holds what office in what organization?

PERSON, POSITION of ORG

- George Marshall, Secretary of State of the United States

PERSON (named|appointed|chose|*etc.*) PERSON Prep? POSITION

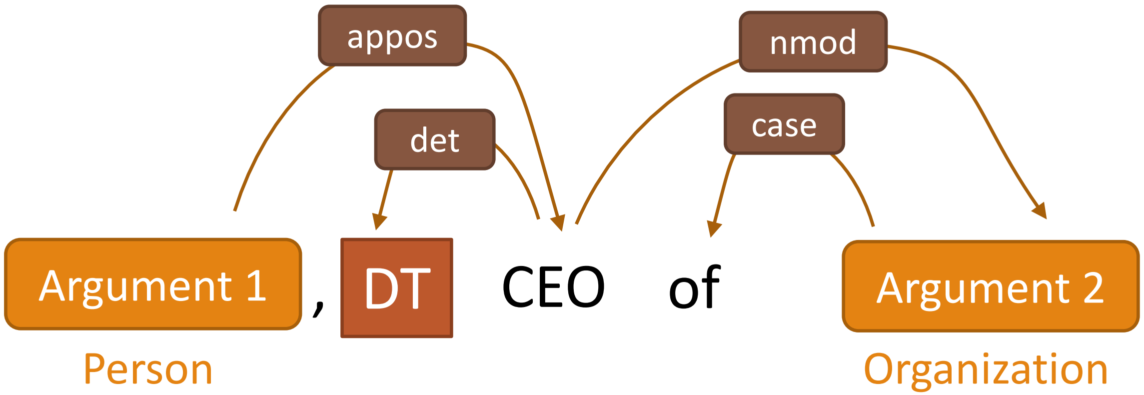
- Truman appointed Marshall Secretary of State

PERSON [be]? (named|appointed|*etc.*) Prep? ORG POSITION

- George Marshall was named US Secretary of State

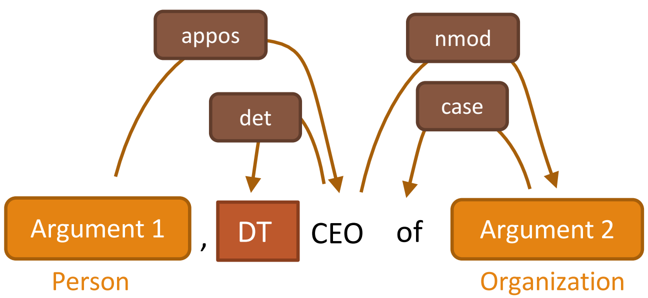
Complex Surface Patterns

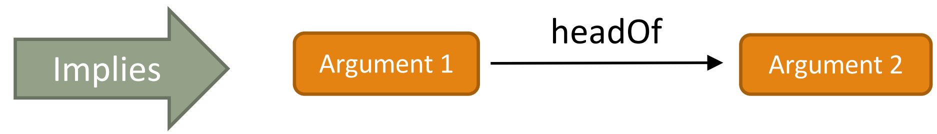
Combine tokens, dependency paths, and entity types to define rules.



Bill Gates, the CEO of Microsoft, said …
Mr. Jobs, the brilliant and charming CEO of Apple Inc., said …
… announced by Steve Jobs, the CEO of Apple.
… announced by Bill Gates, the director and CEO of Microsoft.
… mused Bill, a former CEO of Microsoft.
and many other possible instantiations…

Rule-Based Extraction





Use a collection of rules as the system itself

Variations

Source:
- Manually specified
- Learned from Data

Multiple Rules:
- Attach priorities/precedence
- Attach probabilities (more later)

Hand-built patterns for relations

Pluses

- Human patterns tend to be high-precision
- Can be tailored to specific domains
- Easy to debug: why a prediction was made, how to fix?

Minuses

- Human patterns are often low-recall
- A lot of work to think of all possible patterns!
- Don’t want to have to do this for every relation!
- We’d like better accuracy (*generalization*)

Outline

Introduction to Relation Extraction

Hand-written Patterns

Supervised Machine Learning

Semi and Unsupervised Learning

Supervised Machine Learning

Choose a set of relations we'd like to extract

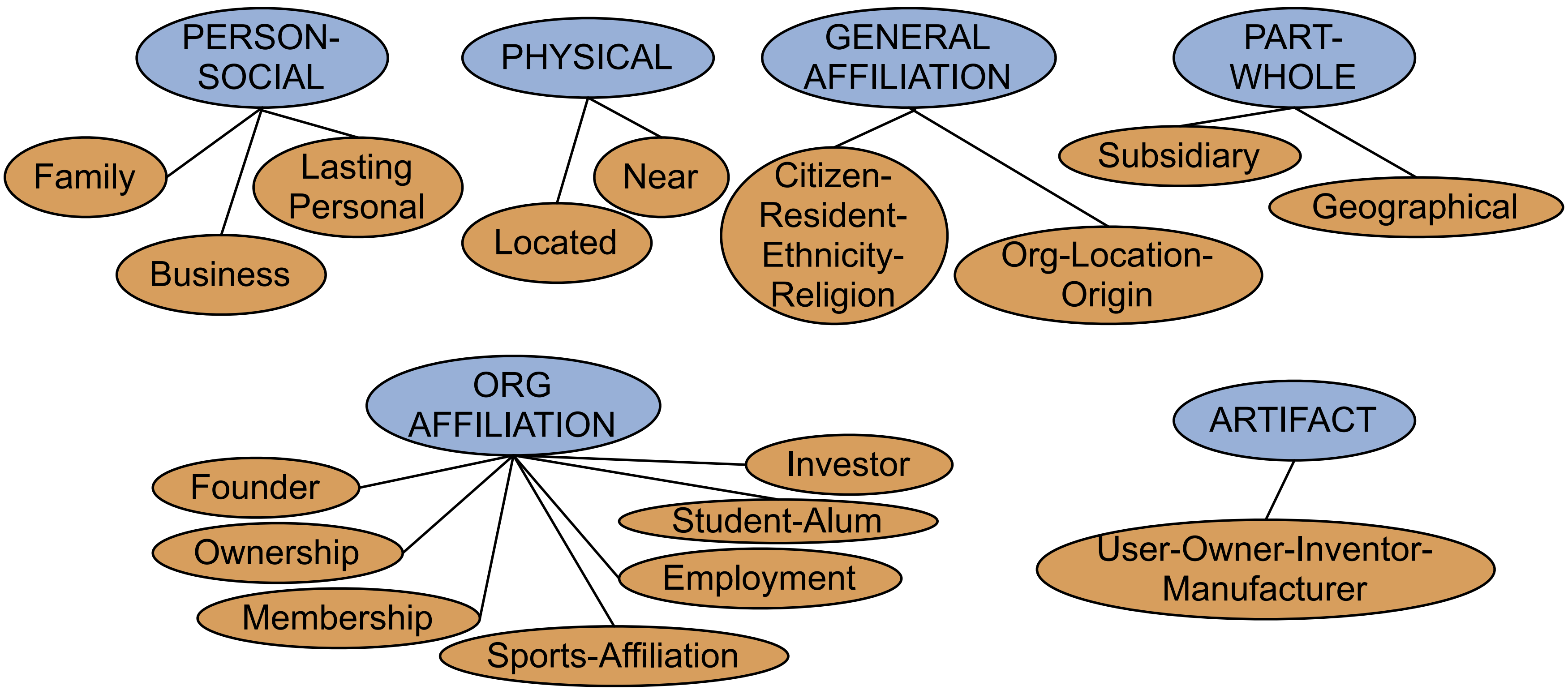
Choose a set of relevant named entities

Find and label data

- Choose a representative corpus
- Label the named entities in the corpus
- Hand-label the relations between these entities
- Break into training, development, and test

Train a classifier on the training set

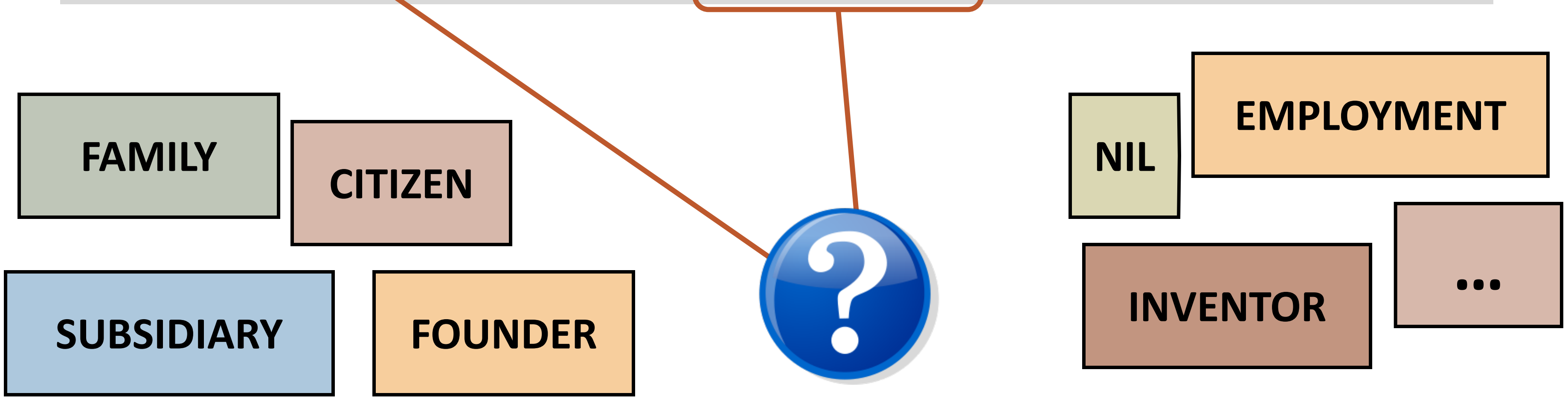
Automated Content Extraction



Relation Extraction

Classify the relation between two entities in a sentence

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.



Word Features for Relation Extraction

American Airlines*, a unit of AMR, immediately matched the move, spokesman* ***Tim Wagner*** *said*
Mention 1 Mention 2

Headwords of M1 and M2, and combination

Airlines Wagner Airlines-Wagner

Bag of words and bigrams in M1 and M2

{American, Airlines, Tim, Wagner, American Airlines, Tim Wagner}

Words or bigrams in particular positions left and right of M1/M2

M2: -1 spokesman

M2: +1 said

Bag of words or bigrams between the two entities

{a, AMR, of, immediately, matched, move, spokesman, the, unit}

Named Entity Type and Mention Level Features

American Airlines*, a unit of AMR, immediately matched the move, spokesman* ***Tim Wagner*** *said*

Mention 1 Mention 2

Named-entity types

- M1: ORG
- M2: PERSON

Concatenation of the two named-entity types

- ORG-PERSON

Entity Level of M1 and M2 (NAME, NOMINAL, PRONOUN)

- M1: NAME [it or he would be PRONOUN]
- M2: NAME [the company would be NOMINAL]

Dependency Parse Features for Relation Extraction

American Airlines, *a unit of AMR, immediately matched the move, spokesman* ***Tim Wagner*** *said*

Mention 1 Mention 2

Base syntactic chunk sequence from one to the other

NP NP PP VP NP NP

Constituent path through the tree from one to the other

NP ↑ NP ↑ S ↑ S ↓ NP

Dependency path

Airlines matched Wagner said

Gazeteer and Trigger word features for relation extraction

Trigger list for family: kinship terms
- parent, wife, husband, grandparent, etc. [from WordNet]

Gazeteer:
- Lists of useful geo or geopolitical words
  - Country name list
  - Other sub-entities

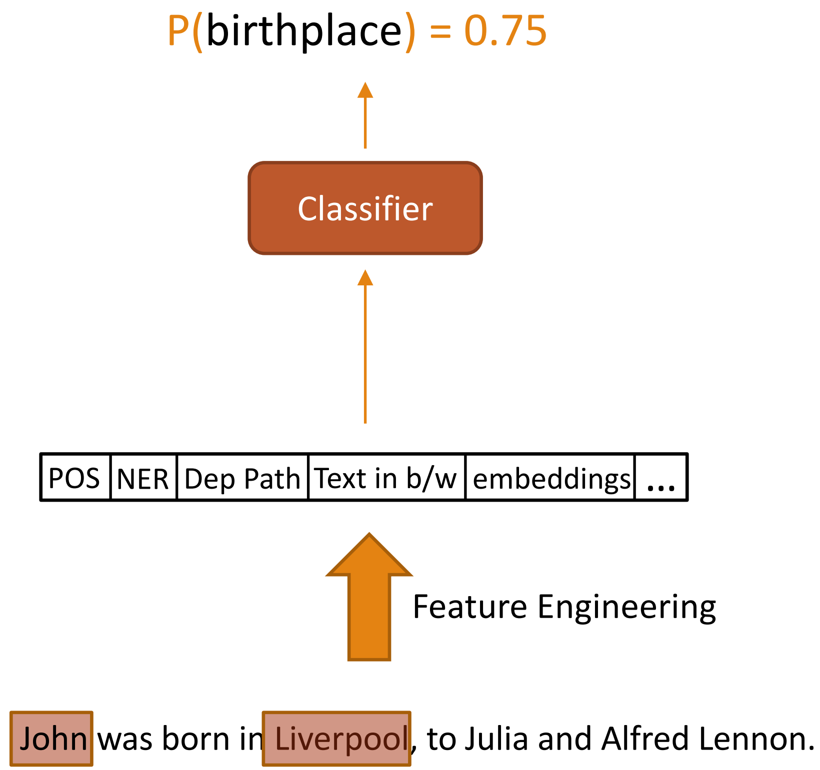
American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Entity-based features	
Entity$_1$ type	ORG
Entity$_1$ head	*airlines*
Entity$_2$ type	PERS
Entity$_2$ head	*Wagner*
Concatenated types	ORGPERS
Word-based features	
Between-entity bag of words	{ *a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* }
Word(s) before Entity$_1$	NONE
Word(s) after Entity$_2$	*said*
Syntactic features	
Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Supervised Extraction

Machine Learning: hopefully, generalizes the labels in the *right way*

Use all of NLP as features: words, POS, NER, dependencies, embeddings



However

Usually, a lot of labeled data is needed, which is expensive & time consuming. Requires a lot of feature engineering!

Supervised Relation Extraction

Pluses

- Can get high accuracies if enough training data
- If test similar enough to training
- Can utilize a number of NLP tasks

Minuses

- Labeling a large training set is expensive
- Supervised models are brittle, don't generalize well to different genres

Outline

Introduction to Relation Extraction

Hand-written Patterns

Supervised Machine Learning

Semi and Unsupervised Learning

Seed-based or bootstrapping approaches to relation extraction

No training set? Maybe you have:

- A few seed tuples or
- A few high-precision patterns

Can you use those seeds to do something useful?

- Bootstrapping: use the seeds to directly learn a relation

Relation Bootstrapping

Gather a set of seed pairs that have the relation

1. Find sentences with these pairs
2. Look at the context between or around the pair and generalize the context to create patterns
3. Use the patterns to gather more pairs
4. Repeat

Bootstrapping Example

<Mark Twain, Elmira> Seed tuple od "died in"

Look for the environments of the seed tuple

"Mark Twain is buried in Elmira, NY."

X is buried in Y

"The grave of Mark Twain is in Elmira"

The grave of X is in Y

"Elmira is Mark Twain's final resting place"

Y is X's final resting place.

Use those patterns to find new tuples

Repeat

Dipre: Extract <author,book> pairs

Start with 5 seeds:

Author	Book
Isaac Asimov	The Robots of Dawn
David Brin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickens	Great Expectations
William Shakespeare	The Comedy of Errors

Find Instances on the Web:

The Comedy of Errors, by William Shakespeare, was
The Comedy of Errors, by William Shakespeare, is
The Comedy of Errors, one of William Shakespeare's earliest attempts
The Comedy of Errors, one of William Shakespeare's most

Extract patterns (group by middle, take longest common prefix/suffix)

`?x , by ?y ,`   `?x , one of ?y ‘s`

Now iterate, finding new seeds that match the pattern

Snowball

Similar iterative algorithm

Organization	Location of Headquarters
Microsoft	Redmond
Exxon	Irving
IBM	Armonk

Group instances w/similar prefix, middle, suffix, extract patterns

- But require that X and Y be named entities
- And compute a confidence for each pattern

.69 ORGANIZATION `{'s, in, headquarters}` LOCATION

.75 LOCATION `{in, based}` ORGANIZATION

Distant Supervision

Combine bootstrapping with supervised learning
- Instead of 5 (or just a few) seeds,
  - Use a large database to get huge # of seed examples
- Create lots of features from all these examples
- Combine in a supervised classifier

Distantly Supervised learning of relation extraction patterns

1	For each relation	Born-In
2	For each tuple in big database	<Edwin Hubble, Marshfield> <Albert Einstein, Ulm>
3	Find sentences in large corpus with both entities	Hubble was born in Marshfield Einstein, born (1879), Ulm Hubble's birthplace in Marshfield
4	Extract frequent features (parse, words, etc)	PER was born in LOC PER, born (XXXX), LOC PER's birthplace in LOC
5	Train supervised classifier using these patterns	$P(\text{born-in} \mid f_1, f_2, f_3, \ldots, f_{70000})$

Distant Supervision Paradigm

Like supervised classification:

- Uses a classifier with lots of features
- Supervised by detailed hand-created knowledge
- Doesn’t require iteratively expanding patterns

Like unsupervised classification:

- Uses very large amounts of unlabeled data
- Not sensitive to genre issues in training corpus

Unsupervised Relation Extraction

Open Information Extraction:

- extract relations from the web with no training data, no list of relations

1. Use parsed data to train a “trustworthy tuple” classifier
2. Single-pass extract all relations between NPs, keep if trustworthy
3. Assessor ranks relations based on text redundancy

(FCI, specializes in, software development)

(Tesla, invented, coil transformer)

(Tesla, inventor of, transformer)

Evaluation of Semi-supervised and Unsupervised Relation Extraction

Since it extracts totally new relations from the web

- There is no gold set of correct instances of relations!
  - Can't compute precision (don't know which ones are correct)
  - Can't compute recall (don't know which ones were missed)

Instead, we can approximate precision (only)

- Draw a random sample of relations from output, check precision manually

$$\hat{P} = \frac{\text{\# of correctly extracted relations in the sample}}{\text{Total \# of extracted relations in the sample}}$$

Can also compute precision at different levels of recall.

- Precision for top 1000 new relations, top 10,000 new relations, top 100,000
- In each case taking a random sample of that set

But no way to evaluate recall

Outline

Introduction to Relation Extraction

Hand-written Patterns

Supervised Machine Learning

Semi and Unsupervised Learning

Upcoming...

Homework

- Homework 3 is due on **February 27**
- Write-up and data has been released.

Project

- Status report due in 1.5 weeks: **March 2, 2017**
- Instructions coming soon
- Only 5 pages

Summaries

- Paper summaries: **February 28**, March 14
- Only 1 page each