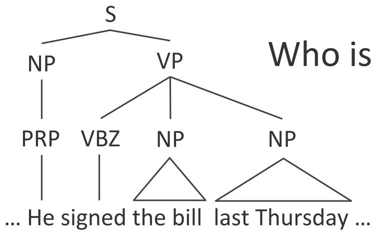
Natural Language Processing

Berkeley
N L P

Coreference Resolution and Entity Linking
UC Berkeley

Sentence-level Analysis



Who is he?

$\exists$ *e. sign(e, he, bill) & date(e, last Thursday)*

Document-level Analysis

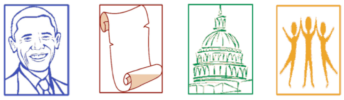
President Barack Obama received the Serve America Act after Congress's vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

Document-level Analysis

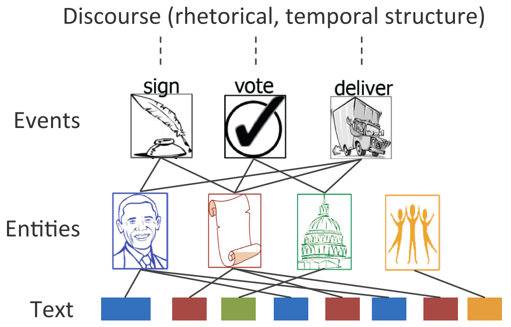
President Barack Obama received the Serve America Act after Congress's vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

Document-level Analysis

President Barack Obama received the Serve America Act after Congress's vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.



Narrative Structure



Entity Analysis

President Barack Obama received the Serve America Act after Congress's vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

Cluster 1	`en.wikipedia.org/wiki/Barack_Obama`
Cluster 2	`.../wiki/Edward_M._Kennedy_Serve_America_Act`
Cluster 3	`.../wiki/United_States_Congress`

Coreference

Input: text (and mentions)

President Barack Obama received the Serve America Act after Congress's vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

Output: clustering of the mentions in text

President Barack Obama received the Serve America Act after Congress's vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

Pragmatics 101

President Barack Obama received the Serve America Act after Congress's vote.

President Barack Obama signed the Serve America Act last Thursday.

President Barack Obama said …

Pragmatics 101

President Barack Obama received the Serve America Act after Congress's vote.

He signed the bill last Thursday.

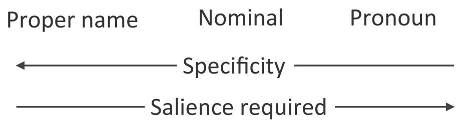
President Barack Obama said …

Pragmatics 101

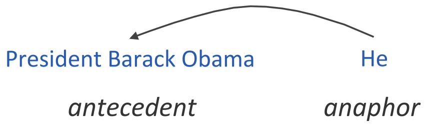
President Barack Obama received the Serve America Act after Congress's vote.

He signed the bill last Thursday.

The president said …



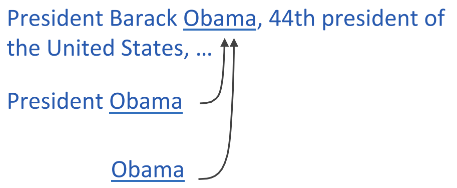
Pragmatics 101



- Coreference is answering the question "who is my antecedent?" for each mention
- Propers, nominals, and pronouns resolve differently!

Proper Names

- Introduce new entities and give information:



President Barack Obama, 44th president of the United States, …

President Obama

Obama

- Main cue: lexical overlap

Pronouns

President Barack Obama received the Serve America Act after Congress's vote. *He* …

President Obama met with Chancellor Merkel. *He* …

President Obama met with President Hollande after *he*…signed the bill. flew to Paris.

- Main cues: agreement, salience

Nominal References

President Obama … The president

Serve America Act … The bill

Barack Obama and Angela Merkel … The leaders

NBC … The network

- Main cues: lexical semantics, world knowledge, salience

What do we need to capture?

- Salience: distance to previous mention
- Semantic compatibility: agreement in number, gender, animacy, semantic type, identity

"A mention refers to the closest compatible antecedent"

- A rule-based system based on this principal won the CoNLL 2011 bakeoff!

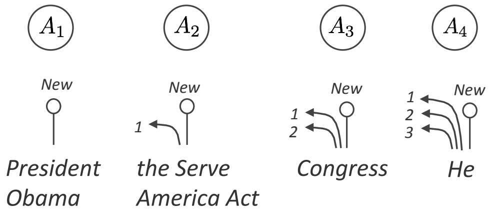
Haghighi and Klein (2009), Raghunathan et al. (2010)

Problem: Robustness

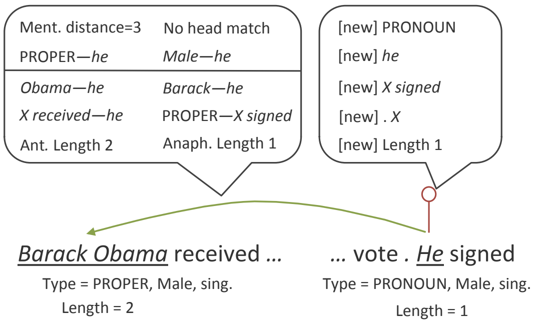
- Number and gender are misidentified
- Generic mentions often don't corefer (*officials*)
- Semantic similarity is a soft concept (sometimes *Washington* and *the US* corefer)
- Even head match is not always reliable (*Gaza Strip* and *Southern Gaza Strip*)

Learning-based Coreference

$$Pr(A_i = a|x) \propto \exp(w^\top f(i,a,x))$$



Features



What else do these capture?

- Centering: progression of mention positions tell us something about discourse status

Barack Obama met with Harry Reid on Monday.
He discussed several key political issues with Reid.
On Tuesday, he announced a new initiative.
~~he~~

  - X *discussed*—X *announced*

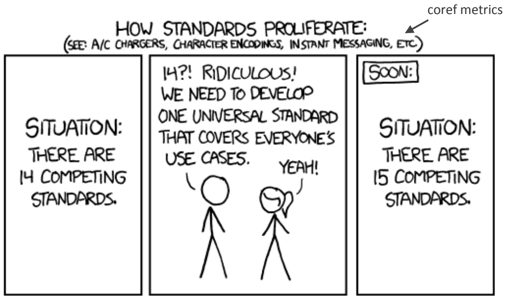
- Definiteness: *the president* is probably a president already in the discourse
  - [new] First word = *the*

Datasets

- OntoNotes dataset: 4000 documents (mix of news, conversations, web) with parses, named entities, coreference
- You have to predict your own entities, and single-mention entities are not annotated

Metrics

coref metrics

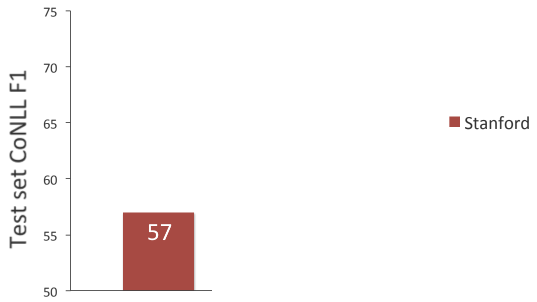

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

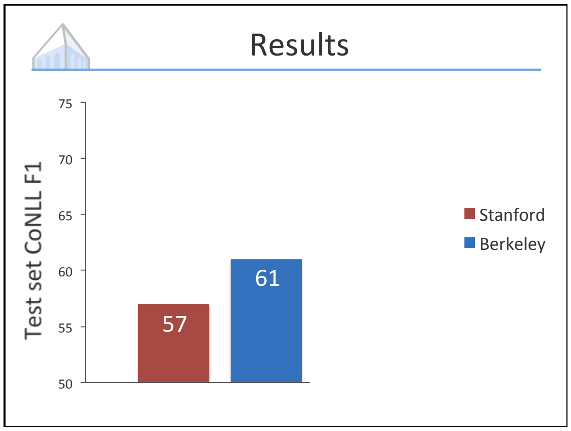
Randall Munroe; http://xkcd.com/927

Metrics

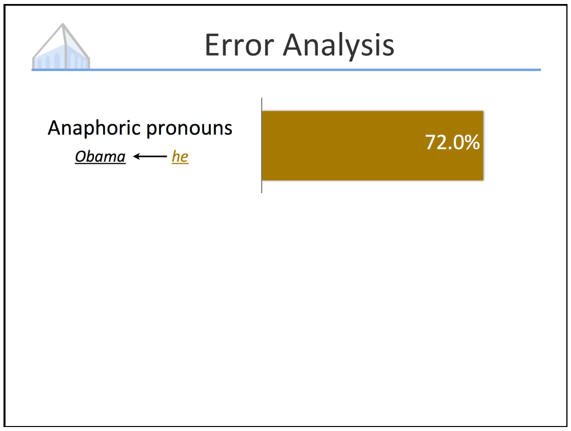
- MUC: "How many antecedents did you get right?" (linear in cluster size)
- B^3: "How many edges in predicted clusters did you get right?" (quadratic in cluster size)
- CEAF: "Do a maximum matching between predicted and gold entities; how close are they?" (???)
- CEAF-M, BLANC, etc.
- CoNLL = (MUC + B^3 + CEAF)/3

Results



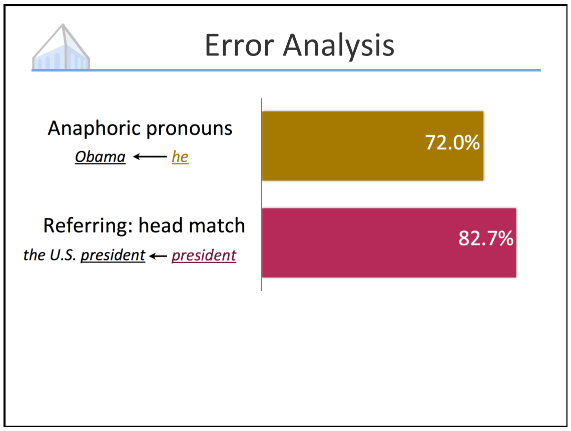


Results



Error Analysis

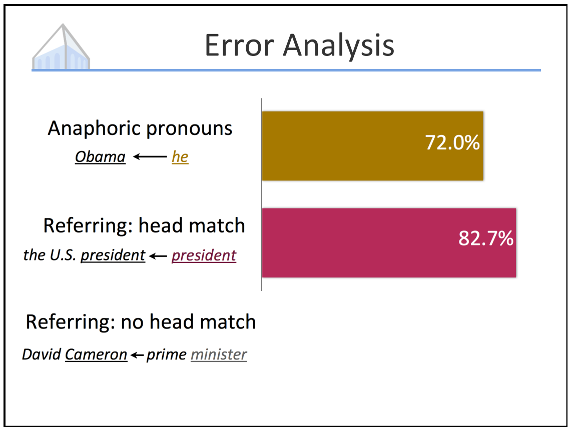
Anaphoric pronouns
Obama ⟵ *he*



Error Analysis

Anaphoric pronouns
Obama ⟵ *he*

Referring: head match
the U.S. president ⟵ *president*

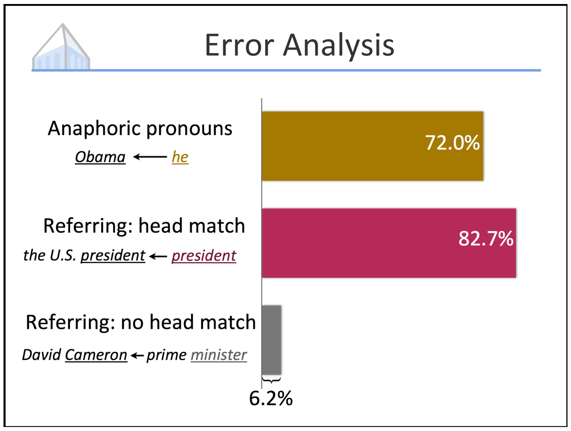


Error Analysis

Anaphoric pronouns
Obama ⟵ *he*

Referring: head match
the U.S. president ⟵ *president*

Referring: no head match
David Cameron ⟵ *prime minister*

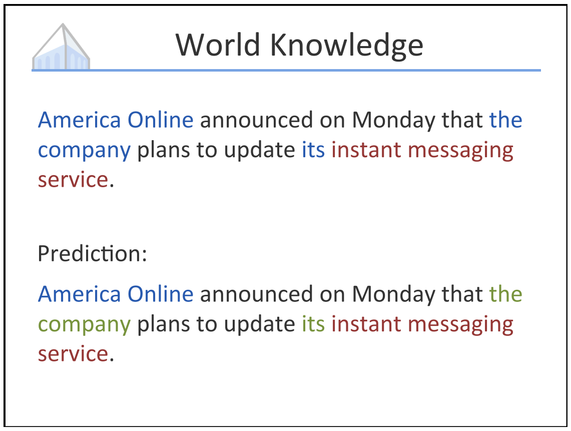


Error Analysis

Anaphoric pronouns
Obama ⟵ *he*

Referring: head match
the U.S. president ⟵ *president*

Referring: no head match
David Cameron ⟵ *prime minister*



World Knowledge

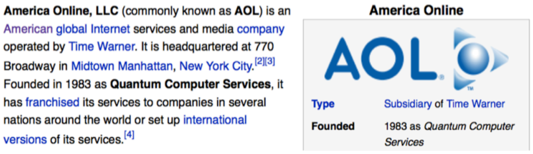
America Online announced on Monday that the company plans to update its instant messaging service.

Prediction:

America Online announced on Monday that the company plans to update its instant messaging service.

World Knowledge

America Online ⟷ **company**



America Online, LLC (commonly known as **AOL**) is an American global Internet services and media company operated by Time Warner. It is headquartered at 770 Broadway in Midtown Manhattan, New York City.[2][3] Founded in 1983 as **Quantum Computer Services**, it has franchised its services to companies in several nations around the world or set up international versions of its services.[4]

America Online

Type	Subsidiary of Time Warner
Founded	1983 as *Quantum Computer Services*

Entity Resolution

Barack Obama en.wikipedia.org/wiki/Barack_Obama

Michael Jordan

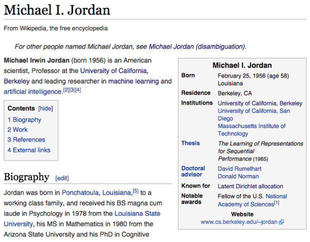
en.wikipedia.org/wiki/Michael_Jordan

en.wikipedia.org/wiki/Michael_I._Jordan

Entity Resolution

- Multiclass decision with 4 million classes
- The outputs are structured objects!

Michael_I._Jordan



Michael I. Jordan

From Wikipedia, the free encyclopedia

For other people named Michael Jordan, see Michael Jordan (disambiguation).

Michael Irwin Jordan (born 1956) is an American scientist, Professor at the University of California, Berkeley and leading researcher in machine learning and artificial intelligence.[2][3][4]

Contents [hide]

1 Biography
2 Work
3 References
4 External links

Biography [edit]

Jordan was born in Ponchatoula, Louisiana,[5] to a working class family, and received his BS magna cum laude in Psychology in 1978 from the Louisiana State University, his MS in Mathematics in 1980 from the Arizona State University and his PhD in Cognitive

Michael I. Jordan

Born	February 25, 1956 (age 58) Louisiana
Residence	Berkeley, CA
Institutions	University of California, Berkeley University of California, San Diego Massachusetts Institute of Technology
Thesis	*The Learning of Representations for Sequential Performance* (1985)
Doctoral advisor	David Rumelhart Donald Norman
Known for	Latent Dirichlet allocation
Notable awards	Fellow of the U.S. National Academy of Sciences[1]

Website
www.cs.berkeley.edu/~jordan

Baseline

Michael Jordan		
	Michael_I._Jordan	0.5
	Michael_Jordan	0.5

Latent Dirichlet allocation

... LDA is an example of a topic model and was first presented as a graphical model for topic discovery by David Blei, Andrew Ng, and Michael Jordan in 2003.[1] ...

Michael_I._Jordan

Basketball

... players who many credit with ushering the professional game to its highest level of popularity: Larry Bird, Earvin "Magic" Johnson, and Michael Jordan. In 2001, the NBA ...

Michael_Jordan

Cucerzan (2007), Milne and Witten (2008)

Choosing the Right Query

professor Michael Jordan

professor Michael Jordan	none
Michael Jordan	Michael_Jordan Michael_I._Jordan
Jordan	Jordan_(country) ...

Durrett and Klein (2014)

Incorporating Context

Michael Jordan gave a talk at the Big Data Bootcamp. The professor covered basic machine learning techniques...

professor: 1
learning: 1
basketball: 0

cosine distance

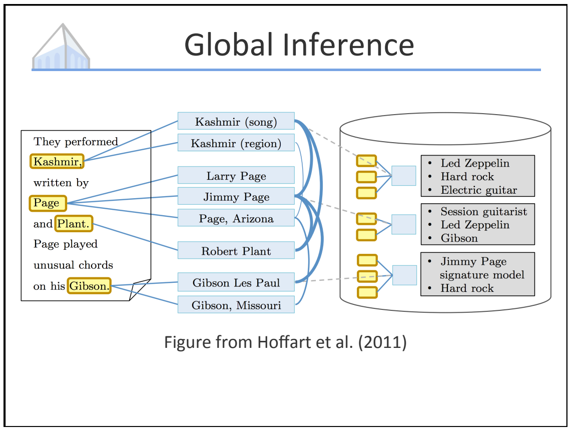
Michael I. Jordan

professor: 12
Bayesian: 5
learning: 10
PhD: 3

Michael Jordan

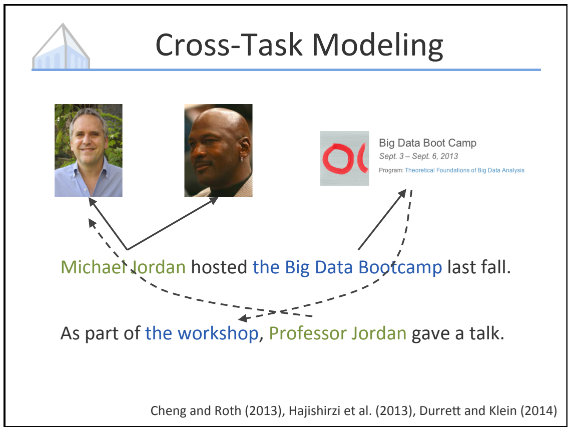
basketball: 50
Bulls: 26
NBA: 30
game: 22

Ratinov et al. (2011)



Global Inference

Figure from Hoffart et al. (2011)

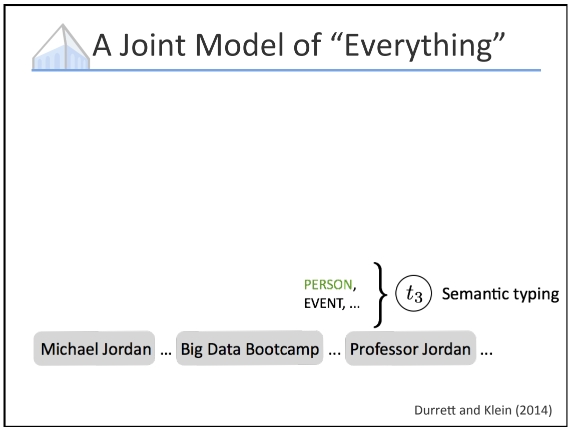


Cross-Task Modeling

Michael Jordan hosted the Big Data Bootcamp last fall.

As part of the workshop, Professor Jordan gave a talk.

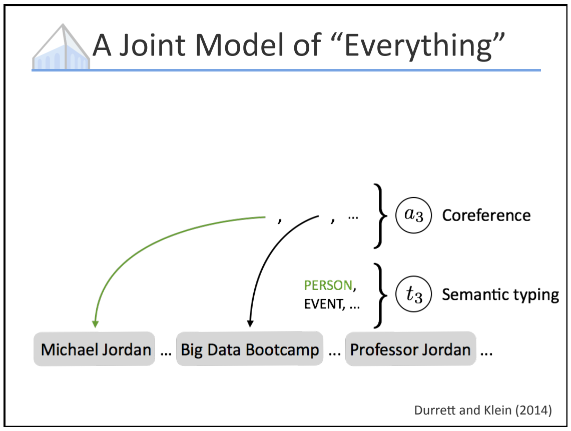
Cheng and Roth (2013), Hajishirzi et al. (2013), Durrett and Klein (2014)



A Joint Model of "Everything"

PERSON,
EVENT, ...
t_3
Semantic typing
Michael Jordan ... Big Data Bootcamp ... Professor Jordan ...

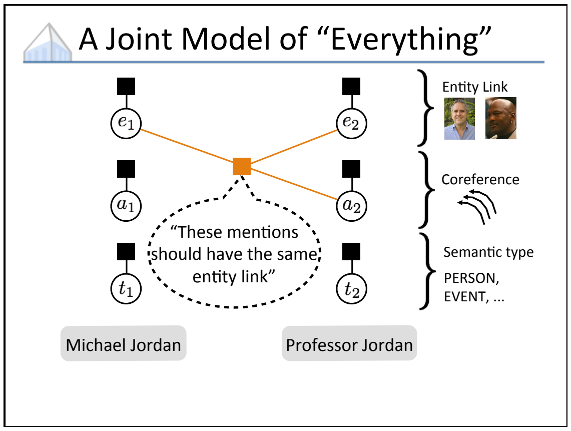
Durrett and Klein (2014)



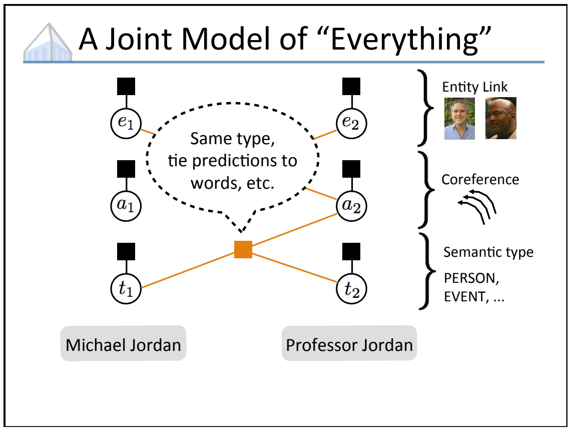
A Joint Model of "Everything"

, , ...
a_3
Coreference
PERSON,
EVENT, ...
t_3
Semantic typing
Michael Jordan ... Big Data Bootcamp ... Professor Jordan ...

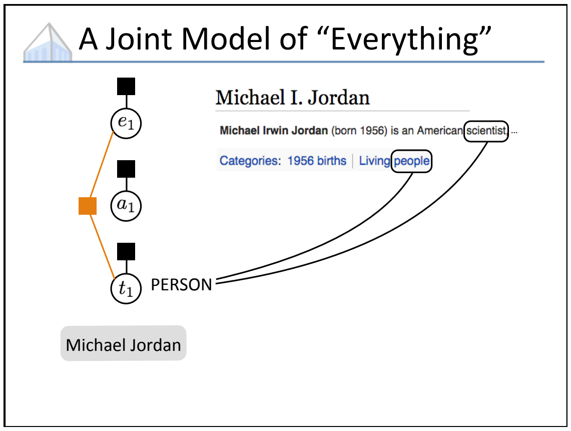
Durrett and Klein (2014)



A Joint Model of "Everything"

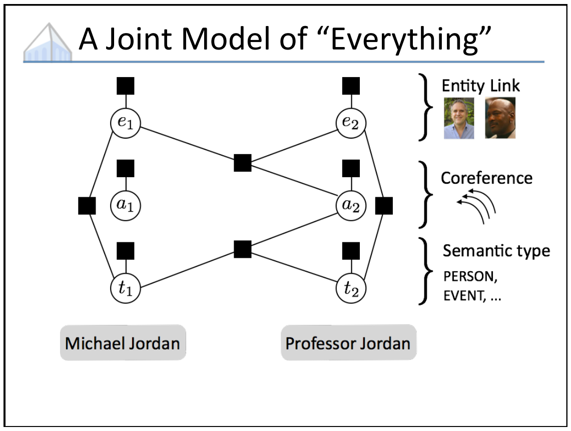


A Joint Model of "Everything"

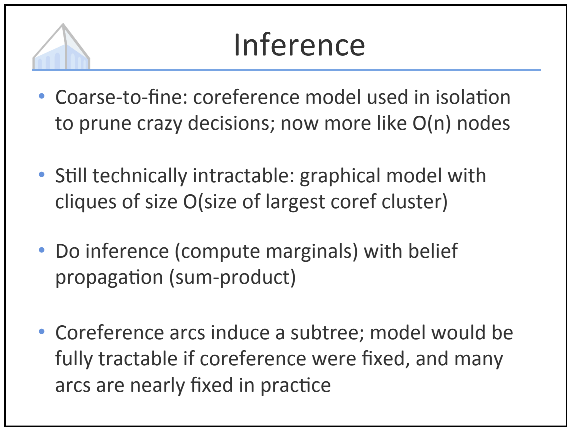


A Joint Model of "Everything"

Michael I. Jordan
Michael Irwin Jordan (born 1956) is an American scientist ...
Categories: 1956 births | Living people
e_1
a_1
t_1
PERSON
Michael Jordan

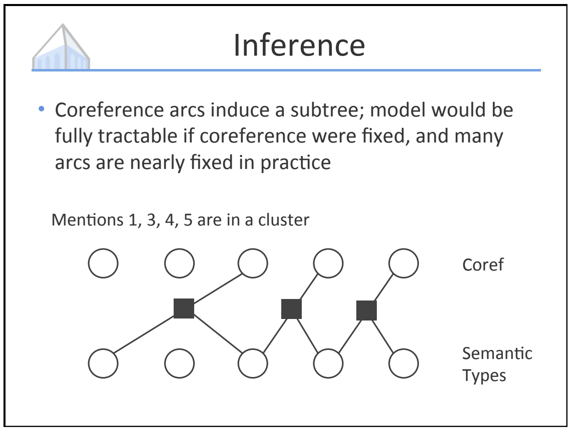


A Joint Model of "Everything"



Inference

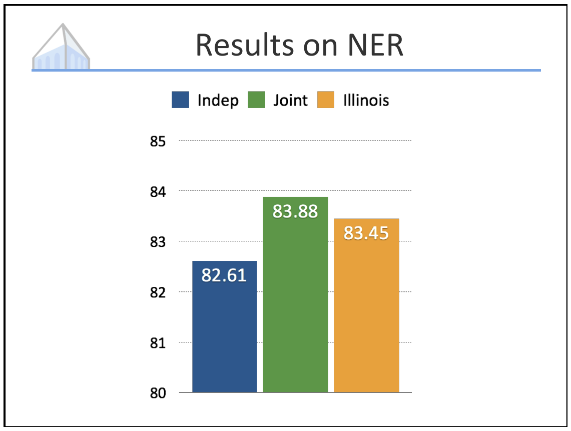
- Coarse-to-fine: coreference model used in isolation to prune crazy decisions; now more like O(n) nodes
- Still technically intractable: graphical model with cliques of size O(size of largest coref cluster)
- Do inference (compute marginals) with belief propagation (sum-product)
- Coreference arcs induce a subtree; model would be fully tractable if coreference were fixed, and many arcs are nearly fixed in practice



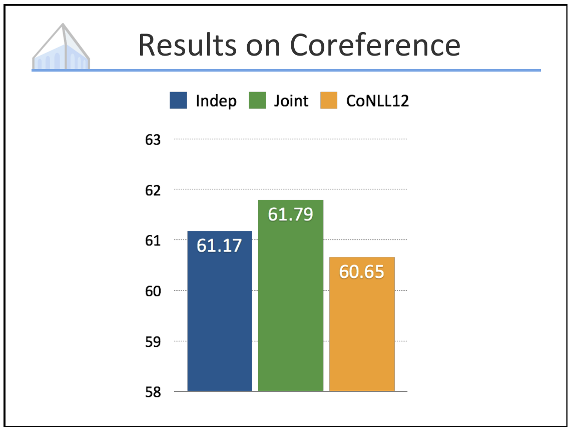
Inference

- Coreference arcs induce a subtree; model would be fully tractable if coreference were fixed, and many arcs are nearly fixed in practice

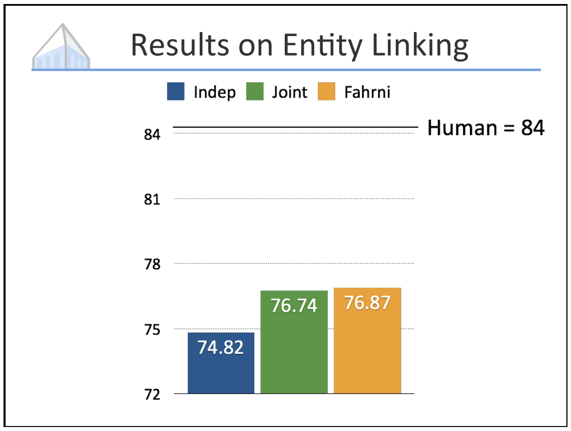
Mentions 1, 3, 4, 5 are in a cluster



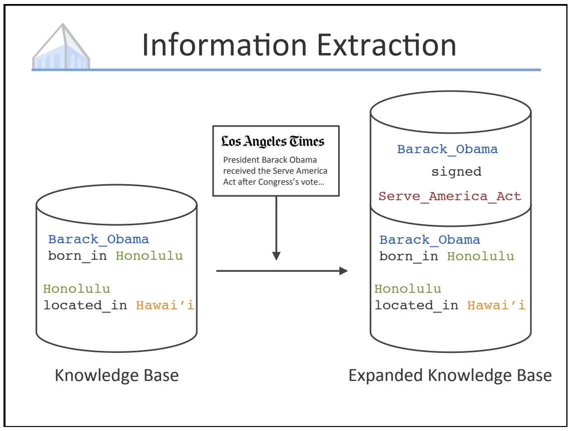
Results on NER



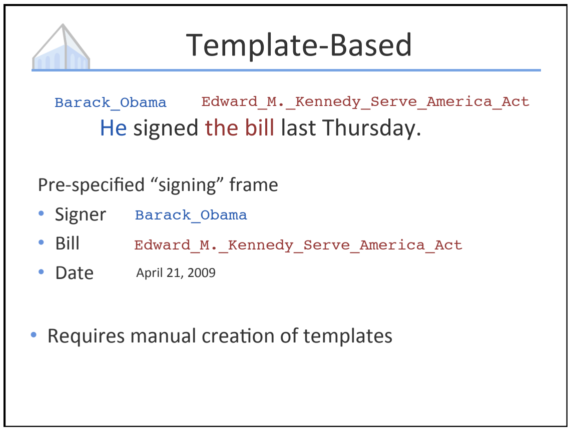
Results on Coreference



Results on Entity Linking



Information Extraction



Template-Based

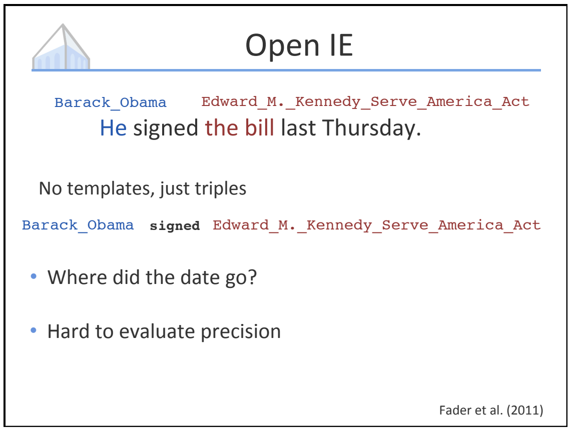
Barack_Obama Edward_M._Kennedy_Serve_America_Act

He signed the bill last Thursday.

Pre-specified "signing" frame

- Signer Barack_Obama
- Bill Edward_M._Kennedy_Serve_America_Act
- Date April 21, 2009

- Requires manual creation of templates



Open IE

Barack_Obama Edward_M._Kennedy_Serve_America_Act

He signed the bill last Thursday.

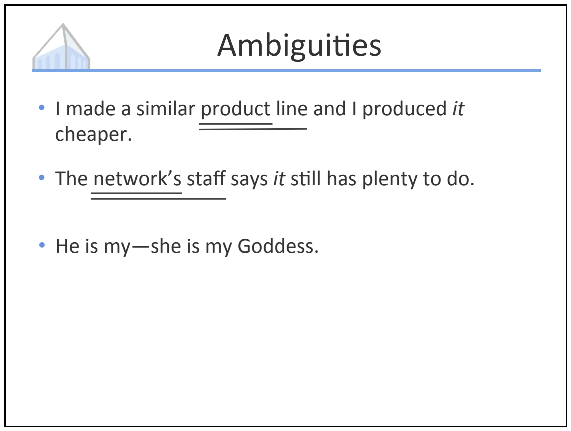
No templates, just triples

Barack_Obama **signed** Edward_M._Kennedy_Serve_America_Act

- Where did the date go?

- Hard to evaluate precision

Fader et al. (2011)



Ambiguities

- I made a similar product line and I produced *it* cheaper.

- The network's staff says *it* still has plenty to do.

- He is my—she is my Goddess.