Natural Language Processing

Berkeley

Coreference Resolution and Entity Linking
UC Berkeley

Sentence-level Analysis

\[
S \\
NP \\
PRP VBZ NP NP \\
\text{... He signed the bill last Thursday ...} \\
\text{e. sign(e, he, bill) \& date(e, last Thursday)}
\]

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

Discourse (rhetorical, temporal structure)

Events

Sign

Vote

Deliver

Entities

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.

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
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
\[ \text{Specificity} \]
\[ \text{Salience required} \]

Coreference

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

• 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 ... unusually.

• 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!

Learning-based Coreference

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

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)
Features

- Ment. distance=3
- No head match
- PROPER—he
- Male—he
- Obama—he
- Barack—he
- PROPER—X signed
- X received—he
- Anaph. Length 1
- [new] PRONOUN
- [new] he
- [new] X signed
- [new] . X
- [new] 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

Metrics

- 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

Results

- MUC: “How many antecedents did you get right?” (linear in cluster size)
- B^2: “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^2 + CEAF)/3

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

Error Analysis

Anaphoric pronouns

<table>
<thead>
<tr>
<th>Obama</th>
<th>he</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>57.0%</td>
</tr>
<tr>
<td>Berkeley</td>
<td>61.0%</td>
</tr>
</tbody>
</table>

Referring: head match

<table>
<thead>
<tr>
<th>the U.S. president</th>
<th>president</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>82.7%</td>
</tr>
<tr>
<td>Berkeley</td>
<td>82.7%</td>
</tr>
</tbody>
</table>

Referring: no head match

<table>
<thead>
<tr>
<th>David Cameron</th>
<th>prime minister</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>6.2%</td>
</tr>
<tr>
<td>Berkeley</td>
<td>6.2%</td>
</tr>
</tbody>
</table>

Error Analysis

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

Entity Resolution

Barack Obama  en.wikipedia.org/wiki/Barack_Obama

Michael Jordan
en.wikipedia.org/wiki/Michael_Jordan

Entity Resolution

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

Baseline

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. 

Basketball
- players who many credit with ushering the professional game to its highest level of popularity: Larry Bird, Ervin "Magic" Johnson, and Michael Jordan

Choosing the Right Query

professor Michael Jordan

Michael Jordan
Michael_I._Jordan
Jordan_(country)

Incorporating Context

Michael I. Jordan
professor: 12
Bayesian: 5
learning: 10
PhD: 3

Michael Jordan
basketball: 50
Bulls: 26
NBA: 30
game: 22
Michael Jordan hosted the Big Data Bootcamp last fall. As part of the workshop, Professor Jordan gave a talk.

A Joint Model of “Everything”

Durrett and Klein (2014)

Same type, tie predictions to words, etc.

Durrett and Klein (2014)
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

Results on NER

Results on Coreference
Results on Entity Linking

<table>
<thead>
<tr>
<th>Indep</th>
<th>Joint</th>
<th>Fahnri</th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>74.82</td>
<td>76.74</td>
</tr>
<tr>
<td>72</td>
<td></td>
<td>76.87</td>
</tr>
</tbody>
</table>

Human = 84

Information Extraction

Barack Obama
born in Honolulu
Hawaii

Barack Obama
Serve America Act

Knowledge Base

Expanded Knowledge Base

Template-Based

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

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.