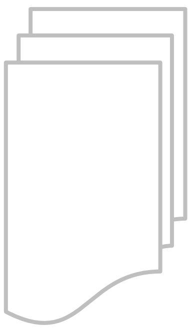
Knowledge Graph Construction from Text

AAAI 2017

Jay Pujara, Sameer Singh, Bhavana Dalvi

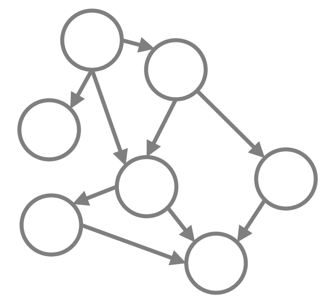
Tutorial Overview

Part 1: Knowledge Graphs



Part 2: Knowledge Extraction

Part 3: Graph Construction



Part 4: Critical Analysis

Tutorial Outline

1. Knowledge Graph Primer [Jay] 
2. Knowledge Extraction from Text
   a. NLP Fundamentals [Sameer] 
   b. Information Extraction [Bhavana] 

Coffee Break 

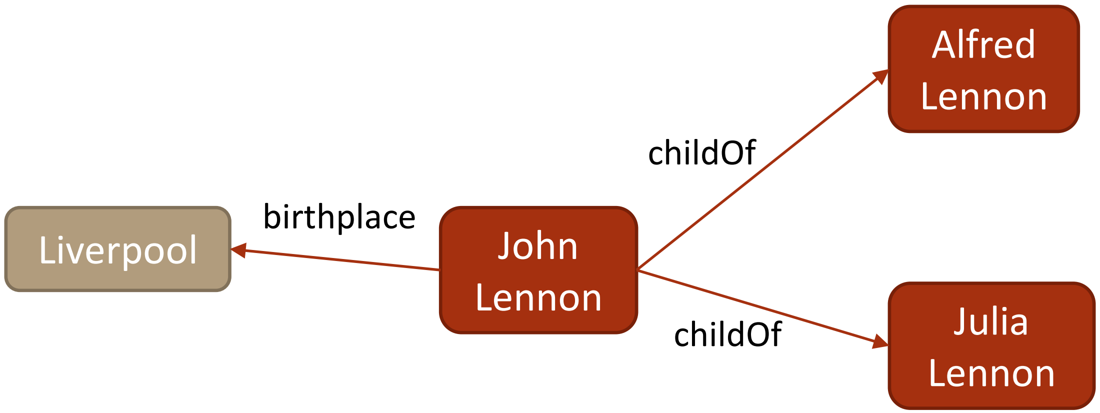
3. Knowledge Graph Construction
   a. Probabilistic Models [Jay] 
   b. Embedding Techniques [Sameer] 
4. Critical Overview and Conclusion [Bhavana] 

What is Knowledge Extraction?

Text

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

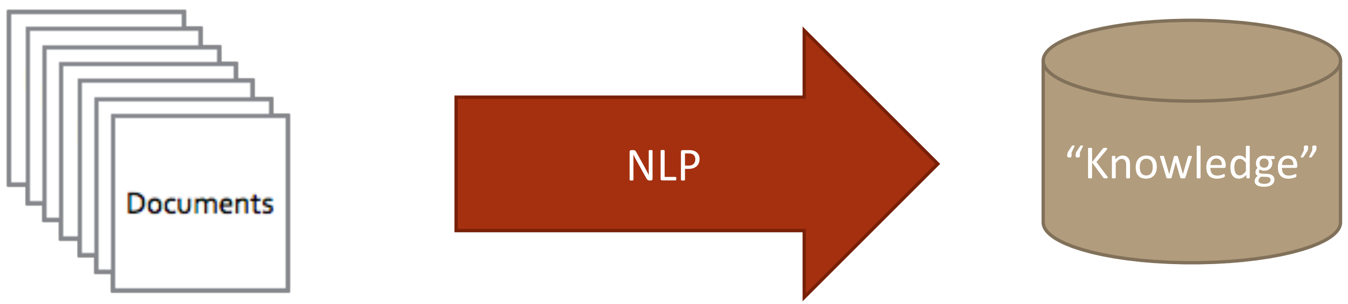
Literal Facts



NLP Fundamentals

EXTRACTING STRUCTURES FROM LANGUAGE

What is NLP?



Unstructured
Ambiguous
Lots and lots of it!

Structured
Precise, Actionable
Specific to the task

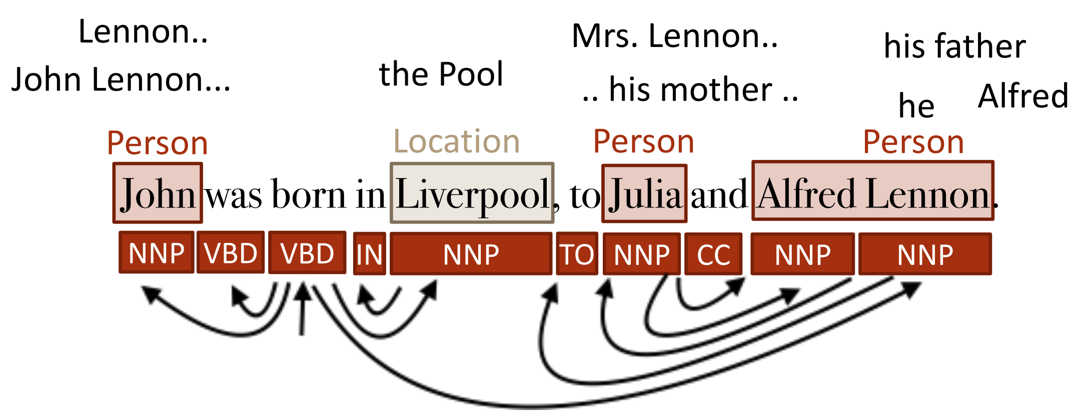
Humans can read them, but
… very slowly
… can’t remember all
… can’t answer questions

Can be used for downstream applications, such as creating Knowledge Graphs!

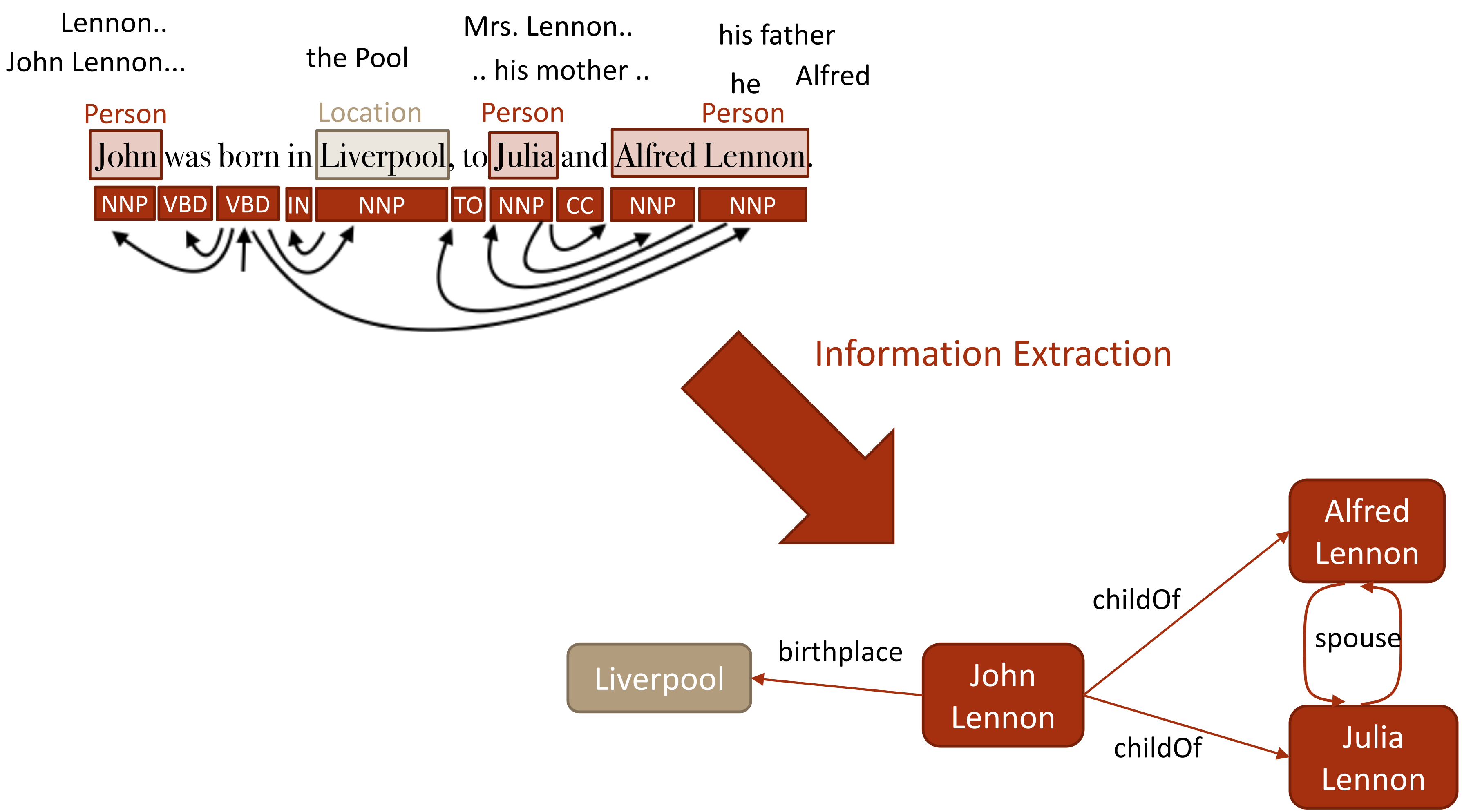
What is NLP?

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

Natural Language Processing



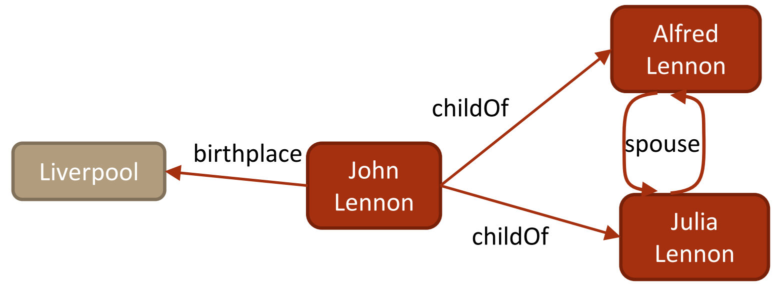
What is Information Extraction?



Breaking it Down

Information Extraction

Entity resolution,
Entity linking,
Relation extraction…



Document

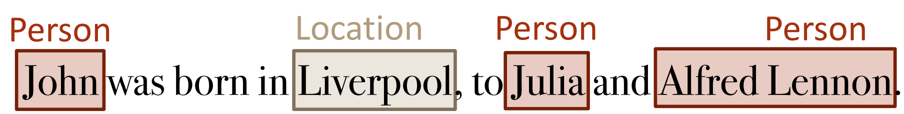
Coreference Resolution…

Lennon..
John Lennon…

the Pool

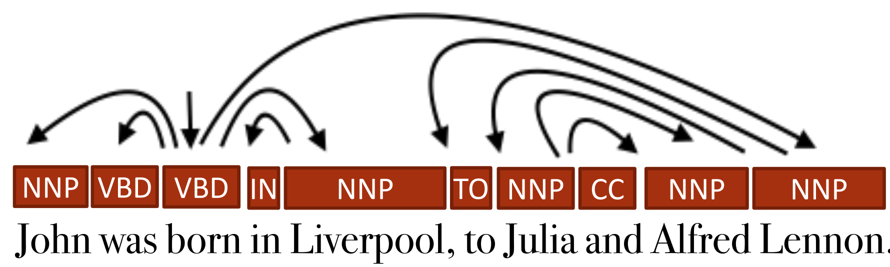
Mrs. Lennon..
.. his mother ..

his father
he
Alfred



Sentence

Dependency Parsing,
Part of speech tagging,
Named entity recognition…



Tokenization & Sentence Splitting

"Mr. Bob Dobolina is thinkin' of a master plan. Why doesn't he quit?"

[Mr.] [Bob] [Dobolina] [is] [thinkin'] [of] [a] [master] [plan] [.]
[Why] [doesn't] [he] [quit] [?]

How it is done:

- Regular expressions, but not trivial
  - Mr., Yahoo!, lower-case
- For non-English, incredibly difficult!
  - Chinese: no "space" character
- Non-trivial for some domains…
  - What is a "token" in BioNLP?

Uses in KG Construction:

- Strictly constrains other NLP tasks
  - Parts of Speech
  - Dependency Parsing
- Directly effects KG nodes/edges
  - Mention boundaries
  - Relations within sentences

Tagging the Parts of Speech

NNP VBD VBD IN NNP TO NNP CC NNP NNP

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

How it is done:

- Context is important!
  - run, table, bar, …
- Label whole sentence together
  - "Structured prediction"
- Conditional Random Fields, ..
- Now: CNNs, LSTMs, …

Uses in KG Construction:

- Entities appear as nouns
- Verbs are very useful
  - For identifying relations
  - For identifying entity types
- Important for downstream NLP
  - NER, Dependency Parsing, …

Detecting Named Entities

Person Location Person Person

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

How it is done:

- Context is important!
  - Georgia, Washington, …
  - John Deere, Thomas Cook, …
  - Princeton, Amazon, …
- Label whole sentence together
  - Structured prediction again

Uses in KG Construction:

- Mentions describes the nodes
- Types are incredibly important!
  - Often restrict relations
- Fine-grained types are informative!
  - Brooklyn: city
  - Sanders: politician, senator

NER: Entity Types

Stanford CoreNLP

3 class:	Location, Person, Organization
4 class:	Location, Person, Organization, Misc
7 class:	Location, Person, Organization, Money, Percent, Date, Time

spaCy.io

PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FACILITY	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LANGUAGE	Any named language.

From Stanford CoreNLP (http://nlp.stanford.edu/software/CRF-NER.shtml)

NER: Entity Types

Fine-grained Types

person	doctor
actor	engineer
architect	monarch
artist	musician
athlete	politician
author	religious_leader
coach	soldier
director	terrorist

organization	terrorist_organization
airline	government_agency
company	government
educational_institution	political_party
fraternity_sorority	educational_department
sports_league	military
sports_team	news_agency

location	body_of_water
city	island
country	mountain
county	glacier
province	astral_body
railway	cemetery
road	park
bridge	

product	camera
engine	mobile_phone
airplane	computer
car	software
ship	game
spacecraft	instrument
train	weapon

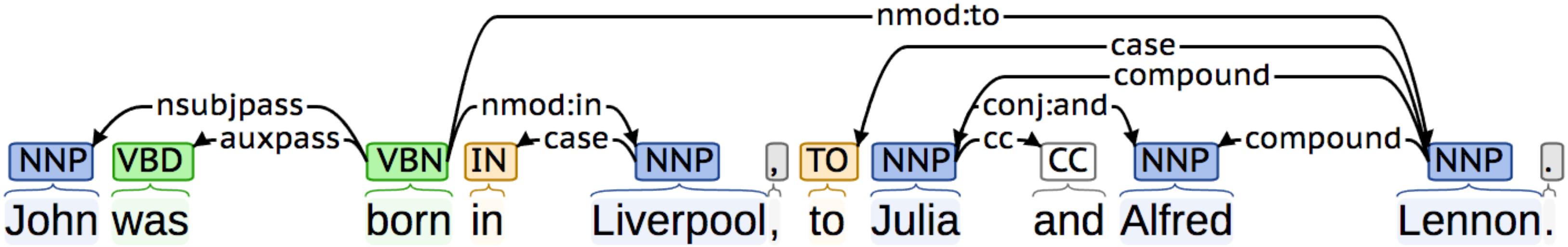
art	written_work
film	newspaper
play	music

event	military_conflict
attack	natural_disaster
election	sports_event
protest	terrorist_attack

building	time	chemical_thing	website
airport	color	biological_thing	broadcast_network
dam	award	medical_treatment	broadcast_program
hospital	educational_degree	disease	tv_channel
hotel	title	symptom	currency
library	law	drug	stock_exchange
power_station	ethnicity	body_part	algorithm
restaurant	language	living_thing	programming_language
sports_facility	religion	animal	transit_system
theater	god	food	transit_line

- More on this later…

Dependency Parsing



How it is done:

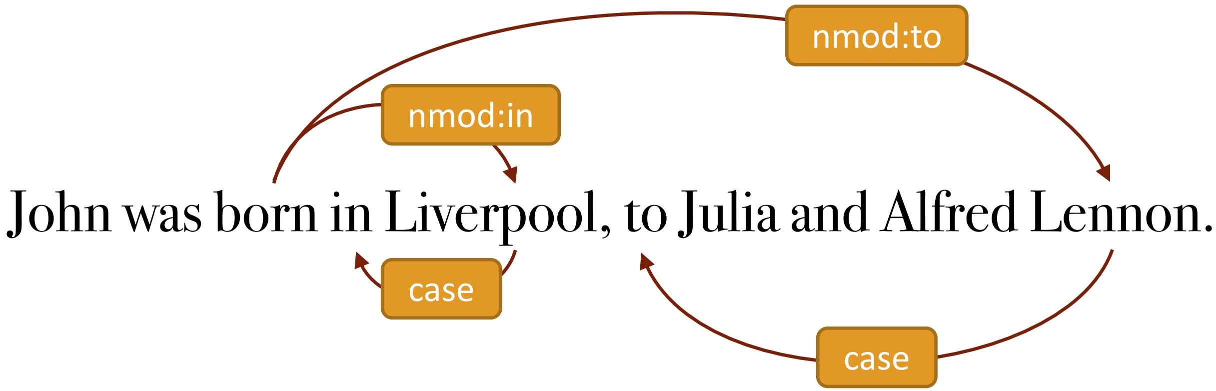
- **Model**: score trees using features
  - **Lexical**: words, POS, …
  - **Structure**: distance, …
- **Prediction**: Search over trees
  - greedy, spanning tree, belief propagation, dynamic prog, …

Uses in KG Construction:

- Incredibly useful for **relations**!
  - What verb is attached?
  - Relation to which mention?
- Incredibly useful for **attributes**!
  - Appositives: “X, the CEO, …”
- Paths are used as **surface relations**

Using http://nlp.stanford.edu:8080/corenlp/process

Dependency Paths



	Text Patterns	Dependency Paths
John, Liverpool	“was born in”	“was born in”
John, Julia	“was born in Liverpool, to”	“was born to”
John, Alfred Lennon	“was born in Liverpool, to Julia and”	“was born to”

Within-document Coreference

He…

Lennon..

John Lennon...

the Pool

Mrs. Lennon..

.. his mother ..

Alfred

his father

he

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

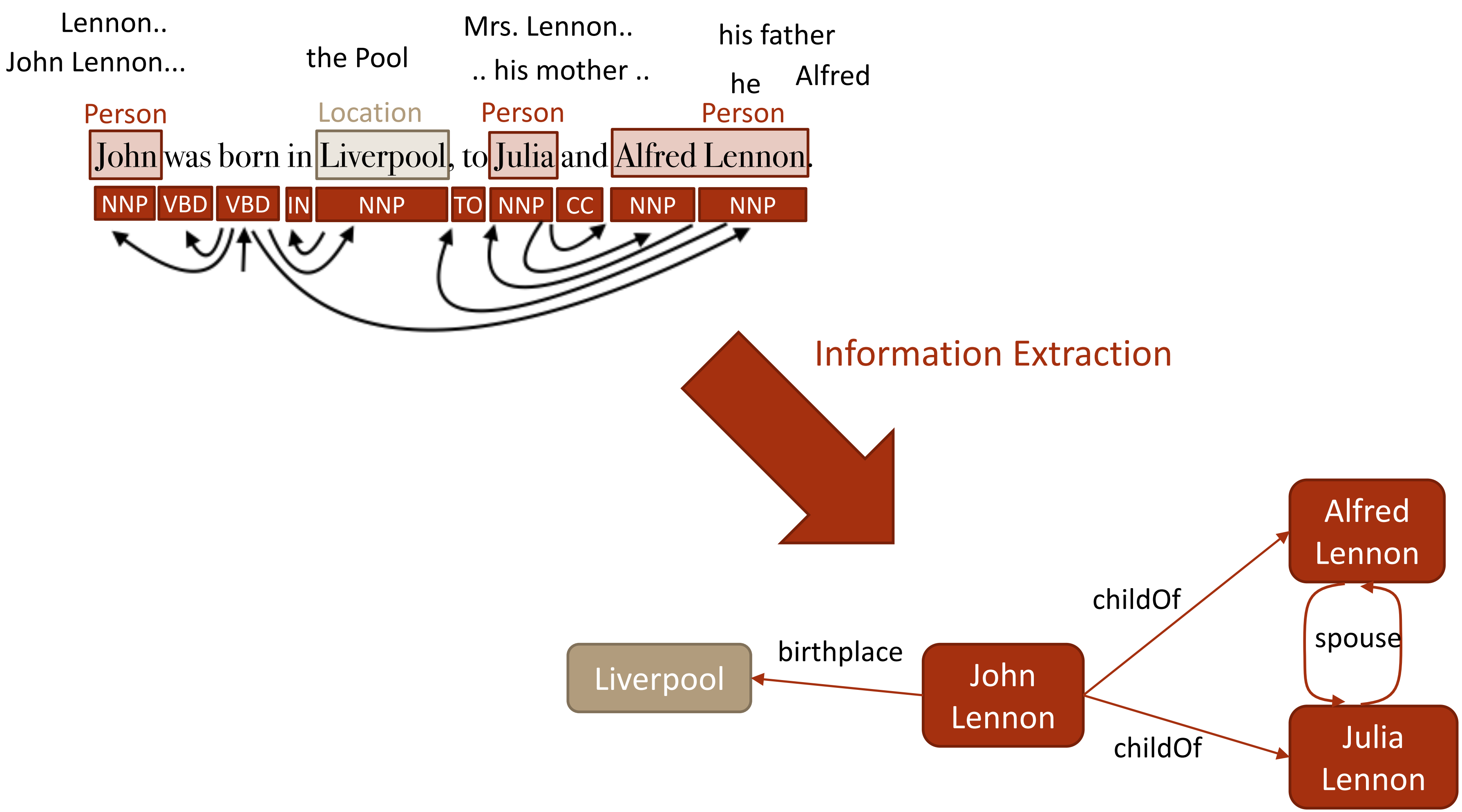
How it is done:

- Mo`del: score pairwise links
  - dep path, similarity, types, …
  - “representative mention”
- Prediction: Search over clusterings
  - greedy (left to right), ILP, belief propagation, MCMC, …

Uses in KG Construction:

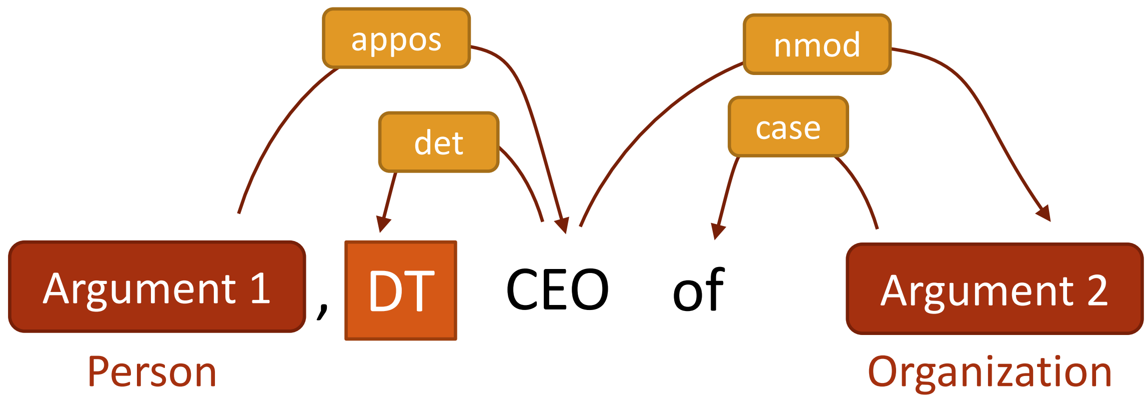
- More context for each entity!
- Many relations occur on pronouns
  - “He is married to her”
- Coref can be used for types
  - Nominals: The president, …
- Difficult, so often ignored

Information Extraction



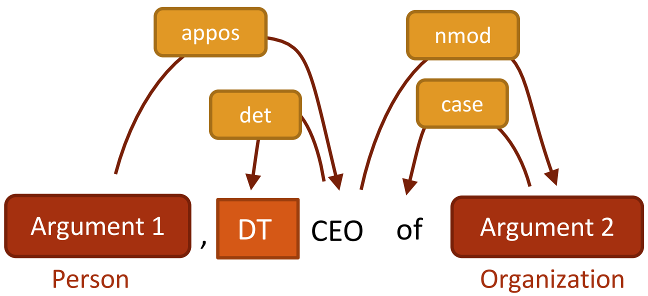
Surface Patterns

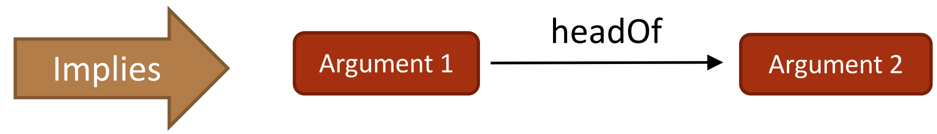
Combine tokens, dependency paths, and entity types to define rules.



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)

High precision: when it fires, it’s correct
Easy to explain predictions
Easy to fix mistakes

However…
Only work when the rules fire
Poor recall: Do not generalize!

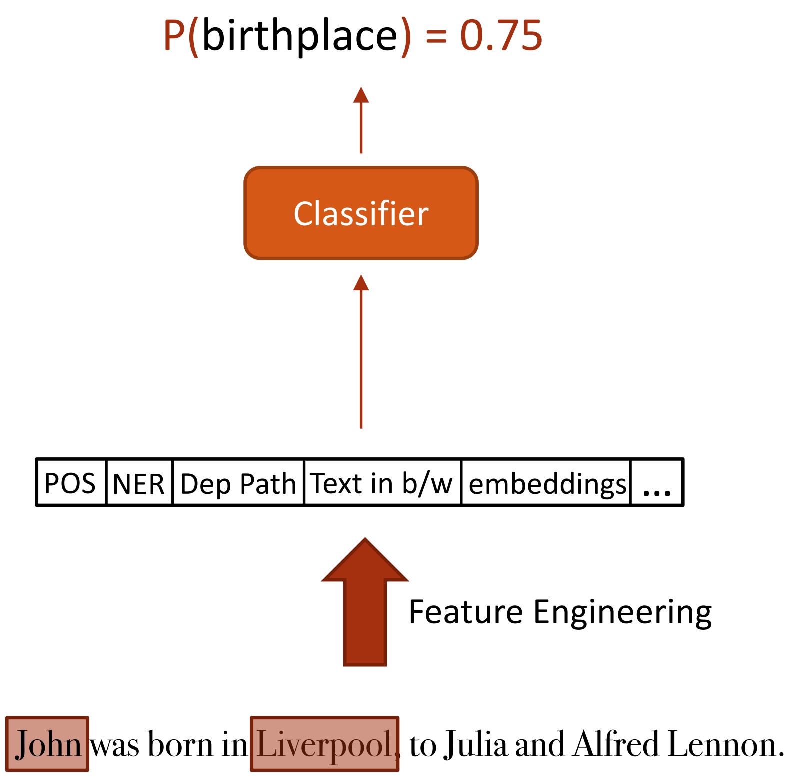
Supervised Extraction

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!



Entity Resolution & Linking

…during the late 60′s and early 70′s, **Kevin Smith** worked with several local…

…the term hip-hop is attributed to **Lovebug Starski**. What does it actually mean…

Like Back in 2008, the Lions drafted **Kevin Smith**, even though Smith was badly…

… backfield in the wake of **Kevin Smith**′s knee injury, and the addition of Haynesworth…

The filmmaker **Kevin Smith** returns to the role of Silent Bob…

Nothing could be more irrelevant to **Kevin Smith**′s audacious ′′Dogma′′ than ticking off…

… The Physiological Basis of Politics,” by **Kevin Smith**, Douglas Oxley, Matthew Hibbing…

Entity Names: Two Main Problems

Entities with Same Name

Same type of entities share names

Kevin Smith, John Smith, Springfield, …

Things named after each other

Clinton, Washington, Paris, Amazon, Princeton, Kingston, …

Partial Reference

First names of people, Location instead of team name, Nick names

Different Names for Entities

Nick Names

Bam Bam, Drumpf, …

Typos/Misspellings

Baarak, Barak, Barrack, …

Inconsistent References

MSFT, APPL, GOOG…

Entity Linking Approach

Washington drops 10 points after game with UCLA Bruins.

Candidate Generation

Washington DC, George Washington, Washington state, Lake Washington, Washington Huskies, Denzel Washington, University of Washington, Washington High School, …

Entity Types — LOC/ORG

Washington DC, ~~George Washington~~, Washington state, Lake Washington, Washington Huskies, ~~Denzel Washington~~, University of Washington, Washington High School, …

Coreference — UWashington, Huskies

~~Washington DC~~, ~~George Washington~~, ~~Washington state~~, ~~Lake Washington~~, Washington Huskies, ~~Denzel Washington~~, University of Washington, ~~Washington High School~~, …

Coherence — UCLA Bruins, USC Trojans

~~Washington DC~~, ~~George Washington~~, ~~Washington state~~, ~~Lake Washington~~, Washington Huskies, ~~Denzel Washington~~, ~~University of Washington~~, ~~Washington High School~~, …

Information Extraction