Natural Language Processing

Berkeley

Coreference Resolution and Entity Linking
UC Berkeley
Who is he?

... He signed the bill last Thursday ...

∃ e. sign(e, he, bill) & date(e, last Thursday)
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.
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.
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

Discourse (rhetorical, temporal structure)

Events
- sign
- vote
- deliver

Entities
- [Image of person]
- [Image of document]
- [Image of building]
- [Image of people]

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.

Cluster 1: en.wikipedia.org/wiki/Barack_Obama

Cluster 2: .../wiki/Edward_M._Kennedy_Serve_America_Act

Cluster 3: .../wiki/United_States_Congress
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.
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 ...
President Barack Obama received the Serve America Act after Congress’s vote. He signed the bill last Thursday.

President Barack Obama said ...
President Barack Obama received the Serve America Act after Congress’s vote.
He signed the bill last Thursday.
The president said ...

Proper name  Nominal  Pronoun

Specificity

Salience required
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, ...

- **Main cue:** lexical overlap

  President **Obama**
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 in 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

\[ Pr(A_i = a | x) \propto \exp(w^T f(i, a, x)) \]

\[ A_1 \quad A_2 \quad A_3 \quad A_4 \]

\[ \text{President Obama} \quad \text{the Serve America Act} \quad \text{Congress} \quad \text{He} \]
### Features

<table>
<thead>
<tr>
<th>Ment. distance=3</th>
<th>No head match</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROPER—<em>he</em></td>
<td>*Male—<em>he</em></td>
</tr>
<tr>
<td>*Obama—<em>he</em></td>
<td>*Barack—<em>he</em></td>
</tr>
<tr>
<td>*X received—<em>he</em></td>
<td>PROPER—<em>X signed</em></td>
</tr>
<tr>
<td>Ant. Length 2</td>
<td>Anaph. Length 1</td>
</tr>
</tbody>
</table>

- **Barack Obama** received ...
  - Type = PROPER, Male, sing.
  - Length = 2

- ... vote. **He** signed
  - Type = PRONOUN, Male, sing.
  - Length = 1
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.

- 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
How standards proliferate:

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)

• B³: “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 + B³ + CEAF)/3
Results

Test set CoNLL F1

Stanford

57
Results

Test set CoNLL F1

- Stanford: 57
- Berkeley: 61
Anaphoric pronouns

\[ \text{Obama} \rightarrow \text{he} \]

72.0%
Error Analysis

Anaphoric pronouns

Obama ← he

72.0%

Referring: head match

the U.S. president ← president

82.7%
Error Analysis

Anaphoric pronouns

Obama ← he

72.0%

Referring: head match

the U.S. president ← president

82.7%

Referring: no head match

David Cameron ← prime minister
Error Analysis

Anaphoric pronouns

Obama ← he

72.0%

Referring: head match

the U.S. president ← president

82.7%

Referring: no head match

David Cameron ← prime minister

6.2%
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.
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.\(^4\)
Entity Resolution

*Barack Obama*  
[en.wikipedia.org/wiki/Barack_Obama](en.wikipedia.org/wiki/Barack_Obama)

*Michael Jordan*  
[en.wikipedia.org/wiki/Michael_Jordan](en.wikipedia.org/wiki/Michael_Jordan)

*Michael I. Jordan*  
[en.wikipedia.org/wiki/Michael_I._Jordan](en.wikipedia.org/wiki/Michael_I._Jordan)
Entity Resolution

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

Michael Jordan

<table>
<thead>
<tr>
<th>Michael_I._Jordan</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael_Jordan</td>
<td>0.5</td>
</tr>
</tbody>
</table>

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.\[1\] ...

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 ...

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

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)

They performed *Kashmir*, written by *Page* and *Plant*. *Page* played unusual chords on his *Gibson*.

- *Kashmir* (song)
- *Kashmir* (region)
- Larry Page
- Jimmy Page
- Page, Arizona
- Robert Plant
- Gibson Les Paul
- Gibson, Missouri

- Led Zeppelin
- Hard rock
- Electric guitar

- Session guitarist
- Led Zeppelin
- Gibson

- Jimmy Page signature model
- Hard rock
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”

Michael Jordan ... Big Data Bootcamp ... Professor Jordan ...

\{t_3\} Semantic typing

Durrett and Klein (2014)
A Joint Model of “Everything”

Michael Jordan, ..., Big Data Bootcamp, ..., Professor Jordan, ...

\( \alpha_3 \) Coreference

\( t_3 \) Semantic typing

Durrett and Klein (2014)
A Joint Model of “Everything”

“These mentions should have the same entity link”

Michael Jordan

Professor Jordan
A Joint Model of “Everything”

Same type, tie predictions to words, etc.

Michael Jordan

Professor Jordan

Entity Link

Coreference

Semantic type

PERSON, EVENT, ...
A Joint Model of “Everything”

Michael I. Jordan

Michael Irwin Jordan (born 1956) is an American scientist...

Categories: 1956 births | Living people

Michael Jordan
A Joint Model of “Everything”

Michael Jordan  Professor Jordan

Entity Link

Coreference

Semantic type
PERSON, EVENT, ...
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

Indep: 82.61
Joint: 83.88
Illinois: 83.45
Results on Coreference

- Indep: 61.17
- Joint: 61.79
- CoNLL12: 60.65
Results on Entity Linking

84

Human = 84

81

78

75

72

Indep Joint Fahrni

74.82 76.74 76.87
**Knowledge Base**

- Barack_Obama
  - born_in Honolulu
- Honolulu
  - located_in Hawai'i

**Los Angeles Times**

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

**Expanded Knowledge Base**

- Barack_Obama
  - signed Serve_America_Act
- Honolulu
  - located_in Hawai'i
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
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.