Information Extraction (IE)

Liad Magen

Sequence Segmentation and Labeling

- Linguistic angle:
  - The S&P 500 rose toward a two-month high, while the Nasdaq 100 jumped more than 2%.

- Independent word combinations = Constituent:
  - The S&P 500
  - 500 rose
  - toward a
  - jumped more than 2%

Sequence Segmentation and Labeling

- Linguistic angle:
  - The S&P 500 rose toward a two-month high, while the Nasdaq 100 jumped more than 2%.

- Independent word combinations = Constituent:
  - The S&P 500
  - 500 rose
  - toward a
  - jumped more than 2%

Substitution test (pro-form substitution)

- Can the phrase be **replaced** by another noun/verb/adj/adv?

The S&P 500 rose toward a two-month high

It rose toward a two-month high

It rose

It rose up

Phrase Types

- Noun Phrase (NP)
- Verb Phrase (VP)
- Propositional Phrase (PP)
- Adjectival Phrase (ADJP)
- Adverbial Phrase (ADVP)

Chunking

Input:

```
The S&P 500 rose toward a two-month high,
while the Nasdaq 100 jumped more than 2%.
```

Output:

[The S&P 500]NP rose toward **[a two-month high]PP**, while the **[Nasdaq 100]NP** jumped **[more than 2%]PP**

Chunking

Input - Tokenized tokens

```
1 The
2 S&P
3 500
4 rose
5 toward
6  a
7 two
8 month
9 high
```

Output – set of triplets:

```
<1, 3, NP>
<6, 9, PP>
```

…

Why do we need to know this?

China's top legislative body on Wednesday passed a resolution allowing for the disqualification of any Hong Kong lawmakers who aren't deemed sufficiently loyal. Chief Executive Carrie Lam's government immediately banished four legislators, prompting the remaining 15 in the 70-seat Legislative Council to resign in masse hours later at a joint press briefing.

Why do we need to know this?

China's top legislative body on **Wednesday** passed a resolution allowing for the disqualification of any **Hong Kong lawmakers** who aren't deemed sufficiently loyal. **Chief Executive Carrie Lam's** government immediately banished four legislators, prompting the remaining 15 in the 70-seat Legislative Council to resign in masse hours later at a joint press briefing.

Useful combinations in NLP: Named Entity Recognition (NER)

China's top legislative body on **Wednesday** passed a resolution allowing for the disqualification of any **Hong Kong lawmakers** who aren't deemed sufficiently loyal. **Chief Executive Carrie Lam's** government immediately banished four legislators, prompting the remaining 15 in the 70-seat Legislative Council to resign en masse hours later at a joint press briefing.

Organization | Location | Person | Temporal

Information Extraction

Paris Whitney Hilton born February 17, 1981 is an American television personality and businesswoman . She is the great-granddaughter of Conrad Hilton , the founder of Hilton Hotels . Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency

Organization | Location | Person | Temporal

Information Extraction

Paris Whitney Hilton born February 17, 1981 is an American television personality and businesswoman . She is the great-granddaughter of Conrad Hilton , the founder of Hilton Hotels . Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with **Donald Trump**'s modeling agency

Organization | Location | Person | Temporal

Information Extraction – medicine

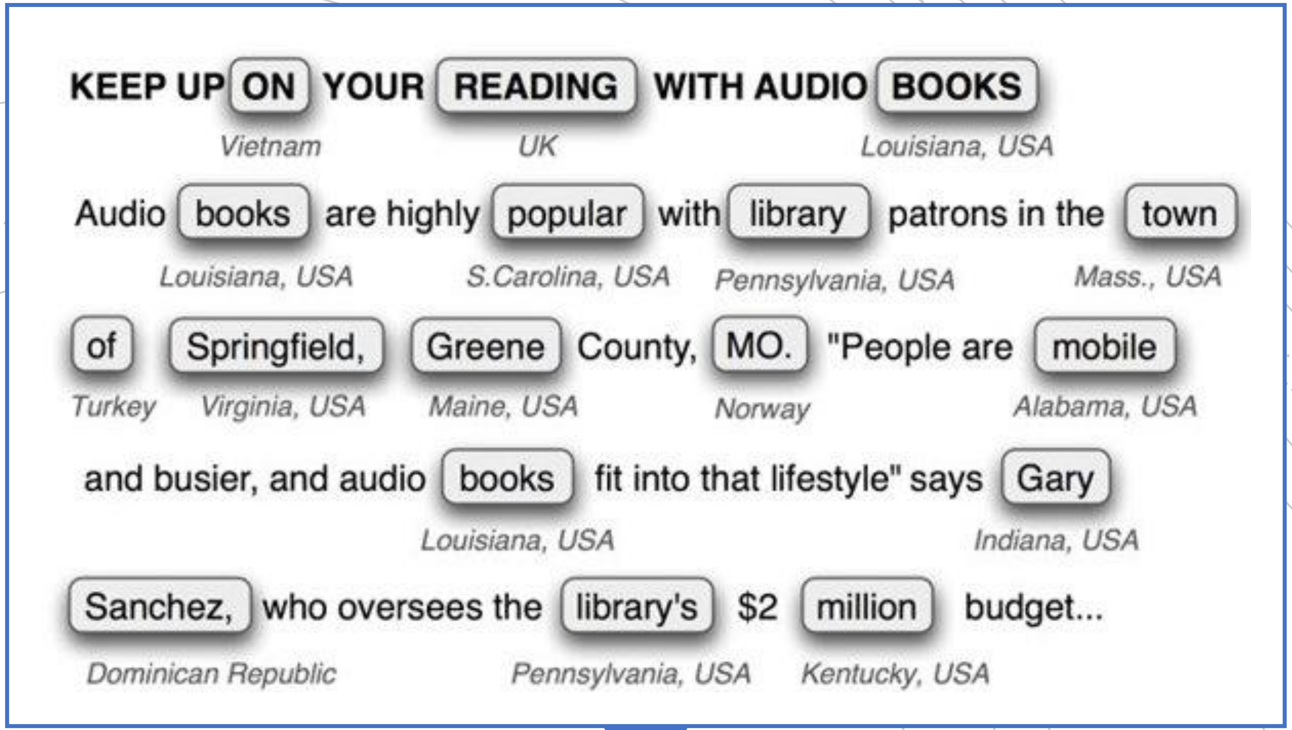
In this study, we observe that granulocyte signatures in the multiple myeloma tumor microenvironment contribute to a more accurate prognosis. This implies that future researchers and clinicians treating patients should quantify tumor microenvironment components, in particular monocytes and granulocytes, which are often ignored in microenvironment studies.

Polymorphonuclear | leukocytes | plasma cancer

And more:

- Housing ads
- Police reports
- Reviews
- News
- Chatbots

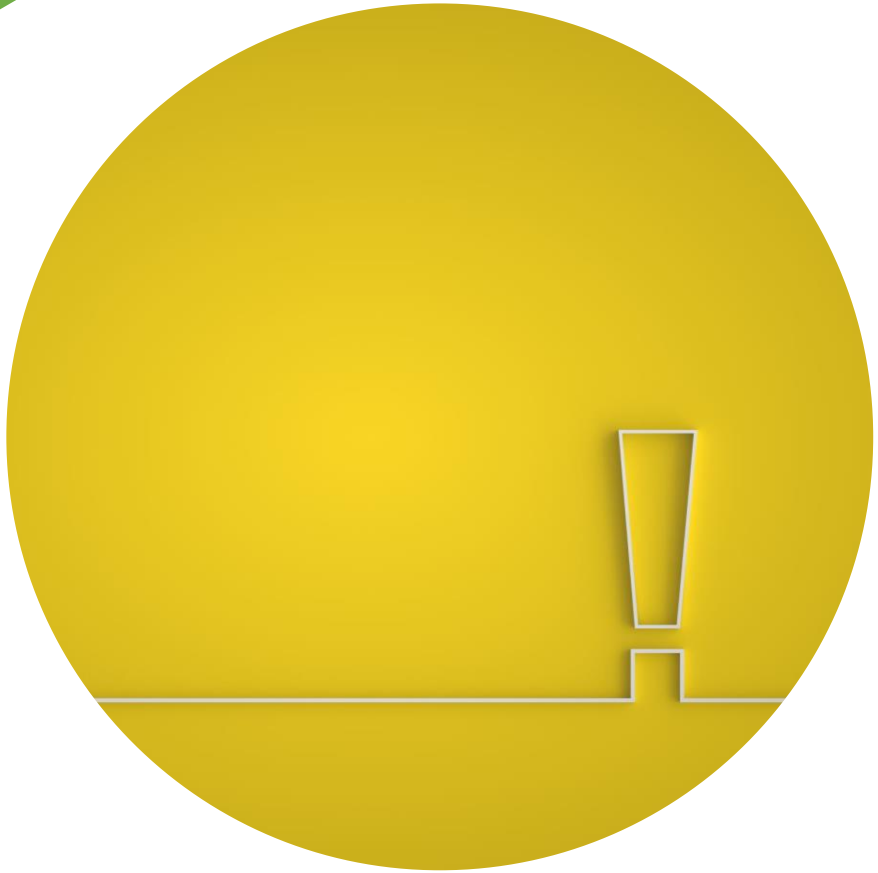
- … Unstructured → Structured



NER is more than a dictionary look-up...

How is it trained?

- Rule-based
- Space classification
- Subsequence classification (semi-CRF)
- Sequence Segmentation as Sequence Labeling (BIO)



CoNNL format

All data files contain one word per line with empty lines representing sentence boundaries. At the end of each line there is a tag which states whether the current word is inside a named entity or not. The tag also encodes the type of named entity. Here is an example sentence:

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

* **CoNLL**: Conference on Computational Natural Language Learning

BIO schema

John	B–PER
Smith	I–PER
lives	O
in	O
New	B–LOC
York	I–LOC

John Smith ⇒ PERSON
New York ⇒ LOCATION

John Snow LABS

NER – CoNNL labeling format: BIO

Some Potential Named Entity Types

- Different annotation schemes for NER use different types
- Common types include:
  - PER—person
  - ORG—Organization
  - LOC—Location
  - GPE—Geopolitical Entity
  - FAC—Facility
  - NAT—Natural phenomenon
- Only tagged when they are *proper names*
- You can also train your own (How?)

Q: How would you evaluate your NER model?

Seq2Seq - BIO

B – Begin

I – Inside

O – Outside

Optional (less used):

S – single

E – End

Paris	**B-PER**
Whitney	**I-PER**
Hilton	**I-PER**
Born	**O**
February	**B-TEMP**
17	**I-TEMP**
1981	**I-TEMP**
…	

Popular Frameworks (partial list)

- spaCy
- Zalando flair
- AllenNLP
- Stanford CoreNLP (CRF-NER)
- BERT

- Train your own:
  - spaCy
  - BERT

Optional Further Reading

- Semi-Markov Conditional Random Fields for IE
- Design Challenges and Misconceptions in NER

Semi-Markov Conditional Random Fields for Information Extraction

Sunita Sarawagi
Indian Institute of Technology
Bombay, India
sunita@iitb.ac.in

William W. Cohen
Center for Automated Learning & Discovery
Carnegie Mellon University
wcohen@cs.cmu.edu

Design Challenges and Misconceptions in Named Entity Recognition[*†‡]

Lev Ratinov **Dan Roth**
Computer Science Department
University of Illinois
Urbana, IL 61801 USA
{ratinov2,danr}@uiuc.edu

Reference Resolution

Unstructured → Structured

- Input: text, empty relational database
- Output: populated relational database

Senator John Edwards is to drop out of the race to become the Democratic party's presidential candidate after consistently trailing in third place. In the latest primary, held in Florida yesterday, Edwards gained only 14% of the vote, with Hillary Clinton polling 50% and Barack Obama on 33%. A reported 1.5m voters turned out to vote.

State	Party	Candidate	Fraction
FL	D	Edwards	0.14
FL	D	Clinton	0.50
FL	D	Obama	0.33

Named Entity Recognition (NER)

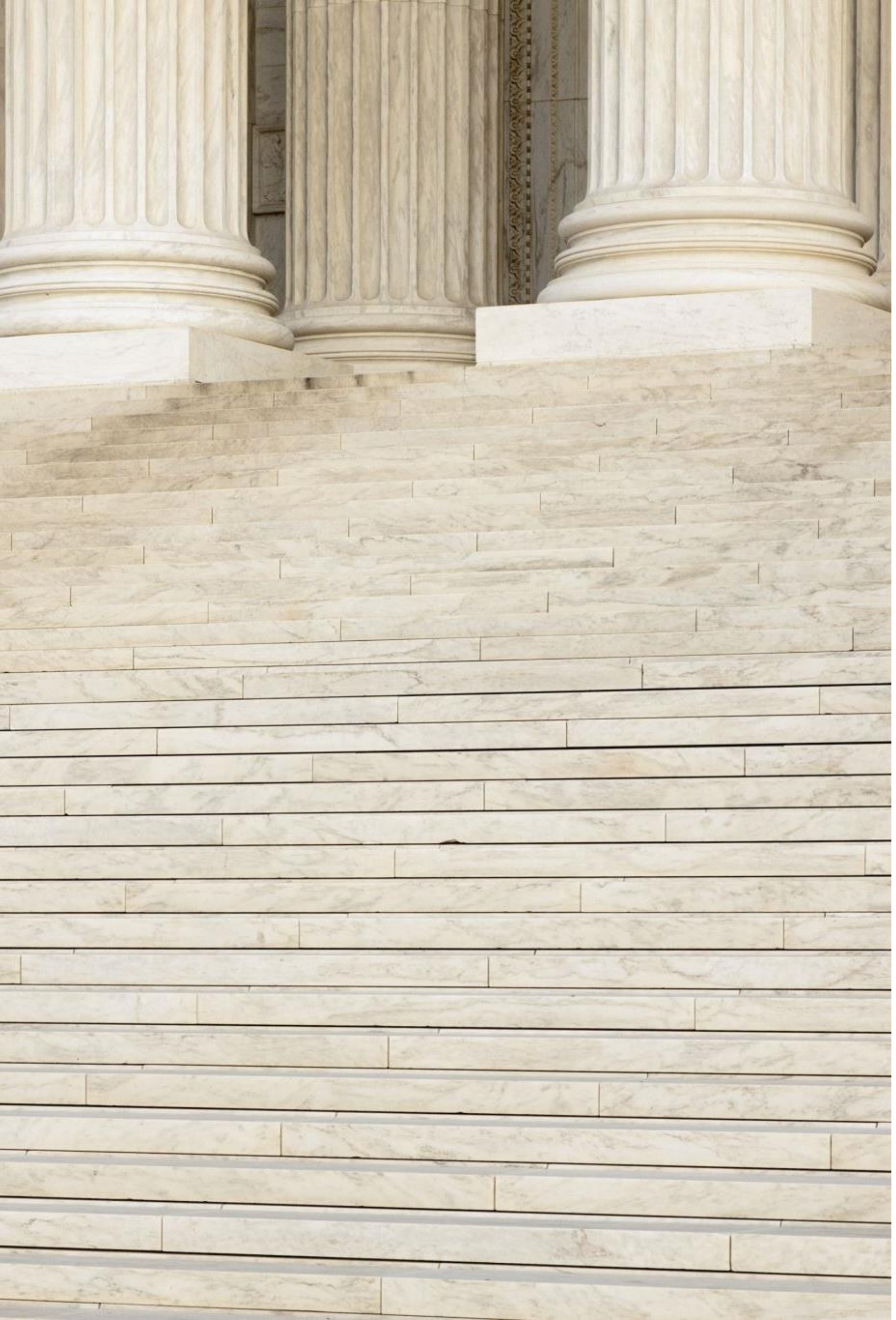
Senator John Edwards is to drop out of the race to become the Democratic party's presidential candidate after consistently trailing in third place. In the latest primary, held in Florida yesterday, Edwards gained only 14% of the vote, with Hillary Clinton polling 50% and Barack Obama on 33%. A reported 1.5m voters turned out to vote.

Named Entity Recognition (NER)

Senator John Edwards is to drop out of the race to become the **Democratic party**'s presidential candidate after consistently trailing in third place. In the latest primary, held in **Florida** yesterday, **Edwards** gained only 14% of the vote, with **Hillary Clinton** polling 50% and **Barack Obama** on 33%. A reported 1.5m voters turned out to vote.

Named Entity Recognition (NER)

Senator John Edwards is to drop out of the race to become the **Democratic party**'s presidential candidate after consistently trailing in third place. In the latest primary, held in **Florida** yesterday, **Edwards** gained only 14% of the vote, with **Hillary Clinton** polling 50% and **Barack Obama** on 33%. A reported 1.5m voters turned out to vote.



Coreference vs Anaphora

Anaphora:

Senator John Edwards …. He ….

Coreference:

Senator John Edwards …. Edwards … The Senator …

Not an easy task…

- Every dancer twisted her knee.
- No dancer twisted her knee

Anaphora, but not coreferential…

We went to see a concert last night.
The tickets were very expensive.

Bridging Anaphora

Coreference Resolution

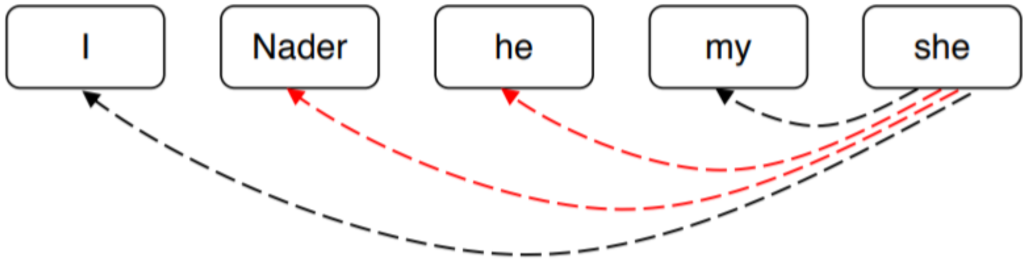
1. Detect candidate mentions (easy)
    1. POS
    2. NER
    3. Syntactic Parser (NP)

2. Cluster the mentions (hard...)
    1. Rule based
    2. Classify pairs (what classifier is it?)
Mention ranking (what classifier is it?)
    3. Clustering pairs
    4. End2End model

Mention Pair

- Train a binary classifier that assigns every pair of mentions a probability of being co-referent: $p(m_i, m_j)$
  - E.g., for “she” look at all **candidate antecedents** (previously occurring mentions) and decide which are co-referent with it.

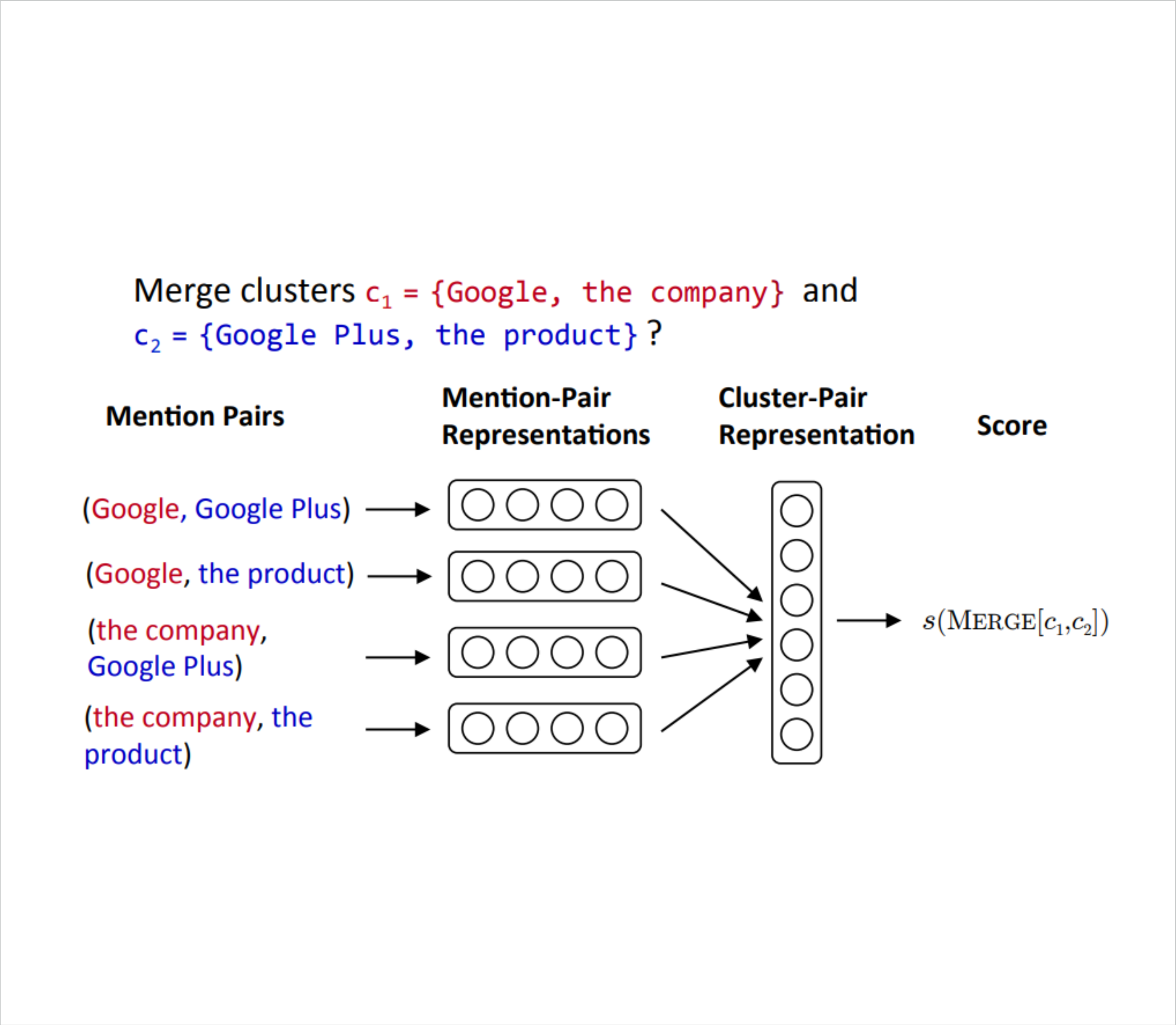
“I voted for Nader because he was most aligned with my values,” she said.

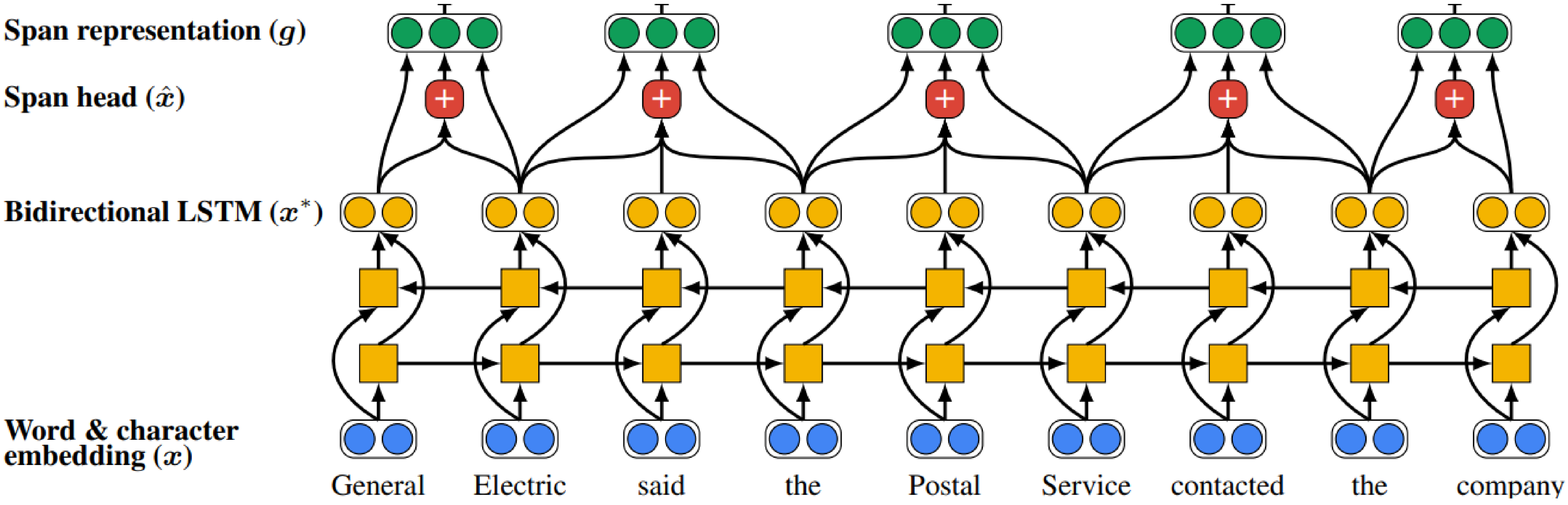


Negative examples: make $p(m_i, m_j)$ to be near 0.

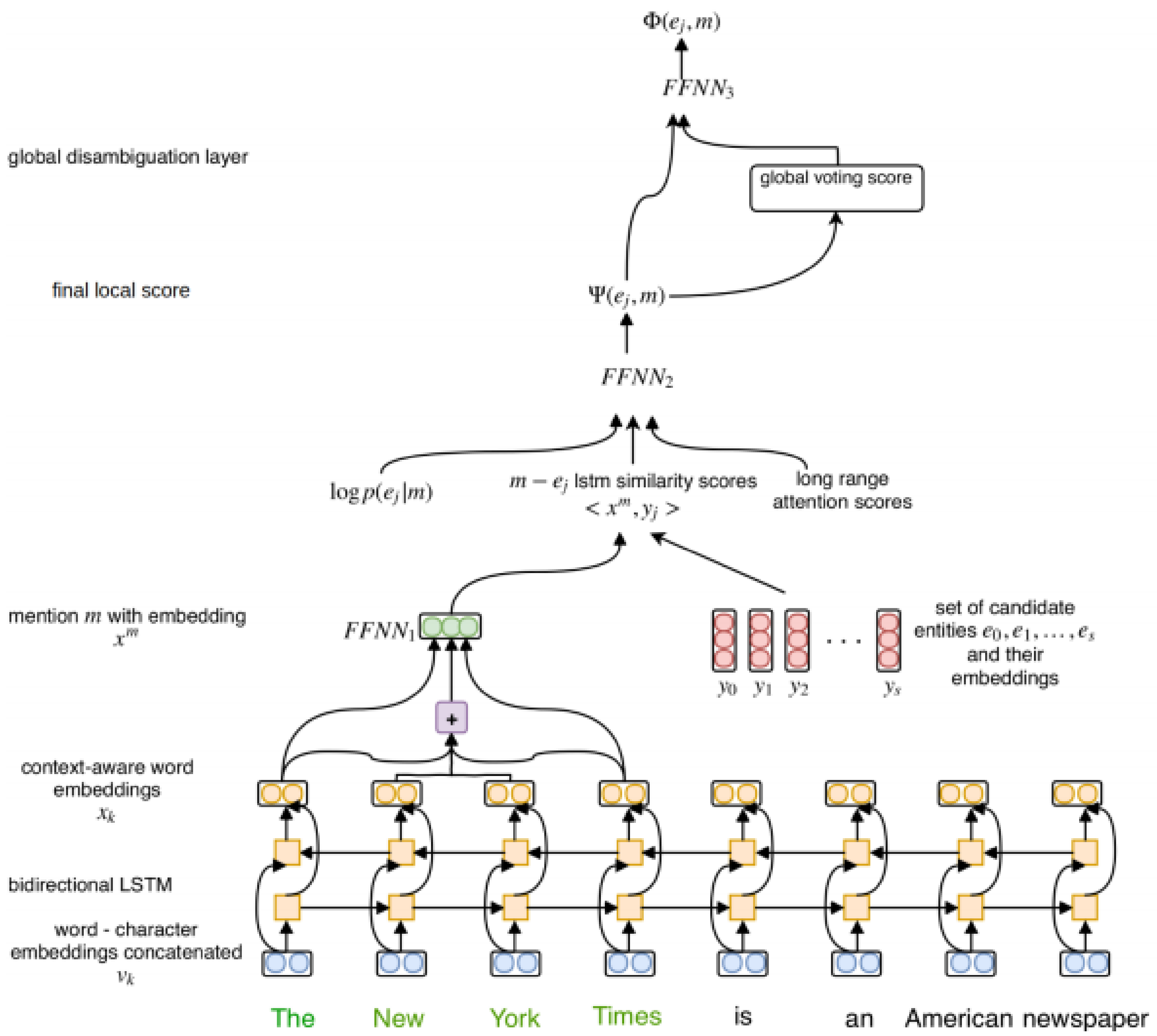
Clustering Pairs

- First, cluster the mentions
- Then, train a classifier to merge pairs if they are indeed referencing the same entity.
- Intuition:
  easy decisions first; hard decisions later.





[End-to-end Neural Coreference Resolution]


$\Phi(e_j, m)$
$FFNN_3$
global disambiguation layer
global voting score
final local score
$\Psi(e_j, m)$
$FFNN_2$
$\log p(e_j|m)$
$m - e_j$ lstm similarity scores
$< x^m, y_j >$
long range attention scores
mention m with embedding x^m
$FFNN_1$
set of candidate entities $e_0, e_1, \dots, e_s$ and their embeddings
y_0 y_1 y_2 y_s
context-aware word embeddings x_k
bidirectional LSTM
word - character embeddings concatenated v_k
The New York Times is an American newspaper

Additional References

- Mention detection in coreference resolution: survey | SpringerLink
- quadrama/gerdracor-coref: German Drama Corpus for Coreference (github.com)
- LILLIE: Information extraction and database integration using linguistics and learning-based algorithms – ScienceDirect
- [2009.08153] End-to-End Neural Event Coreference Resolution (arxiv.org)
- [1606.01323] Improving Coreference Resolution by Learning Entity-Level Distributed Representations (arxiv.org)

F1@MA F1@MI	AIDA A	AIDA B	MSNBC	OKE-2015	OKE-2016	N3-Reuters-12	N3-RSS-500	Derczynski	KORE50
FREME	23.6	23.8	15.8	26.1	22.7	26.8	32.5	31.4	12.3
	37.6	36.3	19.9	31.6	28.5	30.9	27.8	18.9	14.5
FOX	54.7	58.1	11.2	53.9	49.5	52.4	35.1	42.0	28.3
	58.0	57.0	8.3	56.8	50.5	53.3	33.8	38.0	30.8
Babelfy	41.2	42.4	36.6	39.3	37.8	19.6	32.1	28.9	52.5
	47.2	48.5	39.7	41.9	37.7	23.0	29.1	29.8	55.9
Entityclassifier.eu	43.0	42.9	41.4	29.2	33.8	24.7	23.1	16.3	25.2
	44.7	45.0	42.2	29.5	32.5	27.9	22.7	16.9	28.0
Kea	36.8	39.0	30.6	44.6	46.3	17.5	22.7	31.3	41.0
	40.4	42.3	30.9	46.2	46.4	18.1	20.5	26.5	46.8
DBpedia Spotlight	49.9	52.0	42.4	42.0	41.4	21.5	26.7	33.7	29.4
	55.2	57.8	40.6	44.4	43.1	24.8	27.2	32.2	34.9
AIDA	68.8	71.9	62.7	58.7	0.0	42.6	42.6	40.6	49.6
	72.4	72.8	65.1	63.1	0.0	46.4	42.4	32.6	55.4
WAT	69.2	70.8	62.6	53.2	51.8	45.0	45.3	44.4	37.3
	72.8	73.0	64.5	56.4	53.9	49.2	42.3	38.0	49.6
Best baseline	**69.2**	**71.9**	**62.7**	**58.7**	**51.8**	**52.4**	**45.3**	**44.4**	**52.5**
	72.8	**73.0**	**65.1**	**63.1**	**53.9**	**53.3**	**42.4**	**38.0**	**55.9**
base model	86.6	81.1	64.5	54.3	43.6	47.7	44.2	43.5	34.9
	89.1	80.5	65.7	58.2	46.0	49.0	38.8	38.1	42.0
base model + att	86.5	81.9	69.4	56.6	49.2	48.3	46.0	47.9	36.0
	88.9	82.3	69.5	60.7	51.6	51.1	40.5	42.3	42.2
base model + att + global	86.6	82.6	73.0	56.6	47.8	45.4	43.8	43.2	26.2
	89.4	82.4	72.4	61.9	52.7	50.3	38.2	34.1	35.2
ED base model + att + global using Stanford NER mentions	75.7	73.3	71.1	62.9	57.1	54.2	45.9	48.8	40.3
	80.3	74.6	71.0	66.9	58.4	54.6	42.2	42.3	46.0

Still an open task…

Model	English	Chinese
Lee et al. (2010)	~55	~50
Chen & Ng (2012) [CoNLL 2012 Chinese winner]	54.5	57.6
Fernandes (2012) [CoNLL 2012 English winner]	60.7	51.6
Wiseman et al. (2015)	63.3	—
Clark & Manning (2016)	65.4	63.7
Lee et al. (2017)	67.2	--

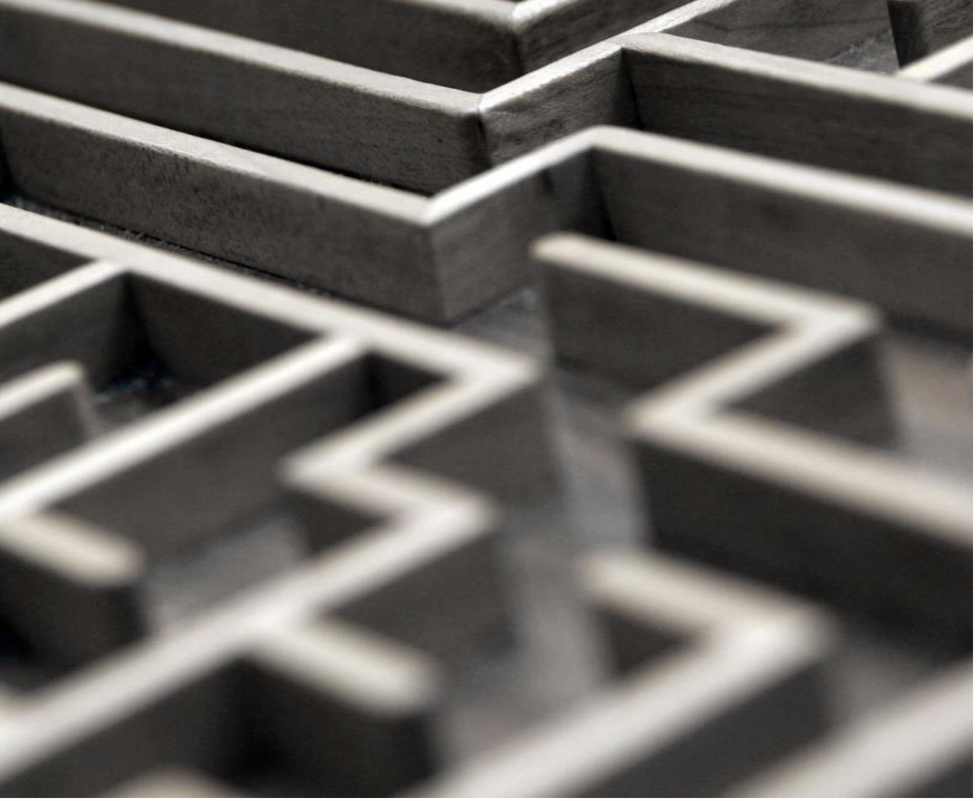
Rule-based system, used to be state-of-the-art!

Non-neural machine learning models

Neural mention ranker

Neural clustering model

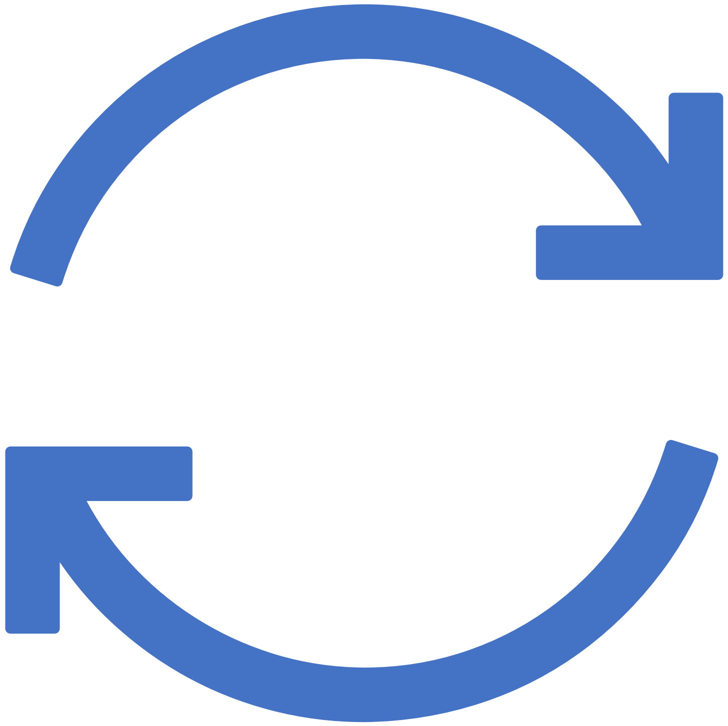
End-to-end neural mention ranker



To sum up

- Coreference is useful but challenging.
- Systems are getting better
  - But the results are still not amazing…
- Try it out:
  - http://corenlp.run/
  - https://huggingface.co/coref/

Relation Extraction



Relation Detection

Senator John Edwards is to drop out of the race to become the **Democratic party'**s presidential candidate after consistently trailing in third place. In the latest primary, held in **Florida** yesterday, **Edwards** gained only 14% of the vote, with **Hillary Clinton** polling 50% and **Barack Obama** on 33%. A reported 1.5m voters turned out to vote.

member_of	
John Edwards	Democratic Party
Hilary Clinton	Democratic Party
Barack Obama	Democratic Party

Relation Extraction Examples

Relations	Types	Examples
Physical-Located	PER-LOC	**He** was in **Tennessee**
Part-Whole-Subsidiary	ORG-ORG	**Alphabet**, the parent company of **Google**
Person-Social-Family	PER-PER	**Yoko**’s husband **John**
Org-AFF-Founder	PER-ORG	**Steve Jobs**, co-founder of **Apple**...
Proximity	LOC-LOC	**Vienna** is not far from **Bratislava**
Geopositional	PER-LOC	**Mozart** was born in **Salzburg**

Seeding Tuples

- Examples
  - Brad is married to Angelina.
  - Bill is married to Hillary.
  - Hillary is married to Bill.
  - Hillary is the wife of Bill.

- Seeds (templates)
  - {PER X} is married to {PER Y}
  - {PER X} is the wife of {PER Y}

Find all X,Y where these templates exist

Extract the X,Ys, find all other X … Y

Seeding Tuples

Bill Gates, the CEO of Microsoft, said …

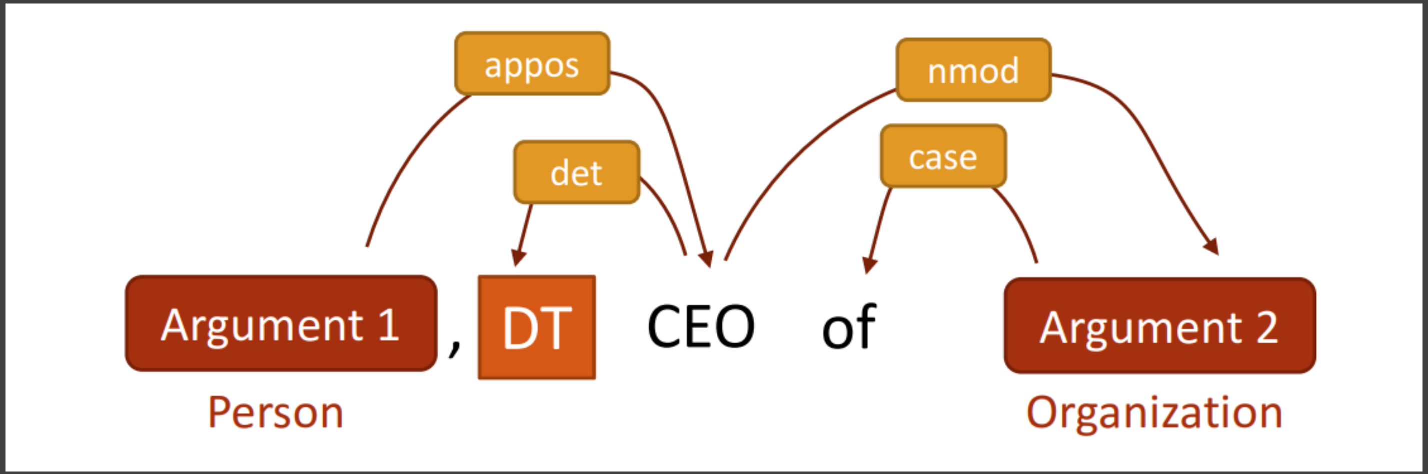
Mr. Jobs, the brilliant and charming CEO of Apple Inc., said …

… announced by Steve Jobs, the CEO of Apple.

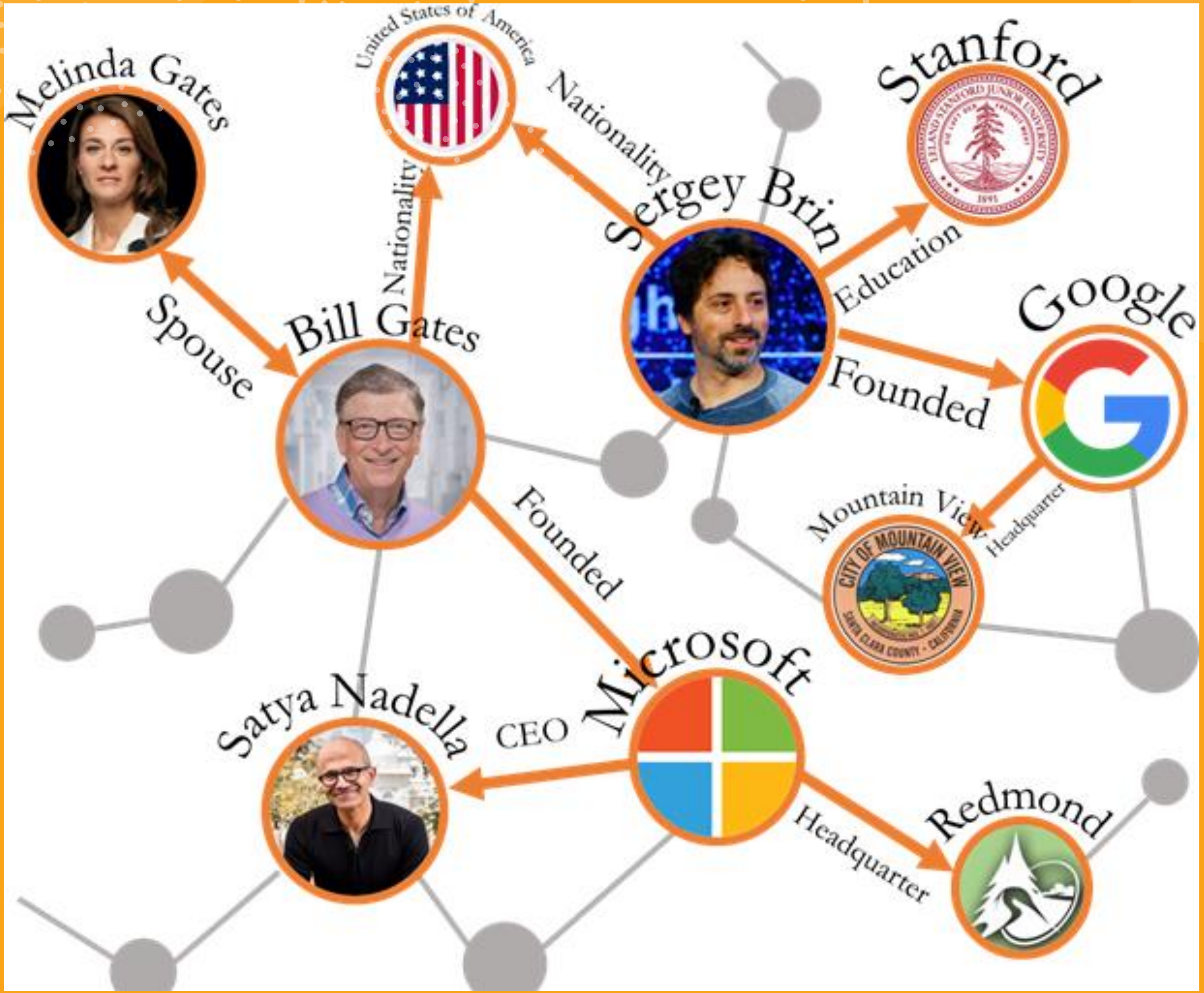
… announced by Bill Gates, the director and CEO of Microsoft.

… mused Bill, a former CEO of Microsoft. and many other possible instantiations…

[PN PER], [DT] [ADJ POS] of [PN ORG]

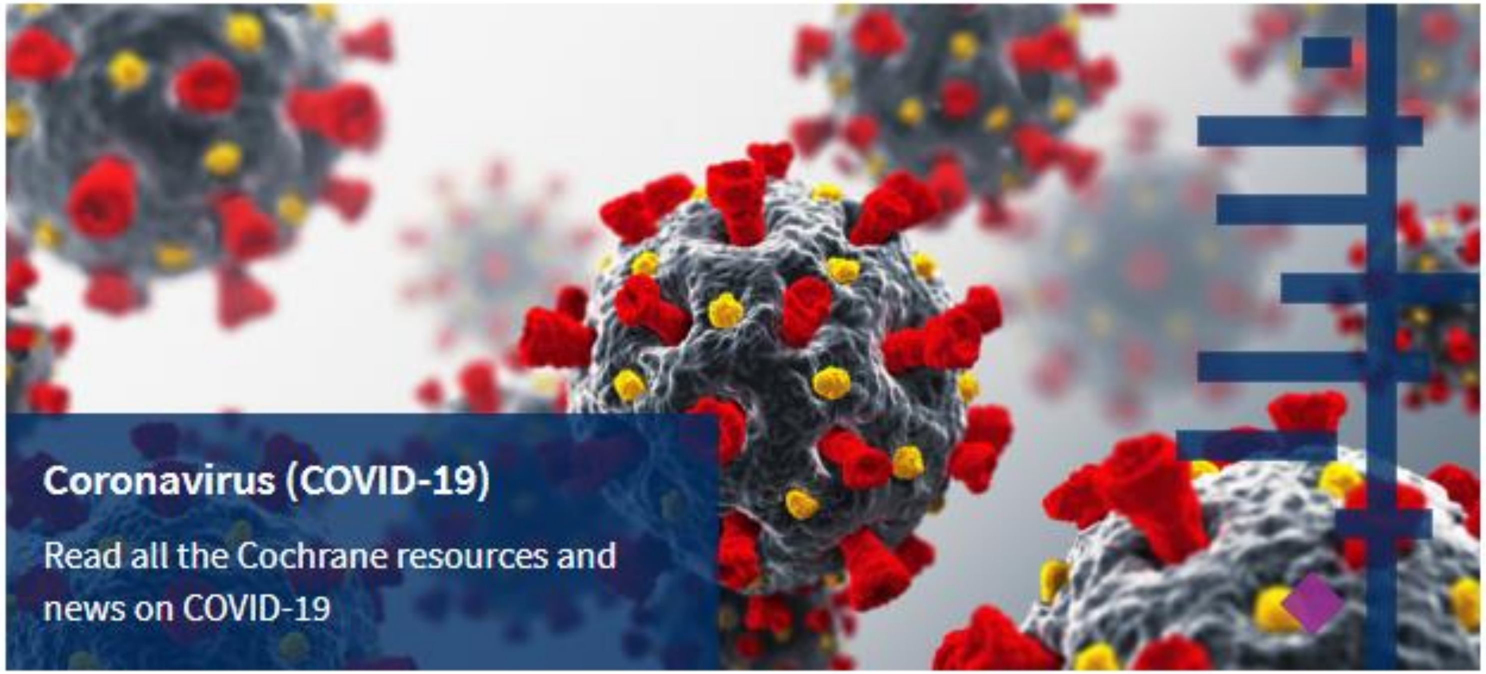


Knowledge Graph



Knowledge Graph

- Nodes = Entities
  - People, organizations, locations, occupations...
  - Protein names, medications, side effects
- Edges = Extracted relations
  - VP, made of NP of NP: father_of, born_at, founder_of,



Our next strategy: let's collaborate

What is Cochrane evidence and how can it help you?

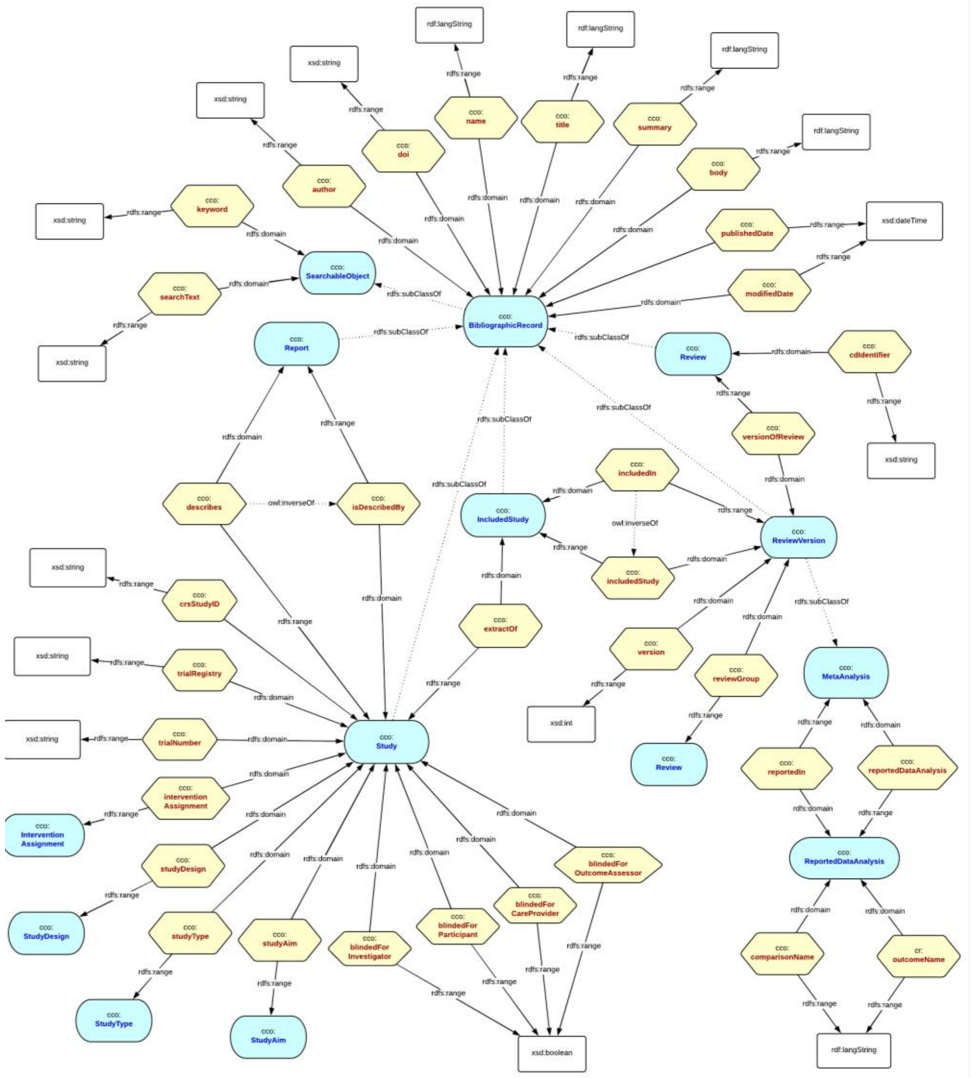
Latest News and Events

Cochrane Clinical

Prof Tracey Howe

Latest Cochrane

Top 10

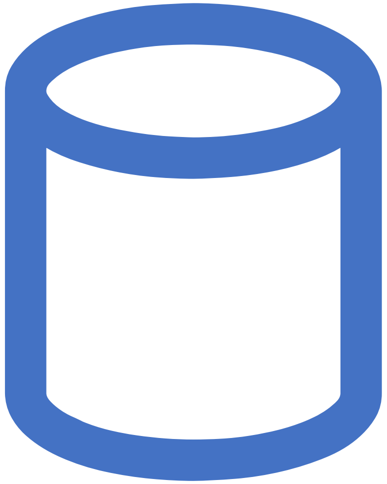


Cochrane respond to COVID-19

https://datalanguage.com/blog/how-knowledge-graph-technology-is-helping-cochrane-respond-to-covid-19

Entity Linking

Given an entity, find the matching reference in the KB



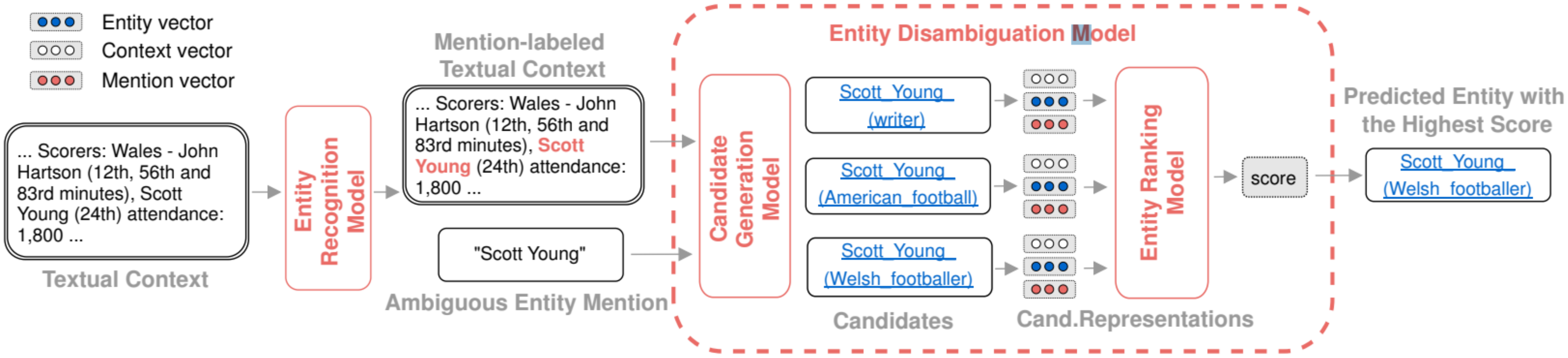
Entity Names

- Different Entities with the same Name
  - Springfield, Kevin Smith
  - Amazon, Paris
- Different Names for the same Entity
  - First names, Location as team name (sport), Nick names
  - Typos/Misspellings: Baarak, Barak, Barrack
  - Inconsistent References: MSFT, APPL, GOOG

Entity Linking

- Candidate Ranking
- Pair-wise Binary Classifier (Unlinkable Mention Prediction)
- End-to-End

Entity Linking with Disambiguation



Entity Linking - Disambiguation

Washington drops 10 points after game with UCLA Bruins.

Entity Linking

Washington drops 10 points after game with UCLA Bruins.

- Fetch Candidates:
  - Washington DC
  - George Washington
  - Washington state
  - Lake Washington
  - Washington Huskies
  - Denzel Washington
  - University of Washington
  - Washington High School …

Entity Linking

Washington drops 10 points after game with UCLA Bruins.

- Fetch Candidates:
- Filter by Entity Type – LOC | ORG:
  - Washington DC
  - ~~George Washington~~
  - Washington state
  - Lake Washington
  - Washington Huskies
  - ~~Denzel Washington~~
  - University of Washington
  - Washington High School …

Entity Linking

Washington drops 10 points after game with UCLA Bruins.

- Fetch Candidates
- Filter by Entity Type – LOC | ORG
- Coreference
    - ~~Washington DC~~
    - ~~George Washington~~
    - ~~Washington state~~
    - ~~Lake Washington~~
    - Washington Huskies
    - ~~Denzel Washington~~
    - University of Washington
    - ~~Washington High School~~

Entity Linking

Washington drops 10 points after game with UCLA Bruins.

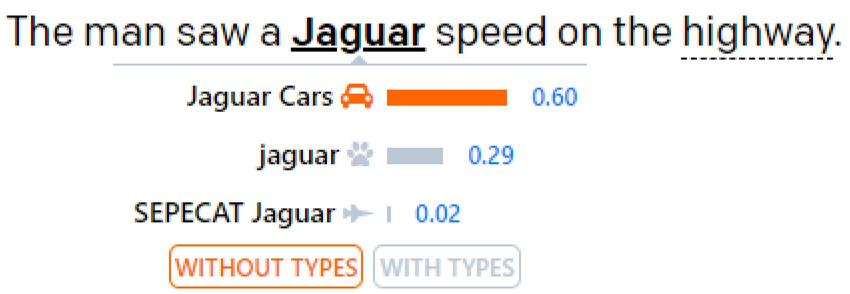
- Fetch Candidates
- Filter by Entity Type – LOC | ORG
- Coreference
- Coherence (UCLA)
  - ~~Washington DC~~
  - ~~George Washington~~
  - ~~Washington state~~
  - ~~Lake Washington~~
  - Washington Huskies
  - ~~Denzel Washington~~
  - ~~University of Washington~~
  - ~~Washington High School~~

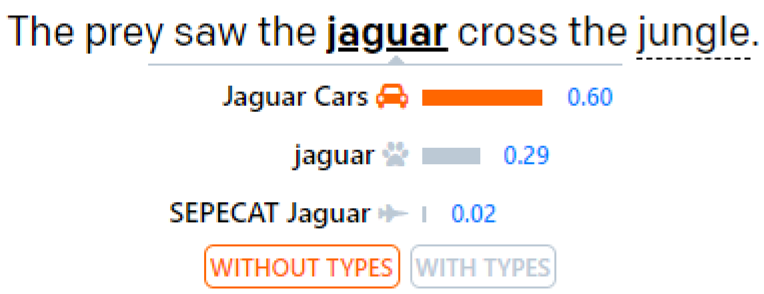
Entity Linking – End2End

The prey saw the **jaguar** cross the jungle

DeepType

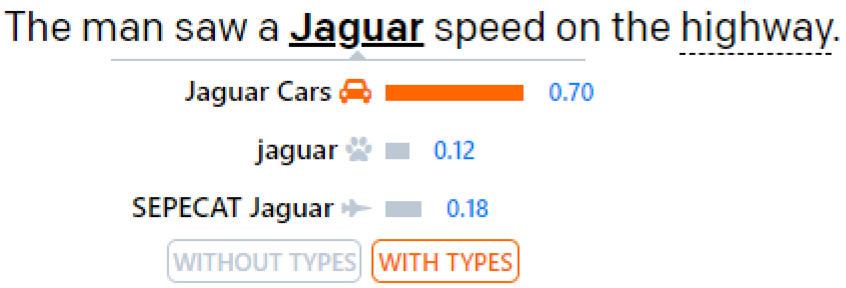
- Multilingual Entity Linking through Neural Type System Evolution (openAI)
- Clustering DB entities into types

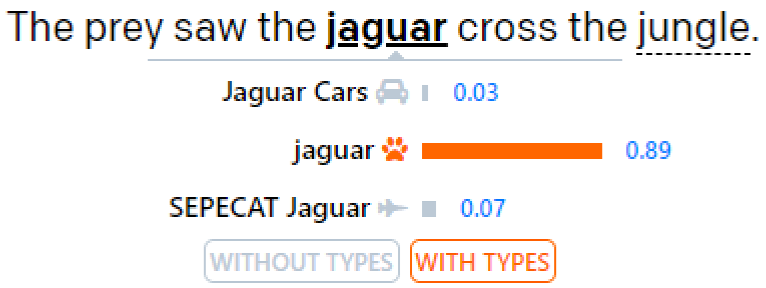




DeepType

- Pick potential categories (~100)
- Classify match of the word + context to category (NN)





DeepType

DeepType: Multilingual Entity Linking by Neural Type System Evolution

Jonathan Raiman
OpenAI
San Francisco, California
raiman@openai.com

Olivier Raiman
Agilience
Paris, France
or@agilience.com

SotA: https://openai.com/blog/discovering-types-for-entity-disambiguation/

MicroPrecision: 94.88

References

- [Named Entity Recognition for Entity Linking: What Works and What's Next (aclanthology.org)]