Document Level Models (1)

Coreference Resolution

Yoav Goldberg
(with additions by Ido Dagan)
Bar Ilan University

Credits for slides by Mihai Surdenau and Marta Recasens
Beyond sentences

Until now

- Mostly low-level components (building blocks)
- Working at sentence level – analyzing each sentence individually

Today (++)

- Looking at the document and the corpus level.
- Still focusing on building-blocks.
Coreference Resolution
Coreference Resolution

Look up in the sky! It's a bird!

It's a plane!

Oh no!

MAXIMUMBLE.COM

©2013 CHRIS HALLBECK
Coreference Resolution

(Which entities are the same?)
What is it?

Coreference resolution = the task of clustering together of expressions that refer to the same entity/concept.

Michelle LaVaughn Robinson Obama is an American lawyer and writer. She is the wife of the 44th and current President of the United States, Barack Obama, and the first African-American First Lady of the United States.
Why is it important?

- Question answering
  - “Who is Barack Obama’s spouse?”
- Information extraction
  - “Find all per:spouse relations between all named entities in this large corpus.”
- News aggregation
  - “What are recent events involving Michelle Obama?”
  - Requires cross-document coreference resolution. More on this soon.
Why is it important?

- Performance doubles for these applications when coreference resolution is used.


Extensions:
event (predicate) coreference, cross document coreference
Real-world example
Chengdu, China (CNN) – The researcher dressed in blue plastic smock, slippers and gloves is having a tough time getting his work done.
Every time Zhang Zhen sets up his camera on a tripod in an effort to document the behavior of one of [[the panda cubs] scattered on [a grassy hillside]], one particularly frisky baby panda comes wobbling towards him, interrupting his shoot. “Mumu!” he yells in frustration, as the four-month old cub rears up on [her] hind legs, lunging towards him. He picks Mumu up and deposits her at [the opposite end] of [the enclosure]. “I’m not sure why she’s been all over me like this. I think she’s excited today,” Zhang says.
Mumu is the oldest of fourteen baby pandas that were born last summer here at the Research Base of [[Giant Panda] Breeding] in [Chengdu, China].
CNN.com, Dec 2013
Stages towards understanding

1. (Pre-processing – sentence boundary, tagging, parsing, …)
2. Entity Extraction (broader sense of entity, not just named)
3. Coreference Resolution
4. Entity Linking (link to an ontology / database / wikipedia)
   - Entity linking is a different related task, partly overlapping coref
Stages towards understanding

1. (Pre-processing – sentence boundary, tagging, parsing, …)
2. Entity Extraction
3. Coreference Resolution
4. Entity Linking (link to an ontology / database / wikipedia)
Chengdu, China (CNN) – The researcher dressed in blue plastic smock, slippers and gloves is having a tough time getting his work done. Every time Zhang Zhen sets up his camera on a tripod in an effort to document the behavior of one of [the panda cubs] scattered on [a grassy hillside]], one particularly frisky baby panda comes wobbling towards him, interrupting his shoot. “Mumu!” he yells in frustration, as the four-month old cub rears up on [her] hind legs, lunging towards him. He picks Mumu up and deposits her at [the opposite end] of [the enclosure]. “I’m not sure why she’s been all over me like this. I think she’s excited today,” Zhang says. Mumu is the oldest of fourteen baby pandas that were born last summer here at the Research Base of [[Giant Panda] Breeding] in [Chengdu, China].
CNN.com, Dec 2013
Chengdu, China (CNN) – The researcher dressed in blue plastic smock, slippers and gloves is having a tough time getting his work done.
Every time Zhang Zhen sets up his camera on a tripod in an effort to document the behavior of one of [the panda cubs] scattered on [a grassy hillside], one particularly frisky baby panda comes wobbling towards him, interrupting his shoot.
“Mumu!” he yells in frustration, as the four-month old cub rears up on [her] hind legs, lunging towards him. He picks Mumu up and deposits her at [the opposite end] of [the enclosure].
“I’m not sure why she’s been all over me like this. I think she’s excited today,” Zhang says.
Mumu is the oldest of fourteen baby pandas that were born last summer here at the Research Base of [[Giant Panda] Breeding] in [Chengdu, China].
CNN.com, Dec 2013
CNN

The researcher

blue plastic smock, slippers and gloves

though time

[his] work

Zhang Zhen

[his] camera

a tripod

an effort

the behavior

one of the panda cubs scattered on a grassy hillside

the panda cubs scattered on . . .

a grassy hillside

One particularly frisky baby panda

him

[his] shoot

Mumu

he

the four-month old cub

[her] hind legs

him

He

Mumu

her

the opposite end of [the enclosure]

I

she

me

I

she

today

Zhang

Mumu

fourteen baby pandas

last summer

here

the Research Base of [[Giant Panda] Breeding] in [Changdu, China]
Stages towards understanding

1. (Pre-processing – sentence boundary, tagging, parsing, …)
2. **Entity Extraction**
3. Coreference Resolution
4. Entity Linking (link to an ontology / database / wikipedia)
Stages towards understanding

1. (Pre-processing – sentence boundary, tagging, parsing, …)
2. Entity Extraction
3. Coreference Resolution
4. Entity Linking (link to an ontology / database / wikipedia)
Chengdu, China
CNN
The researcher
blue plastic smock, slippers and gloves
though time
[his] work
Zhang Zhen
[his] camera
a tripod
an effort
the behavior
one of the panda cubs scattered on a grassy hillside
the panda cubs scattered on . . .
a grassy hillside
One particularly frisky baby panda
him
[his] shoot
Mumu
he
the four-month old cub
[h]er] hind legs
him
He
Mumu
her
the opposite end of [the enclosure]
I
she
me
I
she
today
Zhang
Mumu
fourteen baby pandas
last summer
here
the Research Base of [[Giant Panda Breeding] in [Changdu, China]
Chengdu, China
CNN
The researcher
blue plastic smock, slippers and gloves
though time
[his] work
Zhang Zhen
[his] camera
a tripod
an effort
the behavior
one of the panda cubs scattered on a grassy hillside
the panda cubs scattered on . . .
a grassy hillside
One particularly frisky baby panda
him
[his] shoot
Mumu
he

the four-month old cub
[her] hind legs
him
He
Mumu
her
the opposite end of [the enclosure]
I
she
me
I
she
today
Zhang
Mumu
fourteen baby pandas
last summer
here
the Research Base of [[Giant Panda] Breeding] in [Changdu, China]
Chengdu, China

The researcher

blue plastic smock, slippers and gloves

though time

[his] work

Zhang Zhen

[his] camera

a tripod

an effort

the behavior

One particularly frisky baby panda

him

[his] shoot

Mumu

he

the four-month old cub

[her] hind legs

him

He

Mumu

her

the opposite end of [the enclosure]

I

she

me

I

she

today

Zhang

Mumu

fourteen baby pandas

last summer

here

the Research Base of [[Giant Panda] Breeding] in [Changdu, China]
Chengdu, China

CNN

The researcher

blue plastic smock, slippers and gloves

though time

[his] work

Zhang Zhen

[his] camera

a tripod

an effort

the behavior

one of the panda cubs scattered on a grassy hillside

the panda cubs scattered on . . .

a grassy hillside

One particularly frisky baby panda

him

[his] shoot

Mumu

he

the four-month old cub

[her] hind legs

him

He

Mumu

her

the opposite end of [the enclosure]

I

she

me

I

she

today

Zhang

Mumu

fourteen baby pandas

last summer

here

the Research Base of [[Giant Panda] Breeding] in [Changdu, China]
Chengdu, China
CNN
The researcher
blue plastic smock, slippers and gloves
though time
[his] work
Zhang Zhen
[his] camera
a tripod
an effort
the behavior
one of the panda cubs scattered on a grassy hillside
the panda cubs scattered on . . .
a grassy hillside
One particularly frisky baby panda
him
[his] shoot
Mumu
he
the four-month old cub
[her] hind legs
him
He
Mumu
her
the opposite end of [the enclosure]
l
she
me
I
she
today
Zhang
Mumu
fourteen baby pandas
last summer
here
the Research Base of [[Giant Panda] Breeding] in [Changdu, China]
Chengdu, China
CNN
The researcher
blue plastic smock, slippers and gloves
though time
[his] work
Zhang Zhen
[his] camera
a tripod
an effort
the behavior
one of the panda cubs scattered on a grassy hillside
the panda cubs scattered on . . .
a grassy hillside
One particularly frisky baby panda
him
[his] shoot
Mumu
he
the four-month old cub
[her] hind legs
him
He
Mumu
her
the opposite end of [the enclosure]
I
she
me
I
she
today
Zhang
Mumu
fourteen baby pandas
last summer
here
the Research Base of [[Giant Panda] Breeding] in [Changdu, China]
Coreferene is a clustering task

- Decide on number of clusters
- Assign entity mentions to clusters
Chengdu, China

CNN

The researcher

blue plastic smock, slippers and gloves

though time

[his] work

Zhang Zhen

[his] camera

a tripod

an effort

the behavior

one of the panda cubs scattered on a grassy hillside

the panda cubs scattered on...

a grassy hillside

One particularly frisky baby panda

him

[his] shoot

Mumu

he

the four-month old cub

[her] hind legs

him

He

Mumu

her

the opposite end of [the enclosure]

I

she

me

I

she

today

Zhang

Mumu

fourteen baby pandas

last summer

here

the Research Base of [[Giant Panda] Breeding] in [Changdu, China]
Evaluation

How to evaluate?

Types of mistakes:
- Splitting a cluster
- Merging of cluster
- Incorrect assignment

Which are more important?

How do we design a metric to capture these?
Evaluation

How to evaluate?

- Types of mistakes:
  - Splitting a cluster
  - Merging of cluster
  - Incorrect assignment
Evaluation

How to evaluate?

- Types of mistakes:
  - Splitting a cluster
  - Merging of cluster
  - Incorrect assignment

- Which are more important?

- How do we design a metric to capture these?
Chengdu, China

The researcher

blue plastic smock, slippers and gloves

though time

[his] work

Zhang Zhen

[his] camera

a tripod

an effort

the behavior

one of the panda cubs scattered on a grassy hillside

the panda cubs scattered on . . .

a grassy hillside

One particularly frisky baby panda

him

[his] shoot

Mumu

he

the four-month old cub

[her] hind legs

him

He

Mumu

her

the opposite end of [the enclosure]

I

she

me

I

she

today

Zhang

Mumu

fourteen baby pandas

last summer

here

the Research Base of [[Giant Panda] Breeding] in [Changdu, China]
Evaluation

The $B^3$ F-Score Metric

\[
P = \frac{1}{|\text{docs}|} \sum_{\text{doc} \in \text{docs}} \sum_{m \in \text{doc}} \frac{|g_m \cap p_m|}{|p_m|}
\]

\[
R = \frac{1}{|\text{docs}|} \sum_{\text{doc} \in \text{docs}} \sum_{m \in \text{doc}} \frac{|g_m \cap p_m|}{|g_m|}
\]

\[
F_1 = 2 \frac{PR}{P + R}
\]

$g_m$ gold cluster containing $m$

$p_m$ predicted cluster containing $m$

$|x|$ size of cluster $x$

$|\text{docs}|$ total number of mentions in all docs
Evaluation

The $B^3$ F-Score Metric

\[ P = \frac{1}{|docs|} \sum_{doc \in docs} \sum_{m \in doc} \frac{|g_m \cap p_m|}{|p_m|} \]

\[ R = \frac{1}{|docs|} \sum_{doc \in docs} \sum_{m \in doc} \frac{|g_m \cap p_m|}{|g_m|} \]

\[ F_1 = 2 \frac{P \cdot R}{P + R} \]

$g_m$ gold cluster containing $m$

$p_m$ predicted cluster containing $m$

$|x|$ size of cluster $x$

Eval is open for debate

- $B^3$ is good, but not perfect.
- Other variants exist. Common CoNLL F1 - averaging 3 measures
How to solve?
Chengdu, China
CNN
The researcher
blue plastic smock, slippers and gloves
though time
[his] work
Zhang Zhen
[his] camera
a tripod
an effort
the behavior
One particularly frisky baby panda
him
[his] shoot
Mumu
he
the four-month old cub
[her] hind legs
him
He
Mumu
her
the opposite end of [the enclosure]
I
she
me
I
she
today
Zhang
Mumu
fourteen baby pandas
last summer
here
the Research Base of [Giant Panda Breeding] in [Changdu, China]
Chengdu, China
The researcher
blue plastic smock, slippers and gloves
though time
[his] work
Zhang Zhen
[his] camera
a tripod
an effort
the behavior
one of the panda cubs scattered on a grassy hillside
the panda cubs scattered on . . .
a grassy hillside
One particularly frisky baby panda
him
[his] shoot
Mumu
he

the four-month old cub
[her] hind legs
him
He
Mumu
her
the opposite end of [the enclosure]
I
she
me
I
she
today
Zhang
Mumu
fourteen baby pandas
last summer
here
the Research Base of [[Giant Panda] Breeding] in [Changdu, China]
A typical algorithm

Identify all mentions
A typical algorithm

Iden\%fy all mentions

Compute link scores between all pairs
A typical algorithm

Identify all mentions

Compute link scores between all pairs

Partition this graph into entity clusters
Pairwise Approaches

▶ How to score each pair?
  ▶ Classifier.
  ▶ But how to train? (what are the examples? what are the features?)

▶ How to choose best partition given pair scores?
Choosing a partition

- Choosing a globally optimal clustering under reasonable objectives is NP-hard.
- Resort to heuristics or approximations.
Choosing a partition

- Choosing a globally optimal clustering under reasonable objectives is NP-hard.
- Resort to heuristics or approximations.

A possible algorithm 1

- For each mention in order of appearance
  - Compute scores to previous mentions
  - Decide if starting a new cluster or linking to existing cluster
Choosing a partition

- Choosing a globally optimal clustering under reasonable objectives is NP-hard.
- Resort to heuristics or approximations.

A possible algorithm 1

- For each mention in order of appearance
  - Compute scores to previous mentions
  - Decide if starting a new cluster or linking to existing cluster

A possible algorithm 2

- Assume that each mention has at most one antecedent
- Mentions with 0 antecedents start a new cluster
- Now, search for trees instead of clusters
  - (trees are easy..)
A possible algorithm 1

- For each mention in order of appearance
  - Compute scores to previous mentions
  - Decide if starting a new cluster or linking to existing cluster

Discuss features.

Discuss training examples.

Read!! - sections 2.2.1 & 3:
Discuss potential problems with the pairwise approach.
Chengdu, China
CNN
The researcher
blue plastic smock, slippers and gloves
though time
[his] work
Zhang Zhen
[his] camera
a tripod
an effort
the behavior
one of the panda cubs scattered on a grassy hillside
the panda cubs scattered on . . .
a grassy hillside
One particularly frisky baby panda
him
[his] shoot
Mumu
he

the four-month old cub
[her] hind legs
him
He
Mumu
her
the opposite end of [the enclosure]
I
she
me
I
she
today
Zhang
Mumu
fourteen baby pandas
last summer
here
the Research Base of [[Giant Panda] Breeding] in [Changdu, China]
Some insights learned

- Most algorithms focus on step 2: computing mention-pair scores using machine learning, which is a local operation
  - Poor representation of context: only two mentions considered
- Recent work showed that it is important to address coreference resolution as a global task, where all mentions are modeled jointly
  - This is hard to model using machine learning
- ML models generalize poorly to new words, domains, and languages
  - Annotating coreference is expensive
Idea I

- ✗ machine learning
- ✓ deterministic, rule-based model
- ✓ “baby steps” approach
- ✓ global model
Entity coreference resolution model

• Novel architecture for coreference resolution:
  • “Baby steps” – accurate things first
  • Global – attribute sharing in clusters
  • Deterministic – rule-based model
• Top ranked system at CoNLL-2011 Shared Task:
  • 58.3% (open), 57.8% (closed)
Baby-steps approach

- Multiple passes (or “sieves”) over text
- Precision of each pass is smaller than preceding ones
- Recall keeps increasing with each pass
- Decisions once made cannot be modified by later passes
- Modular architecture
Focus on high recall

Recall increases (precision decreases) as more sieves added

More global decisions

Post processing
The second attack occurred after some rocket firings aimed, apparently, toward the israelis, apparently in retaliation. We’re checking our facts on that one. ... the strike will undermine efforts by palestinian authorities to bring an end to terrorist attacks and does not contribute to the security of israel.
Why multiple sieves?

The second attack occurred after some rocket firings aimed, apparently, toward the israelis, apparently in retaliation. *we’re* checking our facts on that one. ... the strike will undermine efforts by palestinian authorities to bring an end to terrorist attacks and does not contribute to the security of **Israel**.

- number: plural
- animacy: animate
- number: plural
- animacy: unknown
- number: singular
- animacy: inanimate
The second attack occurred after some rocket firings aimed, apparently, toward the israelis, apparently in retaliation. We’re checking our facts on that one. The strike will undermine efforts by Palestinian authorities to bring an end to terrorist attacks and does not contribute to the security of Israel.
Why multiple sieves?

The second attack occurred after some rocket firings aimed, apparently, toward the israelis, apparently in retaliation. We’re checking our facts on that one. ... the strike will undermine efforts by Palestinian authorities to bring an end to terrorist attacks and does not contribute to the security of Israel.
Pass 1 – Mention detection

• Extract all noun phrases (NP) plus pronouns and named entities even in modifier position

• Remove non-referring expressions, e.g., generic “it”, with manually written patterns
  • E.g., *It is possible that*...
Pass 2 – Speaker identification

• Extract speakers and use the info for resolution
  • “....”, she said.

• Positive and negative constraints for following sieves:

  “I voted for Nader because he was most aligned with my values,” she said.
Pass 3 – Exact string match

Exactly the same text:

…TWA's bid for USAir skeptically, seeing it as a ploy to pressure USAir into buying TWA.

The Shahab 3 ground-ground missile: the new addition to Iran’s military capabilities … developed the Shahab 3 ground-ground missile for defense purposes with capabilities ranging from …
Pass 4 – Relaxed string match

String match after dropping the text following the head word:

…Clinton… Clinton, whose term ends in January…
Pass 5 – Precise constructs

Appositives:

… but Bob Gerson, video editor of This Week in Consumer Electronics, says Sony conceives …

Predicate nominatives:

Started three years ago, Chemical's interest-rate options group was a leading force in the field.

Role appositives:

… [[actress] Rebecca Schaeffer] …
… [[painter] Pablo Picasso] …
Pass 5 – Precise constructs

Relative pronouns:
… [the finance street [which] has already formed in the Waitan district] …

Acronyms:
Agence France Presse … AFP

Demonyms/Gentilics:
Israel… Israeli
The Japanese company already has 12% of the total camcorder market, ranking it third behind the RCA and Panasonic brands … The company also plans to aggressively start marketing … The electronics company…

- Coupled with various constraints:
  - No new information in mentions to be resolved
  - No location mismatch, “Lebanon” != “southern Lebanon”
  - No numeric mismatch, “people” != “around 200 people”
  - No i-within-i, e.g., [[Sony Corporation] of America]
Pass 10 – Relaxed head match

• Same constraints as above but anaphora head can match any word in the candidate cluster

“Sanders”
is compatible with the cluster:
{Sauls, the judge, Circuit Judge N. Sanders Sauls}
Pass 11 – Pronoun resolution

- Attributes must agree
  - Number
  - Gender
  - Person
  - Animacy
- Assigned using POS tags, NER labels, static list of assignments for pronouns
- Improved further using gender and animacy dictionaries of Bergsma and Lin (2006), and Ji and Lin (2009)
Post processing

• Discard singleton clusters
  • This is why we could maximize recall in mention detection!
• Discard shorter mentions in appositive patterns
• Discard mentions that appear later in copulative relations

• Implemented to comply with OntoNotes annotations
A run-through example

John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.
John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.
A run-through example

John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.

Speaker identification
A run-through example

John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.
A run-through example

John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.

Precise constructs
John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.

Strict head match
John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.
Mention selection in a given sieve

• In each sieve, we consider for resolution only mentions that are currently first in textual order in their cluster. (in order to decide whether to merge it with an antecedent)

• Most informative!

✓ ✔ ✔

textual order
Features are shared within clusters

- Within a cluster:
  - Union of all modifiers
  - Union of all head words
  - Union of all attributes: number, gender, animacy
- Robustness to missing/incorrect attributes

“a group of students”

- number: singular

“five students”

- number: plural

number: singular, plural
EXPERIMENTS
### Results on older corpora

<table>
<thead>
<tr>
<th>UNSUPERVISED</th>
<th>ACE 2004 Test</th>
<th>ACE NWIRE</th>
<th>MUC6</th>
</tr>
</thead>
<tbody>
<tr>
<td>This work</td>
<td>81</td>
<td>80.2</td>
<td>74.4</td>
</tr>
<tr>
<td>Haghighi and Klein (2009)</td>
<td>79.0</td>
<td>76.9</td>
<td>75.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SUPERVISED</th>
<th>Ace 2004 Test</th>
<th>ACE NWIRE</th>
<th>MUC6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culotta et al. (2007)</td>
<td>79.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bengston and Roth (2008)</td>
<td>80.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Finkel and Manning (2008) +G</td>
<td>-</td>
<td>74.5</td>
<td>64.3</td>
</tr>
</tbody>
</table>

B³ F1 scores of different systems on standard corpora
Results: CoNLL-2011 closed track

CoNLL score = (MUC F1 + $B^3$ F1 + CEAF F1) / 3
Results: CoNLL-2011 open track

CoNLL score = \((\text{MUC F1} + \text{B}^3 \text{ F1} + \text{CEAF F1}) / 3\)
CoNLL-2012 shared task

- Multilingual unrestricted coreference resolution in OntoNotes
  - English, Chinese, Arabic
- Higher barrier of entry
  - 16 submissions vs. 23 submissions in 2011
- But there was significant progress
  - Best score for English increased from 58.3 to 63.4
CoNLL-2012 shared task

- Two out of the top three systems used our system
- Fernandes et al., PUC/IBM Brazil
  - Adapted our system to Chinese and Arabic
  - Reranked the output of our system
  - Best system overall
- Chen and Ng, UT Dallas
  - Adapted our system to Chinese and Arabic
  - Added two ML-based sieves to our system
  - Best for Chinese, top 3 overall
- Proof that our approach is multilingual
## Analysis: Importance of sharing features

<table>
<thead>
<tr>
<th>Model Type</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity-centric model</td>
<td>59.3</td>
</tr>
<tr>
<td>Mention-pair model</td>
<td>55.9</td>
</tr>
</tbody>
</table>

CoNLL F1 in OntoNotes Dev
## Analysis: Importance of multiple sieves

<table>
<thead>
<tr>
<th>Model</th>
<th>CoNLL F1 in OntoNotes Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-pass model</td>
<td>59.3</td>
</tr>
<tr>
<td>Single-pass model</td>
<td>53</td>
</tr>
</tbody>
</table>

CoNLL F1 in OntoNotes Dev
## Analysis: Importance of features

<table>
<thead>
<tr>
<th>Feature</th>
<th>F1 Score</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>59.3</td>
<td></td>
</tr>
<tr>
<td>wo/ Number</td>
<td>56.7</td>
<td>- 2.6</td>
</tr>
<tr>
<td>wo/ Gender</td>
<td>58.9</td>
<td>- 0.4</td>
</tr>
<tr>
<td>wo/ Animacy</td>
<td>58.3</td>
<td>- 1.0</td>
</tr>
<tr>
<td>wo/ NE</td>
<td>58.8</td>
<td>- 0.5</td>
</tr>
</tbody>
</table>

CoNLL F1 in OntoNotes Dev
Idea 1: Conclusions

- Novel architecture for coreference resolution
  - “Baby steps”
  - Global
  - Deterministic
- State of the art results (in multiple languages)
  - Best at CoNLL-2011
  - Two of the top 3 systems at CoNLL-2012 used it
Big-picture conclusions

- Understanding the problem is more important than machine learning
- Model things jointly when you can
Recent Improvements
Taking Coreference Resolution beyond the 60% Performance Barrier

Marta Recasens
Google Research

(Joint work with Matthew Can, Marie-Catherine de Marneffe, Chris Potts, Dan Jurafsky, Eduard Hovy, and M. Antònia Martí)

April 26, 2013 • Carnegie Mellon University

* Same Referent, Different Words: Unsupervised Mining of Opaque Coreferent Mentions. NAACL 2013.
Nestle USA issued a voluntary recall of its Nesquik chocolate powder after being tipped off by an ingredient supplier of possible salmonella contamination.

The Glendale-based company said it was calling back canisters of the product, which is mixed with milk to create a sweet drink, that were made in October and sold nationwide.

Consumers should look for containers bearing an expiration date of October 2014.

Nestle decided to recall the power
Nestle USA issued a voluntary recall of its Nesquik chocolate powder after being tipped off by an ingredient supplier of possible salmonella contamination.

The Glendale-based company said it was calling back canisters of the product, which is mixed with milk to create a sweet drink, that were made in October and sold nationwide.

Consumers should look for containers bearing an expiration date of October 2014.

Nestle decided to recall the powder.
Nestle USA issued a voluntary recall of its Nesquik chocolate powder after being tipped off by an ingredient supplier of possible salmonella contamination.

The Glendale-based company said it was calling back canisters of the product, which is mixed with milk to create a sweet drink, that were made in October and sold nationwide.

Consumers should look for containers bearing an expiration date of October 2014.

Nestle decided to recall the powder.
Nestle USA issued a voluntary recall of its Nesquik chocolate powder after being tipped off by an ingredient supplier of possible salmonella contamination. The Glendale-based company was calling back canisters of the product, which is mixed with milk to create a sweet drink, that were made in October and sold nationwide. Consumers should look for containers bearing an expiration date of October 2014. Nestle decided to recall the powder.
Singleton mentions are hard and common.

- Design a classifier specifically for predicting them.
- Use a different set of linguistically motivated features.

Then, predicted singletons are filtered and not considered by the coreference algorithm, reducing its errors.
What’s hard? Why only \(\sim 60\%\) accuracy?
The unsolved problem of coreference resolution

The flaw was first reported by a security researcher David Emery, who posted his findings to the Cryptome mailing list. [...] The bug has not been corrected by any subsequent updates.

The software is used to turn 2D photos into 3D models; in reality, a person uploads photos taken or stored on an iPad to the Autodesk Cloud, where the actual conversion happens. [...] The app is free, but requires an iPad 2 or better running iOS 5.x.
Autodesk's had its 123D Catch iPad application in the works for quite some time now, but starting today, you'll finally be able to use that Cupertino slate to turn those beautiful snaps into three-dee creations.

Now you can keep up with all of the people you follow with a “best-of” weekly email from Twitter. [...] The micro-blogging service will now be sending out weekly email digests that will feature a summary of your Twitter stream.

The unsolved problem of coreference resolution
• Surface features
  – String/head match
  – Sentence/token distance
• Morphological features
  – Mention is a pronoun/definite/demonstrative/proper noun
• Syntactic features
  – Gender/number agreement
  – Grammatical role
• Semantic features
  – NE type
  – WordNet
  – Wikipedia
  – Others: Yago, lexico-semantic patterns, etc.
Semantic class match
(Soon et al. 01)
WordNet

WordNet paths
(Harabagiu et al. 01, Ng & Cardie 02, Poesio et al. 04, Ponzetto & Strube 06)

IN-GLOSS

S: (n) manufacturer, maker, manufacturing business (a business engaged in manufacturing some product)

SYNONYM

S: (n) manufacturer, maker, manufacturing business (a business engaged in manufacturing some product)
WordNet

WordNet paths
(Harabagiu et al. 01, Ng & Cardie 02, Poesio et al. 04, Ponzetto & Strube 06)
Semantic similarity is not coreference

- Related by WordNet

- Coreferent

- Scientists: people
  - Arm
  - Leg

- Company: allies

- Voters: people
  - Authority: government

- Fire: cause
  - Energy sources: gas supplies
  - Government: chairman
Distributional similarity

Distributional hypothesis (Harris 1954): words that occur in the same contexts tend to have similar meanings.

```plaintext
     aardvark  computer  data  pinch  result  sugar  ...
apricot     0        0       0      1       0     1
pineapple   0        0       0      1       0     1
digital     0        2       1      0       1     0
information 0       1       6      0      4     0
```

```
|    | subj.of.absorb | subj.of.absorb | subj.of_behave | subj.of_behave | ... | pobj.of.inside | pobj.of.into | ... | nmod.of.abnormality | nmod.of.anemia | nmod.of.architecture | nmod.of.architecture | nmod.of.attack | nmod.of.attack | nmod.of.attack | nmod.of.attack | nmod.of.attack | nmod.of.attack | nmod.of.attack | nmod.of_attack | nmod.of.attack | nmod.of.attack | ...
|----|----------------|----------------|----------------|----------------|-----|----------------|-------------|-----|---------------------|----------------|--------------------|--------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-----
| cell| 1              | 1              | 1              | 16             | 30  | 3              | 1           | 1   | 8                   | 1              | 1                  | 1                  | 8              | 1              | 1              | 1              | 1              | 1              | 1              | 1              | 1              | 1              | ...
```
Distributional similarity is still not coreference.
Microsoft has released a new feature.

The search giant has released a new feature.
Google has acquired the company.

Acquisition of Nik Software

The search giant has acquired the company.

Intuition of our solution: Restricted distributional similarity
Restricted distributional semantics

Story: Sprint-blocks-out-employees-vacations

Sprint blocks out vacation days for a major phone announcement.

According to SprintFeed, the carrier is blocking out vacation days for employees.
Comparable corpus

Techmeme (www.techmeme.com)
Zynga’s stock price is **dropping**, so now the San Francisco-based company is attempting to devise creative ways to monetize some of its more popular games. Yesterday we **reported** that Draw Something is hemorrhaging users, but that hasn’t stopped its new parent company from thrusting more ads upon the once wildly-popular game.

Zynga’s latest big-ticket acquisition has already figured out how to draw in users, but now Draw Something has an ad model that brings brands into the picture.

Until recently, the Pictionary-like game had only run spammy banner ads in its free mobile app that, including the paid no-ads version, has amassed a staggering 50 million downloads in five months. Now, with a direct-sales force that’s been on the ground for a whole eight weeks, Draw Something is inserting advertisers’ paid terms into the game for players to literally draw brands.

Here’s how the game works: Pick a word from a list of three, then create a drawing so a Facebook
Zynga’s stock price is on the rise as the San Francisco-based company is attempting to monetize some of its free games. The company was reported to be developing a new form of advertising (on top of its existing banner ads), now inserting words connected with brands, instead of a product that they sell.

With the ink still drying following the acquisition of OMGPOP and its hit app Draw Something, Zynga is putting its new mobile property to work by engaging advertisers and encouraging them to pay for words that allow users to literally draw brands.

Until recently, the Pictionary-like game had only run spammy banner ads in its free mobile app that, including the paid no-ads version, has amassed a staggering 50 million downloads in five months. Now, with a direct-sales force that’s been on the ground for a whole eight weeks, Draw Something is inserting advertisers’ paid terms into the game for players to literally draw brands.

Here’s how the game works: Pick a word from a list of three, then create a drawing so a Facebook
Extraction

- Stanford sentence splitter, tagger, NER
- MaltParser (linear time)

- Top 10 tf*idf ranked verbs for each story

\[
\text{tf}(v, s) * \text{idf}(v, S) = \text{tf}(v, s) * \log \frac{|S|}{|\{s \in S : v \in s\}|}
\]

- Phrasal verbs (\textit{give up} vs. \textit{give away})
- Excluding: light verbs (\textit{do, have, give}...)
  - report verbs (\textit{say, tell}...)
  - copular verbs (\textit{seem, become}...)
- WordNet synonyms are included (\textit{release, publish}...)
Assumption: In a story, the same verb refers to the same event

Subjects and objects are clustered, respectively

- Passive constructions ($X \text{ compromised } Y \rightarrow Y \text{ has been compromised}$)
- Ergative verbs ($X \text{ scattered } Y \rightarrow Y \text{ scattered}$)
- Nominalizations from NomBank ($\text{acquire } \rightarrow \text{Google’s acquisition of Sparrow}$)

Exclude same-head NPs and pronouns
Coreference relations

- Android phones  products
- pictures  shots
- Mark Zuckerberg  the hoodie-wearing Facebook co-founder

Bad relations

① Parsing errors

- [attacks against Chrome]_{s} exploit ...  [the full details on the]_{s} exploit

② Algorithm violations (one verb ≠ one event)

- Remove [spam from the emails]_{o} ...
- Remove [the test accounts]_{o}

③ Text extraction errors

- </li> <li id="gadgets"> <a href=http://www.thetechherald.com/> Networking </a>
Filters for:

• Parsing errors
  – Non-nominal head

• Algorithm violations
  – NE – NE
  – Negation
  – Enumeration
  – Numbers
  – Temporals

• Text extraction errors
  – Mention length
  – Sentence length
  – Ill-formed sentence

\( \text{shopping ( ah} \)
Generalization

- Remove determiners
  
  `the promotion` >> `promotion`

- relative, -ing, -ed clauses
  
  `the device available online from Google` >> `device`

- Keep adjectives and prepositional modifiers
  
  `online piracy`

  `distribution of pirated material`

- Generalize NE to types
  
  `Cook’s departure` >> `PERSON’s departure`

- Lemmas
  
  `data` >> `datum`

128,492 coreferent pairs
Generalization

• Frequency counts

- (rule, limitation)  5  
- (phone, experience)  1  
- (FBI, agent)  20

- (company, HP)  35  
- (company, price)  12

• Normalized PMI (Bouma 2009) \([-1, 1]\)

\[
PMI(x, y) = \ln \frac{p(x, y)}{p(x)p(y)} \quad \text{NPMI}(x, y) = \frac{PMI}{-\ln p(x, y)}
\]

- (rule, limitation)  0.417  
- (phone, experience)  –0.152  
- (FBI, agent)  0.566

- (company, HP)  0.203  
- (company, price)  –0.053
Dictionary snapshot

offering, IPO
password, login information
user, consumer
firm, company
phone, device
Apple, company
iPad, slate
Android, platform
site, company
app, software
agreement, wording
platform, code
filing, complaint

search, search result
update, change
bug, issue
Google, search giant
search algorithm, search engine
hardware key, digital lock
content, photo
rule, limitation
coupon, sale
medical record, medical file
device, developer
version, handset
Groupon, company
Dictionary snapshot

• Synonymy
  
  *user, consumer*

• Hypernymy
  
  *Google, company*

• Metonymy
  
  *cloud, users*

• General nouns
  
  *bug, issue*

• World knowledge
  
  *Google, search giant*
Stanford coreference system
(Lee et al. 2011)
Existing Tools

Two good tools (available for download) are:

- Stanford Coreference System
  - Sieve + singletons
- Illinois Coreference Package
  - Pairwise classification with strong features
The coreference resolution task

- Definition
- Evaluation
- Pairwise / machine learning approach
  - Features
  - Constructing training examples
- Rule-based systems (Sieve, baby-steps)
  - Many smart decisions
  - Global constraints
- Targetting specific problems
  - A method for detecting singletons
  - “semantic” knowledge acquisition from large corpus

Extensions: event (predicate) coreference, cross document coreference