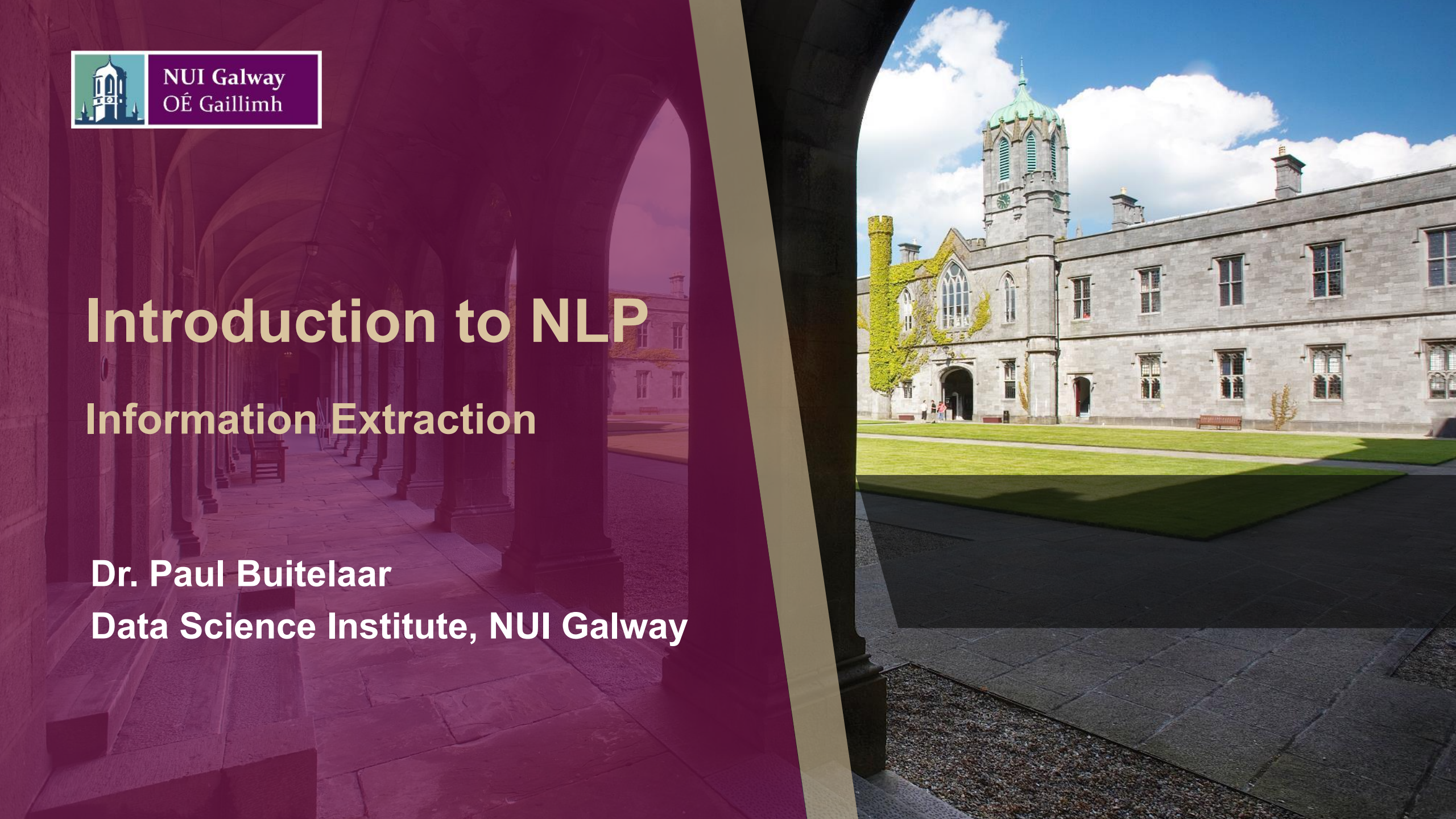


NUI Galway
OÉ Gaillimh

Introduction to NLP

Information Extraction

Dr. Paul Buitelaar
Data Science Institute, NUI Galway

Learning Outcomes of This Lecture

Understand the basic idea and application of information extraction

Understand information extraction sub-tasks, in particular entity recognition / entity linking and relation extraction

Gain insight into developing a relation extraction approach, in particular by use of supervised, semi-supervised and unsupervised methods

Gain insight into evaluating information extraction approaches

Overview

Information Extraction (IE)

IE Approaches

Named Entity Recognition and Entity Linking

Relation Extraction

Evaluation of IE

Overview

Information Extraction (IE)

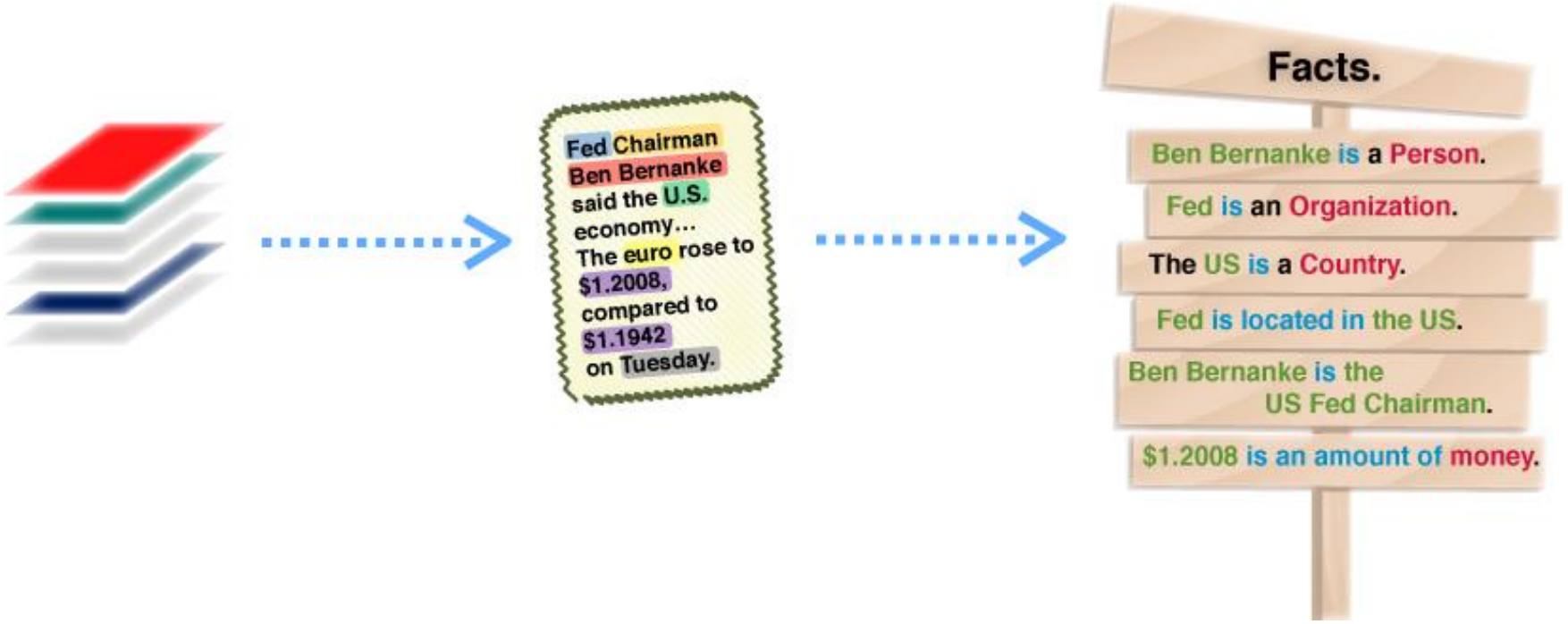
IE Approaches

Named Entity Recognition and Entity Linking

Relation Extraction

Evaluation of IE

Information Extraction – Basic Idea



Some Applications of Information Extraction

Semantic normalization

- Normalize variations across a data set to a canonical form

Database curation

- Complete and/or correct a database with information from textual data

Fact extraction

- Extract facts (entities, relations, events) from textual data

Semantic Normalization

Normalize instances to a canonical form, e.g.

Ireland, IRL, IE, Eire ⇒ *IRL*

Obama, Barack, President Obama ⇒ *POTUS44*

Often for statistical purposes to aggregate data across variations

Semantic Normalization

Normalize different entity instances to a class, e.g.

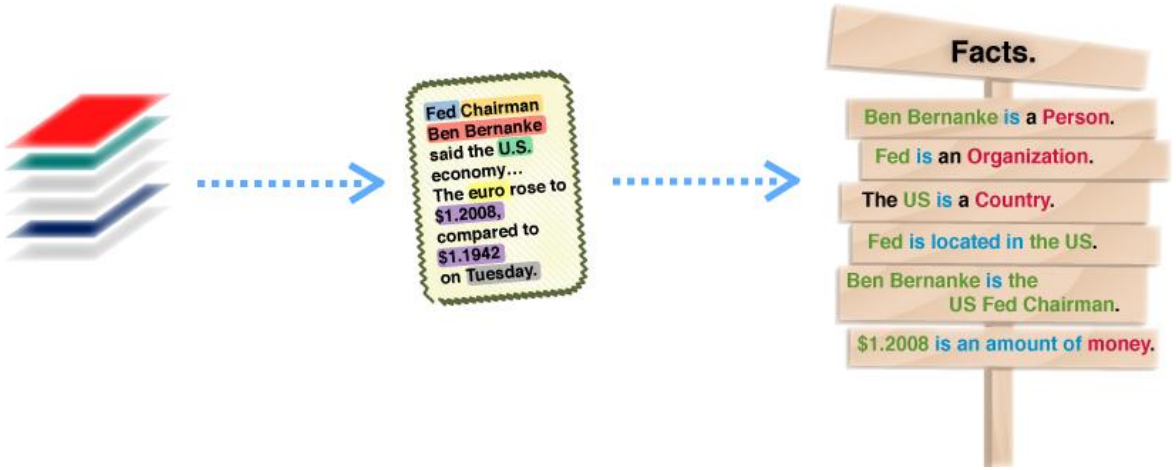
1930, 2015, 1066	⇒ *YEAR*
Ireland, Germany, India, the Netherlands	⇒ *COUNTRY*
IBM, Google, Apple, Irish Times, Amazon	⇒ *COMPANY*
Obama, Trump, Clinton, Bush, Reagan, Nixon	⇒ *US-PRESIDENT*

Database Curation

Completion and/or correction of a database

Curate information on missing or incorrect attributes (‘address’, ‘position’, …)

IE provides semi-automated support by checking a DB against related textual data



Fact Extraction - Entities

Named Entity Recognition (NER) systems typically cover small set of entity types

PERSON	*Donald Trump, Barack Obama, the Queen, …*
ORGANIZATION	*NASA, IBM, Dell, Galway City Council, …*
LOCATION	*Dublin, Galway, Ireland, Lower Dangan, …*
TIME	*12pm, August 20th, Xmas day, …*

Fact Extraction - Other Entities

TECHNOLOGY	*Java* *Python* *WordPress* *...*
PRODUCT	*iPhone* *Samsung Galaxy S9* *Dell XPS 15* *...*

Fact Extraction - Relations

We can also identify and extract relations between entities

> *"In March 2014, Facebook CEO Mark Zuckerberg agreed to acquire Oculus VR for US$2.3 billion in cash and stock."* – Wikipedia
>
> ACQUISITION (ORG *Facebook*, ORG *Oculus VR*, TIME *2014*)

Overview

Information Extraction (IE)

IE Approaches

Named Entity Recognition and Entity Linking

Relation Extraction

Evaluation of IE

Information Extraction Approaches

Lexical lookup

Match text with **manually defined lists (Gazetteers)** of common lexical variants for entities

Rules

Match text with **manually defined extraction patterns** for identifying entities/relations

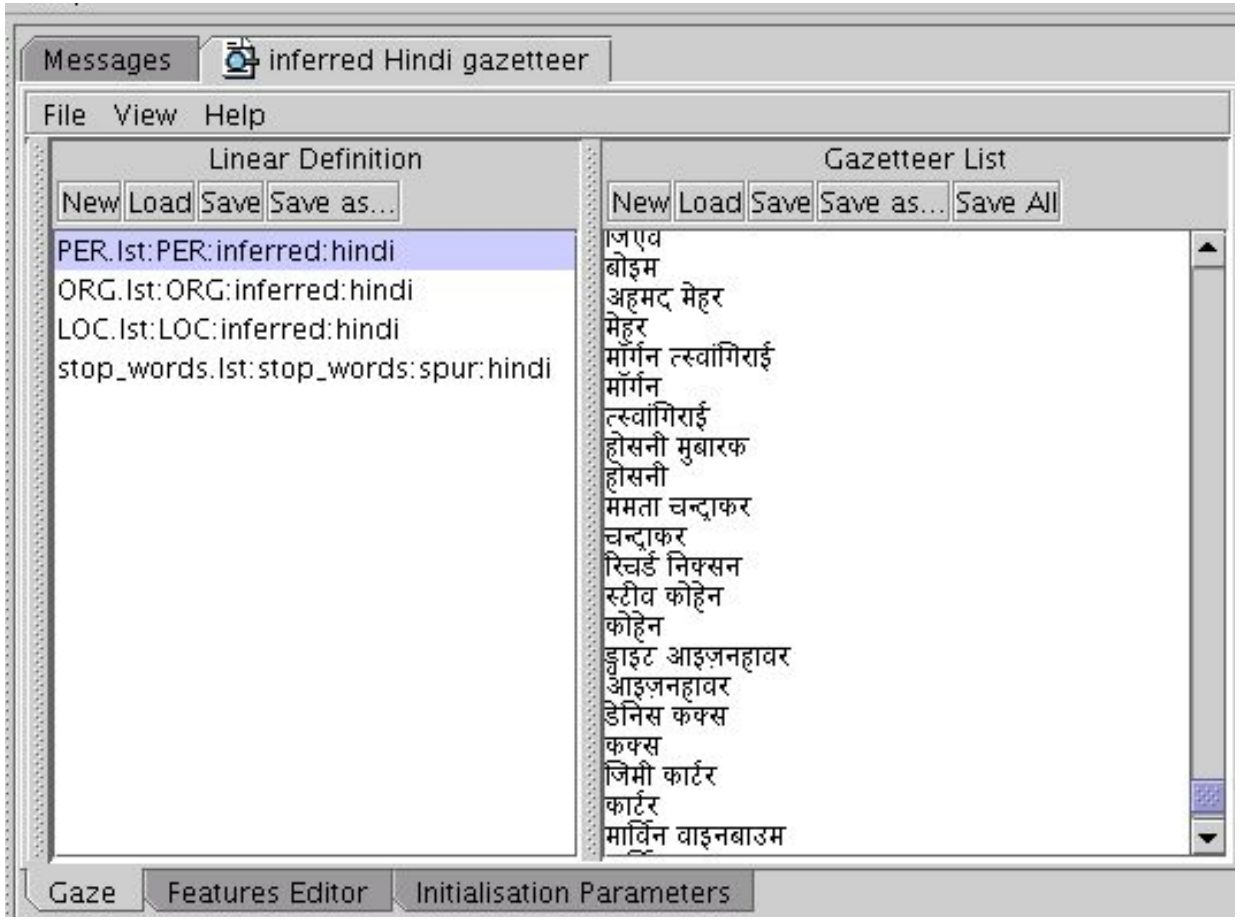
Machine learning

Supervised, semi-supervised, or unsupervised learning of patterns and lexical variants

Jurafsky & Martin Ch17

Lexical Lookup - Gazetteer

List of lexical variants for an entity (India, Ireland) or entity type (LOCATION)



More on gazetteers in Gate: https://gate.ac.uk/sale/tao/splitch13.html#x18-32600013

Lexical Lookup - Ambiguity

A man was caught stealing from an [CITY Aberdeen] gym.

A win for [TEAM Aberdeen] would take them back level on points.

Challenges of IE with Lexical Lookup

Gazetteer needed for each new application domain

- Manual definition – expensive, labour-/time-intensive

How to resolve ambiguities, e.g.

Aberdeen	*CITY*	
	TEAM	
USA	*COUNTRY*	*United States of America*
	COMPANY	*United Space Alliance*
	ORGANIZATION	*University of South Alabama*

Rules - Hearst Patterns

Most commonly used rules in IE are based on so-called ‘Hearst patterns’

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...

Marti A. Hearst, Automatic acquisition of hyponyms from large text corpora (1992). Proceedings of the 14th International Conference on Computational Linguistics (COLING), Nantes, France, pp. 539-545

Hyponymy Extraction with Hearst Patterns

[NP The bow lute], such as [NP the Bambara ndang], is plucked and has an individual curved neck.

IF NP_0 such as {NP_1 , NP_2 … , (and | or)} NP_n THEN for all NP_i $1 \leq i \leq n$, hyponym (NP_i , NP_0)

hyponym (Bambara ndang, bow lute)

Marti A. Hearst, Automatic acquisition of hyponyms from large text corpora (1992). Proceedings of the 14th International Conference on Computational Linguistics (COLING), Nantes, France, pp. 539-545

Relation Extraction with Hearst Patterns

For example: relation **POS**ition, between a **PER**son and an **ORG**anisation

Annotated Corpus: *[**PER** George Marshall] was appointed [**ORG** US] [**POS** Secretary of State]*

*[**PER** George Marshall] was named [**ORG** US] [**POS** Secretary of State]*

Extraction Rule: IF **PER** was (appointed|named|…) **ORG POS** THEN **POS(PER,ORG)**

Extracted Relation: Secretary_of_State:**POS** (George_Marshall:**PER**, US:**ORG**)

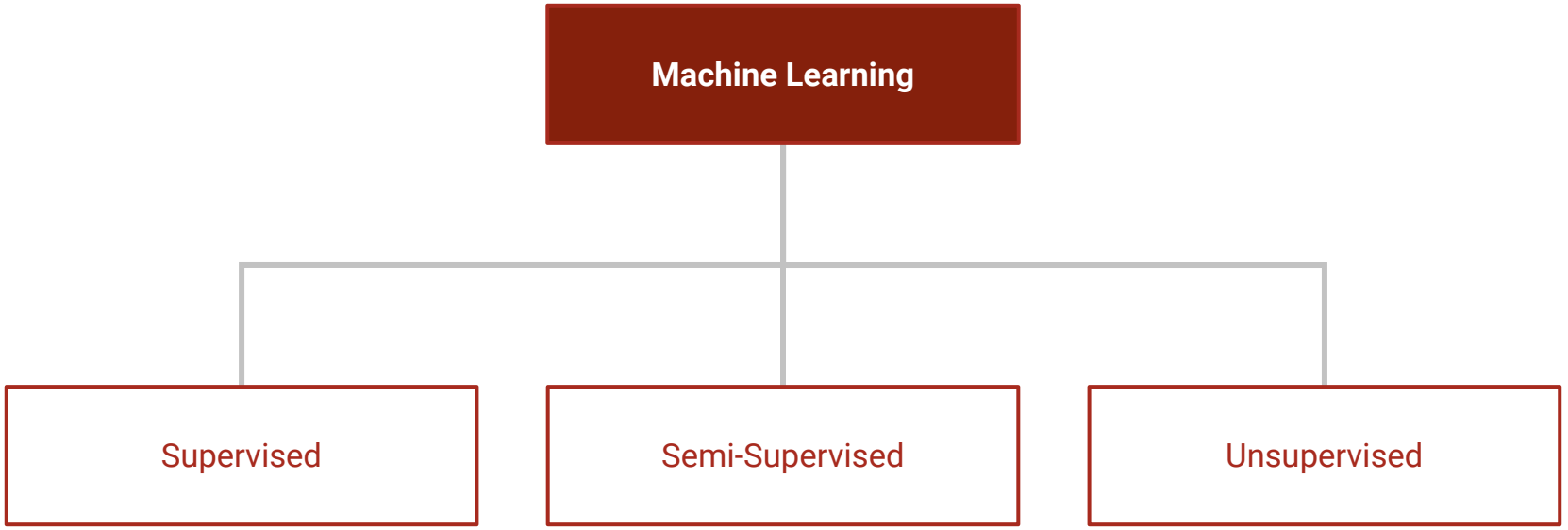
Challenges of Rule-based IE

Rules need to be defined for each new application domain

- Manual definition – expensive, labour-/time-intensive
- Expertise required on domain semantics as well as domain-specific language

Rule execution can be slow

Machine Learning



Supervised ML for IE

Supervised training on labeled data (manual annotation)

Challenges

- Manual annotation required for each new application domain
- Agreement between annotators may be low for complex entities/relations

Inter-Annotator Agreement (IAA)

IAA is a measure of the level of agreement between multiple annotators

Implemented for two annotators using Cohen's kappa coefficient

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

Pr(a): observed agreement

Pr(e): expected agreement

https://corpuslinguisticmethods.wordpress.com/2014/01/15/what-is-inter-annotator-agreement/

IAA Example

Suppose two annotators are asked to annotate a dataset of images of **puppies** and **fried chicken**

Example adapted from https://people.ischool.berkeley.edu/~dbamman/nlpF17/slides/5_truth_ethics.pdf

IAA Example

To calculate Cohen kappa:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

annotator B \ annotator A	puppy	fried chicken
puppy	6	3
fried chicken	2	5

Pr(a): observed agreement 11/16 = 68.75%

Pr(e): expected agreement P(A = Puppy) P(B = Puppy) + P(A = FrCh) P(B = FrCh) = (8/16*9/16) + (8/16*7/16) = 0.5

K = 0.6875 - 0.5 / 1 - 0.5 = 0.375

Example adapted from https://people.ischool.berkeley.edu/~dbamman/nlpF17/slides/5_truth_ethics.pdf

Interpreting IAA

K value	Strength of Agreement (Landis & Koch scale)
< 0.21	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good
0.81 - 1.00	Very Good

Optimizing IAA to K > 0.80 is preferred for data quality

To improve K:

- adjust annotation guidelines for re-annotation
- discuss disagreements between annotators

Semi-Supervised ML for IE

Semi-supervised training on ‘weakly’ labeled data on known examples to bootstrap annotation and subsequent supervised training

Known examples can be taken from existing databases or through ‘distant supervision’ on data sources such as Wikipedia

Unsupervised ML for IE

Bottom-up, data-driven extraction without semi/supervised labeled data

Very scalable and cheap to implement and maintain

Hard to interpret results

Overview

Information Extraction (IE)

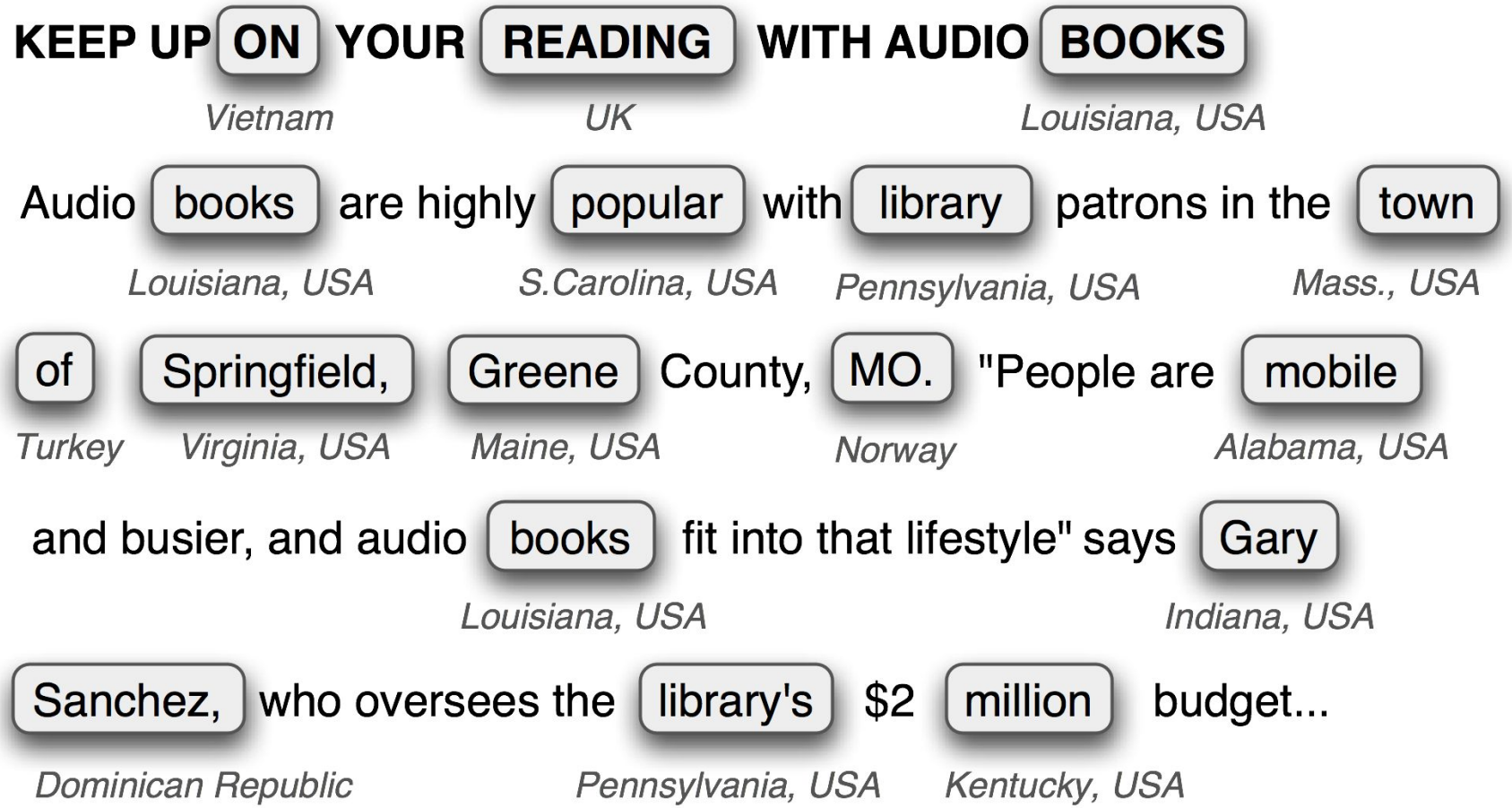
IE Approaches

Named Entity Recognition and Entity Linking

Relation Extraction

Evaluation of IE

Named Entities



http://www.nltk.org/book/ch07.html

Named Entity Recognition

Named Entity Recognition (NER) systems cover a small set of basic NE types, e.g.

PERSON	*Donald Trump, Barack Obama, the Queen, …*
ORGANIZATION	*NASA, IBM, Dell, Galway City Council, …*
LOCATION	*Dublin, Galway, Ireland, Lower Dangan, …*
TIME	*12pm, August 20th, Xmas day, …*

NE Annotation - ENAMEX (XML)

Deere & Co. *said it reached a tentative agreement with the* ***machinist union*** *at its* ***Horicon Wis.*** *plant, ending a month-old strike.*

*<ENAMEX TYPE="**ORGANIZATION:CORPORATION**">**Deere & Co**.</ENAMEX>*

said it reached a tentative agreement with the

*<ENAMEX TYPE=" **ORGANIZATION:OTHER**">**machinist**</ENAMEX>*

*<ENAMEX TYPE="**ORGANIZATION:OTHER**">**union**</ENAMEX>*

at its

*<ENAMEX TYPE="**GEO-POLITICAL ENTITY:CITY**">**Horicon**</ENAMEX> ,*

*<ENAMEX TYPE="**GEO-POLITICAL ENTITY:STATE_PROVINCE**">**Wis**.</ENAMEX> ,*

plant, ending month-old strike.

NE annotation - BIO

NEs annotated with beginning (**B**) and span (**I**nside/**O**utside)

Optionally with entity type (ORGanization, PERson, ...)

Beginning of entity (-TYPE)	B(-ORG)
Inside of entity (-TYPE)	I(-ORG)
Outside of entity	O

Deere	*B-ORG*
&	*I-ORG*
Co.	*I-ORG*
said	*O*
it	*O*
reached	*O*
a	*O*
tentative	*O*
agreement	*O*
with	*O*
the	*O*
machinist	*B-ORG*
union	*I-ORG*
…	*…*

NE Annotation Guidelines

Guidelines instruct human annotators, e.g. ACE (Automatic Content Extraction)

PERSON	single individual or a group
ORGANIZATION	corporations, agencies, and other groups of people defined by an established organizational structure
GEO-POLITICAL ENTITY	geographical regions defined by political and/or social groups
LOCATION	geographical areas, landmasses, bodies of water, geological formations

https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

Entity Linking

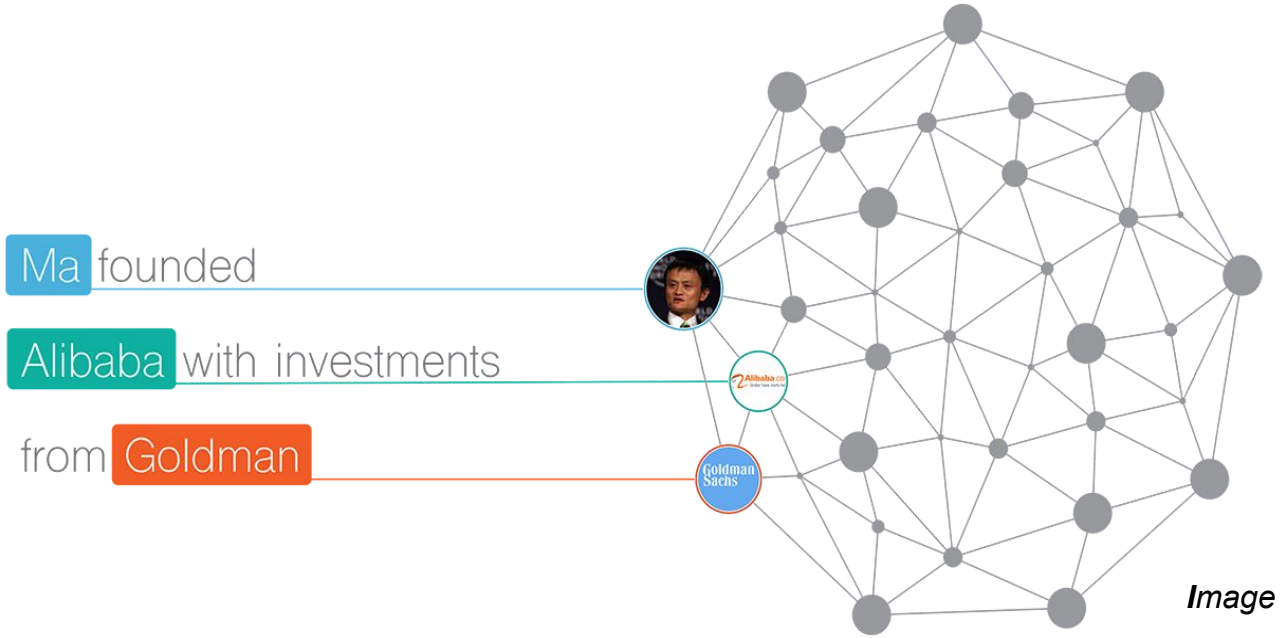


Image by Ambiverse

Beyond small set of basic NE types (Person, Location, Time, …)

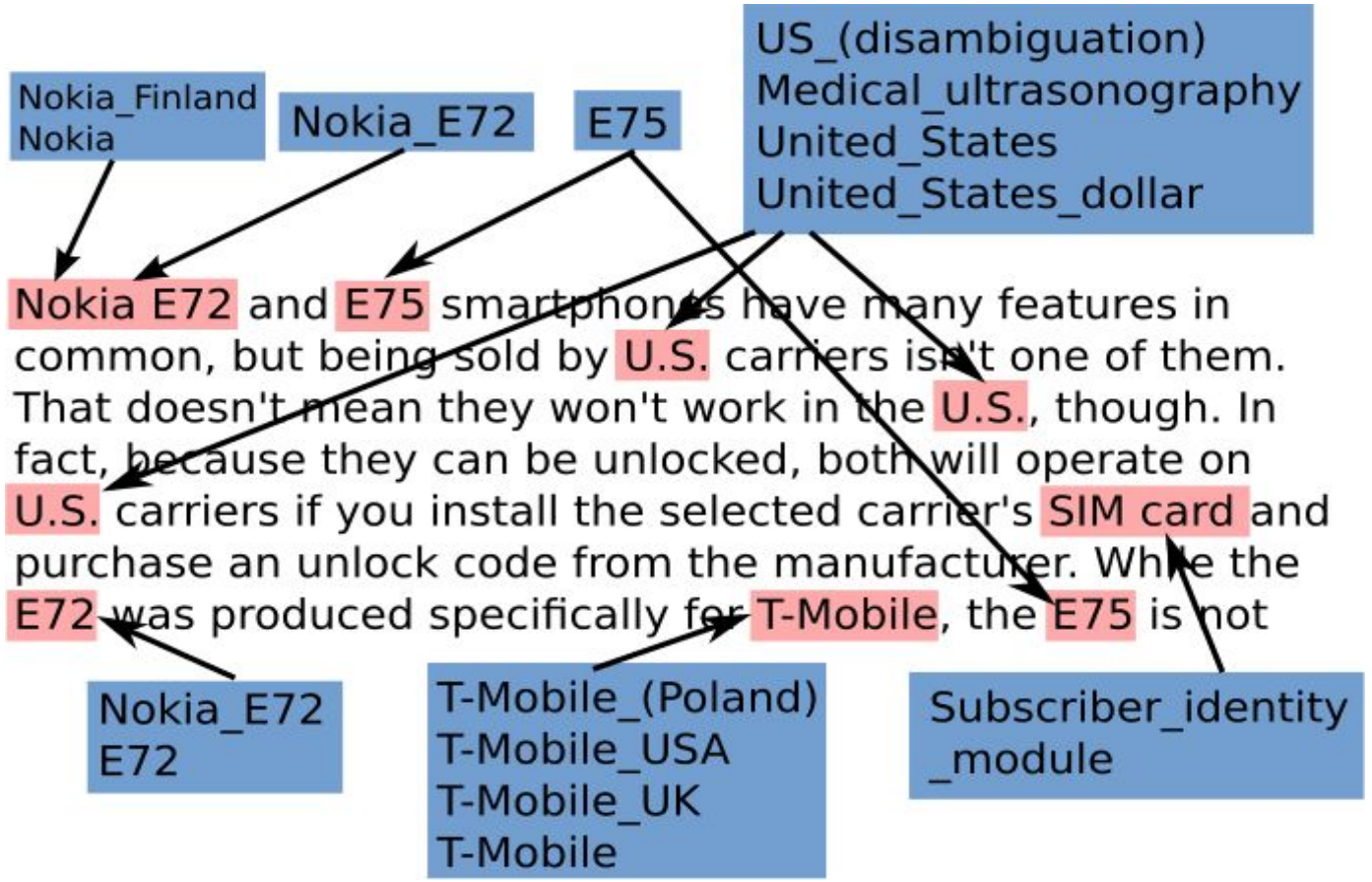
Ground entities in large-scale knowledge graphs (‘DBpedia, YAGO, …’)

Significantly expands the number of entity types

Increases level of ambiguity

https://wiki.dbpedia.org/
https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

Entity Linking – Example



Entity Linking – Mentions

*"**Nokia E72** and **E75** smartphones have many features in common, but being sold by **U.S.** carriers isn't one of them."*

Entity Linking – References & Types

*"**Nokia E72** and E75 smartphones have many features in common, but being sold by U.S. carriers isn't one of them."*

https://en.wikipedia.org/wiki/Nokia_E72 Category:Smartphones

Entity Linking – References & Types

*"Nokia E72 and **E75** smartphones have many features in common, but being sold by U.S. carriers isn't one of them."*

https://en.wikipedia.org/wiki/Nokia_E75	Category:Smartphones
https://en.wikipedia.org/wiki/European_route_E75	Category:International_road_network

Entity Linking – References & Types

*"Nokia E72 and E75 smartphones have many features in common, but being sold by **U.S.** carriers isn't one of them."*

https://en.wikipedia.org/wiki/United_States	Category:Countries_in_North_America
https://en.wikipedia.org/wiki/US_Airways	Category:American_Airlines_Group
https://en.wikipedia.org/wiki/University_of_Salzburg	Category:Universities_in_Austria

Entity Linking – Disambiguation Features

Context around the entity

Popularity of the entity

Coherence across different entities

Features can be **Text-based, Graph-based**

Entity Linking – Context, Text-based

Context words within a window (e.g. Wikipedia pages for ‘Multinational Companies’)

Royal Dutch Shell plc (LSE: RDSA, RDSB), commonly known as **Shell**, is a British-Dutch oil and gas company headquartered in the Netherlands and incorporated in the United Kingdom.[2] It is one of the six oil and gas "supermajors" and the sixth-largest company in the world measured by 2016 revenues (and the largest based in Europe).[1] Shell was first in the 2013 Fortune Global 500 list of the world's largest companies;[3] in that year its revenues were equivalent to 84% of the Netherlands' $556 billion GDP.[4]

incorporated
headquartered
largest

Volkswagen Aktiengesellschaft (German: [ˈfɔlksˌvaːgn̩]), known internationally as the **Volkswagen Group**, is a German multinational automotive manufacturing company headquartered in Wolfsburg, Lower Saxony, Germany and indirectly majority owned by the Austrian Porsche-Piech family.[7][8] It designs, manufactures and distributes passenger and commercial vehicles, motorcycles, engines, and turbomachinery and offers related services including financing, leasing and fleet management. In 2016, it was the world's largest automaker by sales, overtaking Toyota and keeping this title in 2017, selling 10.7 million vehicles.[9] It has maintained the largest market share in Europe for over two decades.[10] It ranked sixth in the 2017 Fortune Global 500 list of the world's largest companies. Volkswagen Group sells passenger cars

sales
headquartered
largest

Entity Linking – Popularity, Text-based

Size of textual description (e.g. compare short vs. long Wikipedia page)

Nokia E75

From Wikipedia, the free encyclopedia

The **Nokia E75** is a smartphone from the Eseries range with a side sliding QWERTY keyboard and also front keypad.[1]

European route E75

From Wikipedia, the free encyclopedia

European route E 75 is part of the International E-road network, which is a series of main roads in Europe.

The E 75 starts at the town of Vardø on the Barents Sea and it runs south through Finland, Poland, Czech Republic, Slovakia, Hungary, Serbia, Macedonia, and Greece. The road ends after about 5,639 kilometres (3,504 mi) at the town of Sitia on eastern end of the island of Crete in the Mediterranean Sea[1], it being the most southerly point reached by an E-road.

From the beginning of the 1990s until 2009, there was no ferry connection between Helsinki and Gdańsk. However, Finnlines started a regular service between Helsinki and Gdynia. It is also possible to take a ferry from Helsinki to Tallinn and drive along the E67 from Tallinn to Piotrków Trybunalski in Poland and then continue with the E75.

Entity Linking – Popularity, Graph-based

Page rank, in/outdegree (e.g. how central is this entity in the Wikipedia graph?)

Image by ontotext

Entity Linking – Coherence, Graph-based

Intersection of sub-graphs, distance, category

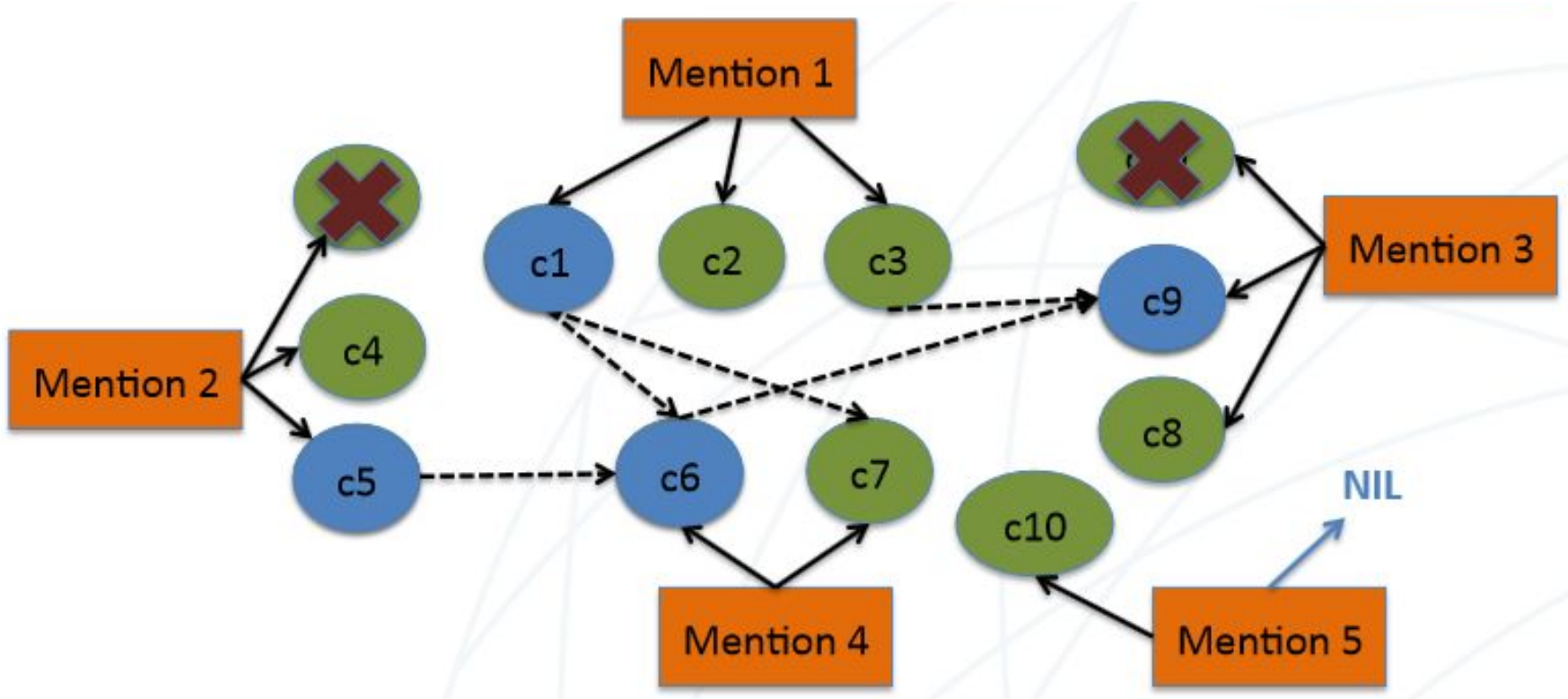


Image by Bianca Pereira

Overview

Information Extraction (IE)

IE Approaches

Named Entity Recognition and Entity Linking

Relation Extraction

Evaluation of IE

Relation Extraction

Extract relations between entities such as PERson has POSition in ORGanisation

[PER George Marshall] served as [ORG US] [POS Secretary of State] .

[PER Mike Pompeo] was sworn in as [POS Director] of the [ORG CIA] .

Relation Annotation - Pos/Neg Instances

Positive instance for PER/ORG/POS > POS (PER, ORG)

George Marshall served as US Secretary of State.

Negative instance for PER/ORG/POS > POS (PER, ORG)

The ambassador said that Rex Tillerson will chair a United Nations meeting.

Relation Annotation - Positive Instances

[PER George Marshall] served as [ORG US] [POS Secretary of State] .

Secretary of State (George Marshall, US)

[PER Mike Pompeo] was sworn in as [POS Director] of the [ORG CIA] .

Director (Mike Pompeo, CIA)

Relation Annotation - Negative Instances

The [POS ambassador] said that [PER Rex Tillerson] will chair a [ORG United Nations] meeting .

ambassador (Rex Tillerson, United Nations)

[PER Pompeo] spoke at the [ORG American Enterprise Institute] with [POS ambassador] Haley .

ambassador (Pompeo, American Enterprise Institute)

Relation Extraction - Features

Positive features

[PER George Marshall] ***served as*** *[ORG US] [POS Secretary of State] .*

[PER Mike Pompeo] was ***sworn in as*** *[POS Director] of the [ORG CIA] .*

Negative features

The [POS ambassador] ***said that*** *[PER Rex Tillerson] will chair a [ORG United Nations] meeting .*

[PER Pompeo] ***spoke at*** *the [ORG American Enterprise Institute] with [POS ambassador] Haley .*

Semi-Supervised Relation Extraction

Exploit existing resources for semi-supervised relation extraction, e.g.

- databases
- ‘distant supervision’ with Wikipedia

Semi-Supervised RelExtr - Databases

For example, database seeds for relation PERSON-IN-POSITION (PER, ORG, POS):

PERSON-IN-POSITION (George Marshall, United States, Secretary of State)

PERSON-IN-POSITION (Rex Tillerson, United States, Secretary of State)

PERSON-IN-POSITION (Rex Tillerson, ExxonMobil, CEO)

PERSON-IN-POSITION (Mike Pompeo, Central Intelligence Agency, Director)

Challenges:

Annotation in text still needs to be established

Database seeds for relations provide only positive instances

Semi-Supervised RelExtr – Distant Supervision

Wikipedia

- Info-box provides seeds
- Corresponding text can serve as ‘annotation’ for training purposes

Rex Wayne Tillerson (born March 23, 1952) is an American former government official and former energy executive who served as the 69th United States Secretary of State from February 1, 2017, to March 31, 2018, under President Donald Trump.[1][2][3] Originally a civil engineer, Tillerson joined Exxon in 1975. He rose to become chairman and chief executive officer of ExxonMobil, holding that position from 2006 until 2017, when he left to join the Trump administration.

Rex Tillerson

69th United States Secretary of State	
In office	February 1, 2017 – March 31, 2018[a]
President	Donald Trump
Deputy	John Sullivan
Preceded by	John Kerry
Succeeded by	Mike Pompeo
Personal details	
Born	Rex Wayne Tillerson March 23, 1952 (age 66) Wichita Falls, Texas, U.S.

Distant Supervision

Info-box seeds

Rex Tillerson

69th United States Secretary of State

In office: February 1, 2017 – March 31, 2018

Corresponding text

Rex Wayne Tillerson (born March 23, 1952) is an American former government official and former energy executive who served as the 69th United States Secretary of State from February 1, 2017, to March 31, 2018, under President Donald Trump.

Distant Supervision

Info-box seeds

Rex Tillerson

69th United States Secretary of State

In office: February 1, 2017 – March 31, 2018

Corresponding text

[Rex Wayne Tillerson] *(born March 23, 1952) is an American former government official and former energy executive who served as the [**69th United States Secretary of State]** from **[February 1, 2017, to March 31, 2018]**, under President Donald Trump.*

Distant Supervision

John Forbes Kerry (born December 11, 1943) is an American politician who served as the 68th United States Secretary of State from 2013 to 2017. A member of the Democratic Party, he previously served as a United States Senator from Massachusetts from 1985 until 2013. He was the Democratic nominee in the 2004 presidential election, losing to Republican incumbent George W. Bush.

John Kerry

68th United States Secretary of State

In office
February 1, 2013 – January 20, 2017

President	Barack Obama
Deputy	William Joseph Burns Wendy Sherman (Acting) Tony Blinken
Preceded by	Hillary Clinton
Succeeded by	Rex Tillerson

Distant Supervision

Barack Hussein Obama II (/bəˈrɑːk huːˈseɪn oʊˈbɑːmə/ (listen);[1] born August 4, 1961) is an American politician who served as the 44th President of the United States from January 20, 2009, to January 20, 2017. A member of the Democratic Party, he was the first African American to assume the presidency and previously served as a United States Senator from Illinois (2005–2008).

Barack Obama

44th President of the United States

In office
January 20, 2009 – January 20, 2017

Vice President Joe Biden

Preceded by George W. Bush

Succeeded by Donald Trump

United States Senator
from Illinois

Distant Supervision

Positive instances for relation **Person-in-Position(PER,POS,FROM,TO)**

[PER Rex Wayne Tillerson] *(born March 23, 1952) is an American former government official and former energy executive who served as the* ***[POS 69th United States Secretary of State]*** *from* ***[FROM February 1, 2017]****, to* ***[TO March 31, 2018]****, under President Donald Trump.*

✓ **Person-in-Position** (Rex Tillerson:**PER**, US Secretary of State:**POS**, 2017:**FROM**, 2018:**TO**)

[PER John Forbes Kerry] *(born December 11, 1943) is an American politician who served as the* ***[POS 68th United States Secretary of State]*** *from* ***[FROM 2013]*** *to* ***[TO 2017]****.*

✓ **Person-in-Position** (John Kerry:**PER**, US Secretary of State:**POS**, 2013:**FROM**, 2017:**TO**)

[PER Barack Hussein Obama II] *(born August 4, 1961) is an American politician who served as the* ***[POS 44th President of the United States]*** *from* ***[FROM January 20, 2009]****, to* ***[TO January 20, 2017]****.*

✓ **Person-in-Position** (Barack Obama:**PER**, President of US:**POS**, 2009:**FROM**, 2017:**TO**)

Distant Supervision

Negative instances for relation **Person-in-Position**

Miles Dewey Davis III (May 26, 1926 – September 28, 1991) was an American jazz trumpeter, bandleader, and composer. He is among the most influential and acclaimed figures in the history of jazz and 20th century music. Davis adopted a variety of musical directions in a five-decade career that kept him at the forefront of many major stylistic developments in jazz.[1]

Miles Davis

Davis photographed in his New York City home by Tom Palumbo, c. 1955–1956

Background information

Birth name	Miles Dewey Davis III
Born	May 26, 1926 Alton, Illinois, U.S.
Died	September 28, 1991 (aged 65) Santa Monica, California, U.S.
Genres	Jazz
Occupation(s)	Musician · bandleader · composer
Instruments	Trumpet · flugelhorn
Years active	1944–1975 · 1980–1991

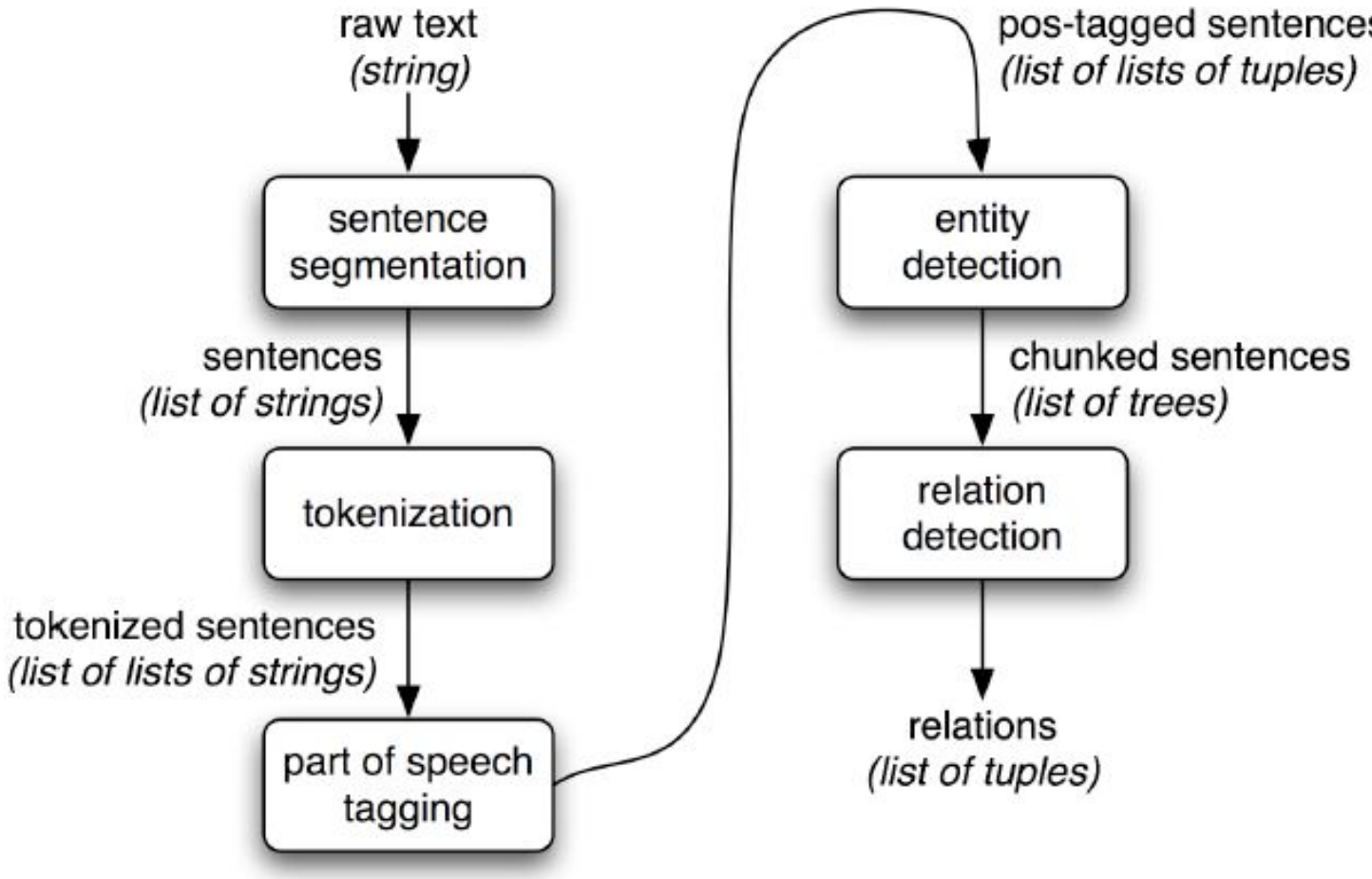
Distant Supervision

Negative instance for relation **Person-in-Position(PER,POS,FROM,TO)**

[PER Miles Dewey Davis III] *(****[FROM May 26, 1926]*** *–* ***[TO September 28, 1991]****) was an American jazz trumpeter,* ***[POS bandleader]****, and composer.*

Person-in-Position (Miles Dewey Davis III:**PER**, bandleader:**POS**, 1926:**FROM**, 1991:**TO**)

Unsupervised Relation Extraction - OpenIE



http://www.nltk.org/book/ch07.html

Corpus Data

John Smith is an expert in Java programming.

Bill Johnson knows Python.

Mary Jones has excellent skills in JavaScript.

PoS Tagging

John/NNP Smith/NNP is an expert/NN in Java/NNP programming.

Bill/NNP Johnson/NNP knows/VBZ Python /NNP.

Mary/NNP Jones/NNP has excellent skills/NN in JavaScript/NNP.

Named Entity Recognition

[NE John/NNP Smith/NNP] *is an expert/NN in [**NE Java/NNP]** programming.*

[NE Bill/NNP Johnson/NNP] *knows/VBZ [**NE Python/NNP]**.*

[NE Mary/NNP Jones/NNP] *has excellent skills/NN in [**NE JavaScript/NNP]**.*

Relation Extraction

*[NE John/NNP Smith/NNP] is an **expert**/NN in [NE Java/NNP] programming.*

*[NE Bill/NNP Johnson/NNP] **knows**/VBZ [NE Python/NNP].*

*[NE Mary/NNP Jones/NNP] has excellent **skills**/NN in [NE JavaScript/NNP].*

Relation Extraction

*[NE **John**/NNP **Smith**/NNP] is an **expert**/NN in [NE **Java**/NNP] programming.*

*[NE **Bill**/NNP **Johnson**/NNP] **knows**/VBZ [NE **Python**/NNP].*

*[NE **Mary**/NNP **Jones**/NNP] has excellent **skills**/NN in [NE **JavaScript**/NNP].*

expert (John Smith, Java)

knows (Bill Johnson, Python)

skills (Mary Jones, JavaScript)

Challenges of Unsupervised Relation Extraction

Very little, or no, semantics (what is exactly extracted?)

Hard to interpret results

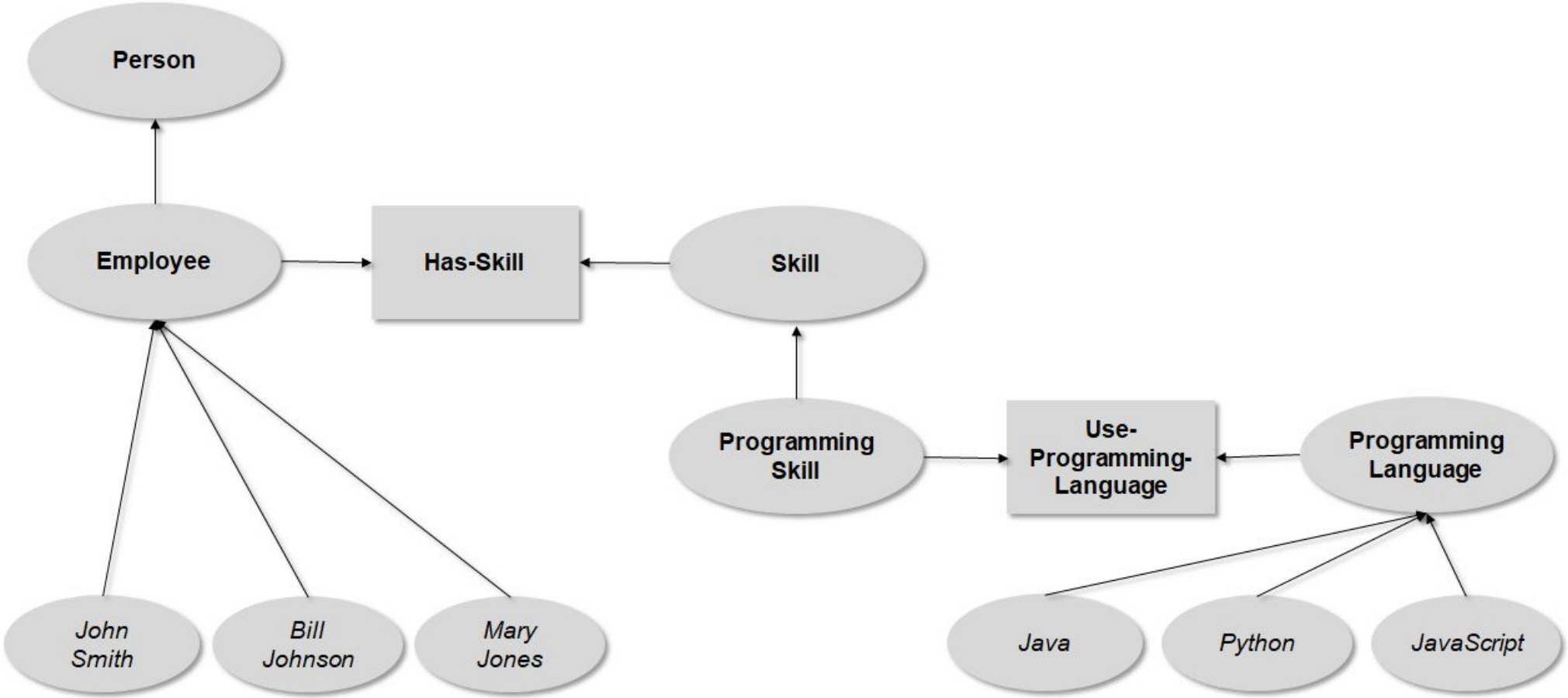
Knowledge Base Population (KBP)

KBP has a knowledge graph against which extraction results can be integrated

No knowledge graph in OpenIE

https://nlp.cs.rpi.edu/kbp/2018/

Knowledge Graph - example



Open IE vs. KBP

[NE John/NNP Smith/NNP] is an expert/NN in [NE Java/NNP] programming.

Open IE expert (John Smith, Java)

KBP **Has-Skill** (John Smith:**Employee**, Java:**Programming Language**)

[NE Bill/NNP Johnson/NNP] knows/VBZ [NE Python/NNP].

Open IE knows (Bill Johnson, Python)

KBP **Has-Skill** (Bill Johnson:**Employee**, Python:**Programming Language**)

[NE Mary/NNP Jones/NNP] has excellent skills/NN in [NE JavaScript/NNP].

Open IE skills (Mary Jones, JavaScript)

KBP **Has-Skill** (Mary Jones:**Employee**, JavaScript:**Programming Language**)

Overview

Information Extraction (IE)

IE Approaches

Named Entity Recognition and Entity Linking

Relation Extraction

Evaluation of IE

Information Extraction Evaluation

Evaluation metrics derive from Information Retrieval: Precision, Recall, F-Score

Evaluation against a “Gold Standard”: human annotated (labelled) data items

IE Evaluation – Measures

Precision $$P = \frac{\#\ correctly\ extracted\ items}{Total\ \#\ of\ extracted\ items}$$

Recall $$R = \frac{\#\ correctly\ extracted\ items}{Total\ \#\ of\ gold\ items}$$

F-Score ('weighted harmonic mean') $$F = \frac{2 \times P \times R}{P + R}$$

IE Evaluation – Example

# Gold Standard items (GS)	**40**
# Extracted items (EX)	**60**
# Correctly extracted items (CEX)	**20**

P = CEX/EX = 20/60 = **0.33**

R = CEX/GS = 20/40 = **0.50**

F = 2*P*R / P+R = 2*0.33*0.5 / 0.33+0.5 = 0.33/0.83 = **~0.40**

IE Evaluation - True/False Positives/Negatives

# Gold Standard items (GS)	**40**
# Extracted items (EX)	**60**
# Correctly extracted items (CEX)	**20**

TP: **20** (CEX)	FN: **20** (GS-CEX)
FP: **40** (EX-CEX)	TN

$$\text{Precision} = \frac{tp}{tp + fp}$$

P = TP/(TP+FP) = 20/60 = **0.33**

$$\text{Recall} = \frac{tp}{tp + fn}$$

R = TP/(TP+FN) = 20/40 = **0.50**

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

Lab of this Week

Exercises in information extraction

QA