Lecture 16: Coreference Resolution
Determining who is who and what is what

Harvard IACS
Chris Tanner
"You would rather have a Lexus or justice? A dream or some substance? A Bimmer, a necklace, or freedom?"
-- Dead Prez
ANNOUNCEMENTS

- HW4 is out
- HW2 and Phase 2 are Quiz 5 have been graded
- Research Project Phase 3 due Oct 28 (Thurs) @ 11:59pm
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
These systems hinge upon understanding what you’re saying (discourse) and the meaning of it (semantics)
Also necessary for information retrieval, question-answering, document summarization, etc

“TL;DR crypto stocks are surging”

Event coreference for information extraction. Humphreys et al., 1997
Question answering based on semantic structures. Narayanan and Harabagiu, 2004
Sub-event based multi-document summarization. Daniel et al., 2003
Opinion

The Freeing of the Ever Given

The stuck container ship became the butt of online jokes, but it was no minor crisis.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
Opinion

The Freeing of the Ever Given

The stuck container ship became the butt of online jokes, but it was no minor crisis.

The New York Times
By Serge Schmemann
April 1, 2021

In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.

The Freeing of the Ever Given

The stuck container ship became the butt of online jokes, but it was no minor crisis.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.

Coreference Resolution

The task of determining which words all refer to the same underlying real-world thing.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.

Coreference Resolution

The task of determining which words all refer to the same underlying real-world thing is easy for humans.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.
Types of referring expressions

Indefinite noun phrases:
  • I saw an incredible oak tree today

Definite noun phrases:
  • I read about it in the New York Times

Pronominal mentions:
  • Emily aced the quiz, as she expected

Nominal mentions and names:
  • The amazing marathoner, Des Linden, is a true inspiration

Demonstratives:
  • These pretzels are making me thirsty.

This, that, these, those
Good models should be able to perform coreference resolution across multiple documents
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.

SUEZ, Egypt (AP) — Experts boarded the massive container ship Tuesday that had blocked Egypt’s vital Suez Canal and disrupted global trade for nearly a week, seeking answers to a single question that could have billions of dollars in legal repercussions: What went wrong?
And handle **events**
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on the shore and a virtual world beyond.

SUEZ, Egypt (AP) — Experts boarded the massive container ship Tuesday that had blocked Egypt’s vital Suez Canal and disrupted global trade for nearly a week, seeking answers to a single question that could have billions of dollars in legal repercussions: What went wrong?
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set Ever Given—a virtual world beyond.

SUEZ, Egypt (AP) — Experts boarded the massive container ship Tuesday that had blocked Egypt’s vital Suez Canal and disrupted global trade for nearly a week, seeking answers to a single question that could have billions of dollars in legal repercussions: What went wrong?
mammoth barge
it
Ever Given
its
container ship

Experts

Egyptians

Egyptian mud

wedged
.set afloat
blocked
drawing cheers

disrupted

wrenching

seeking

succeeded
Takeaway #1

Coreference resolution determines which mentions all refer to the same underlying entity or event, and is ultimately a clustering task.
Early research demonstrated highly-effective rule-based entity coref systems

<table>
<thead>
<tr>
<th>Ordered sieves</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Mention Detection Sieve</td>
</tr>
<tr>
<td>2. Discourse Processing Sieve</td>
</tr>
<tr>
<td>3. Exact String Match Sieve</td>
</tr>
<tr>
<td>4. Relaxed String Match Sieve</td>
</tr>
<tr>
<td>5. Precise Constructs Sieve (e.g., appositives)</td>
</tr>
<tr>
<td>6-8. Strict Head Matching Sieves A-C</td>
</tr>
<tr>
<td>9. Proper Head Word Match Sieve</td>
</tr>
<tr>
<td>10. Alias Sieve</td>
</tr>
<tr>
<td>11. Relaxed Head Matching Sieve</td>
</tr>
<tr>
<td>12. Lexical Chain Sieve</td>
</tr>
<tr>
<td>13. Pronouns Sieve</td>
</tr>
</tbody>
</table>

Table 1: The sieves in our system; sieves new to this paper are in bold.
Rule 1: cluster together all entity mentions that are identical

The Ever Given cargo ship has been stuck for the past six days. While reports of Ever Given started to …
Rule 10: cluster together all entity mentions that are aliases according to Wikipedia

Donald Glover, better known as Childish Gambino, has written and produced an incredible TV series titled Atlanta.
Then, many systems threw tons of manually-defined features into their models.

Narrowing the Modeling Gap: A Cluster-Ranking Approach to Coreference Resolution. Rahman and Ng, JAIR 2011

Improving Coreference Resolution by Learning Entity-Level Distributed Representations. Clark and Manning. ACL 2016

Additional Mention Features: The type of the mention (pronoun, nominal, proper, or list), the mention’s position (index of the mention divided by the number of mentions in the document), whether the mentions are contained in another mention, and the length of the mention in words.

Document Genre: The genre of the mention’s document (broadcast news, newswire, web data, etc.).

Distance Features: The distance between the mentions in sentences, the distance between the mentions in intervening mentions, and whether the mentions overlap.

Speaker Features: Whether the mentions have the same speaker and whether one mention is the other mention’s speaker as determined by string matching rules from Raghunathan et al. (2010).

String Matching Features: Head match, exact string match, and partial string match.
Then, many systems threw tons of manually-defined features into their models.

Improving Coreference Resolution by Learning Entity-Level Distributed Representations. Clark and Manning. ACL 2016


Entity Coreference (2011 – present)

Takeaway #2

Research has largely relied on ML models w/ many manually-defined features.

Strong results but clear limitations.
Actress Lindsay Lohan finally checked into court-mandated rehab at the Betty Ford Center late Thursday.

Lindsay Lohan checked into the Betty Ford Clinic in Rancho Mirage, California on Thursday night, for what is to be a three-month stay, her rep confirms to People.
SameLemma: if two mentions have the same lemma (base form), classify them as being coref!

<table>
<thead>
<tr>
<th>Original word</th>
<th>Lemmatization</th>
</tr>
</thead>
<tbody>
<tr>
<td>running</td>
<td>run</td>
</tr>
<tr>
<td>ran</td>
<td>run</td>
</tr>
</tbody>
</table>

This shouldn’t work so well, but it does.
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
Coreference resolution
Coreference resolution
An earthquake with a preliminary magnitude of 4.6 was recorded in the North Bay this morning, according to the U.S. Geological Survey. The quake occurred at 2:09 a.m. about 14 miles north-northeast of Healdsburg and had a depth of 1.2 miles. It was followed by a 2.9 aftershock at 2:12 a.m. and a 2.2 at 2:15 a.m... there are no reports of injuries or major damage.

4.6 Magnitude Quake Rattles Sonoma County
Early Thursday
An earthquake measuring 4.6 rattled Sonoma and Lake counties early Thursday, according to the U.S. Geological Survey. The quake occurred at 2:09 a.m., about 14 miles northeast of Healdsburg, on the Maacama Fault with a depth of 12 miles. A Sonoma County Sheriff’s dispatcher said around 7 a.m. that there had been no reports of damage or injuries.
Coreference resolution

Mention Detection → Mention Pair Model → Clustering

corpus

cluster₁
- mammoth
- barge
- it
- Ever Given
- its
- container ship

cluster₂
- Experts
Mention Detection → Mention Pair Model → Clustering

Coreference resolution

cluster₁
- mammoth barge
- it
- Ever Given
- its
- container ship

cluster₂
- Experts
Mention Pair Model

Calculates a coref probability for all pairs of mentions

- shook
- shook
- recorded
- rattles
- according to
- rattles
<table>
<thead>
<tr>
<th>Feature</th>
<th># dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word embeddings</td>
<td>300</td>
</tr>
<tr>
<td>Lemma embeddings</td>
<td>300</td>
</tr>
<tr>
<td>Dependency Parse embeddings</td>
<td>400</td>
</tr>
<tr>
<td>Character embeddings</td>
<td>100</td>
</tr>
<tr>
<td>Part-of-speech embeddings</td>
<td>300</td>
</tr>
</tbody>
</table>
Conjoined CNN

Conjoined CNN

**Distance Score:** $L^2$ norm

**Loss Function:** Contrastive Loss

\[
(1 - Y) \frac{1}{2} (D_W)^2 + (Y) \frac{1}{2} \{max(0, m - D_W)\}^2
\]

Two identical networks with tied weights

Conjoined CNN

<table>
<thead>
<tr>
<th>Pair</th>
<th>Score</th>
<th>Coref</th>
<th>Coref</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_{17}, m_2$</td>
<td>0.0</td>
<td>erupted</td>
<td>erupted</td>
</tr>
<tr>
<td>$m_{17}, m_4$</td>
<td>0.0</td>
<td>erupted</td>
<td>erupted</td>
</tr>
<tr>
<td>$m_5, m_{923}$</td>
<td>0.03</td>
<td>announced</td>
<td>announce</td>
</tr>
<tr>
<td>$m_{78}, m_57$</td>
<td>0.05</td>
<td>erupt</td>
<td>erupted</td>
</tr>
</tbody>
</table>

**0.5 threshold**

<table>
<thead>
<tr>
<th>Pair</th>
<th>Score</th>
<th>Coref</th>
<th>Coref</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_{801}, m_{39}$</td>
<td>0.97</td>
<td>revealed</td>
<td>broke into</td>
</tr>
<tr>
<td>$m_{26}, m_{48}$</td>
<td>0.98</td>
<td>handed down</td>
<td>confirmed</td>
</tr>
</tbody>
</table>

**Accuracy:** 92.4
**Precision:** 55.8
**Recall:** 71.2
**F1 Score:** 62.8
LibSVM and FFNN received same features as CCNN, plus relational features (e.g., cosine sim., dot-product, WordNet)

Coreference resolution
Coreference resolution
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
Clustering

Goal is to return clusters from the fully-connected graph

```
Shook

0.9
0.71
0.2

0.2
0.08

Recorded

0.82
0.23

0.08

According to

0.12

Rattles

0.52
0.73
0.15
```
Goal is to return clusters from the fully-connected graph.
Nearly 100% of past systems simply performed agglomerative clustering.

We want:

• More holistic, cluster-to-cluster predictions
• Less sensitivity to non-uniformity across topics
• No additional stopping parameter
• Prevention against an all-subsuming cluster
Clustering

- min-pair distance: $\min_{m_i, m_j} d(m_i, m_j)$
- avg-pair distance: $\frac{\sum_{m_i, m_j} d(m_i, m_j)}{||C_x|| ||C_y||}$
- max-pair distance: $\max_{m_i, m_j} d(m_i, m_j)$
- size of candidate cluster: $\frac{||C_x|| + ||C_y||}{\sum_z ||C_z||}$

Coreference resolution
Outline

Coreference Resolution

Conjoined CNN

Neural Clustering

Results

Improvements

Leveraging Data

No Data

Better Data

Additional Research
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
Gold Test Mentions

... as Peter Capaldi stepped into Matt Smith’s soon to be vacant ...

Predicted Test Mentions

... as Peter Capaldi stepped into Matt Smith’s soon to be vacant ...
Feature Ablation (full system)

Lemma + Character Embeddings yields the best performance
### Table 4.6: Coreference Systems’ clustering performance on the ECB+ test set, using the predicted mentions and testing procedure from Choubey and Huang [20]. Our CCNN models use only the Lemma + Character Embedding features. FFNN denotes a Feed-Forward Neural Network Mention-Pair model. AGG denotes Agglomerative Clustering. Our models’ scores represent the average from 50 runs, with standard deviation denoted by ( ).
Using **Gold Mentions**

<table>
<thead>
<tr>
<th></th>
<th>Within-Document</th>
<th>Cross-Document</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MUC</td>
<td>B^3</td>
<td>CEAF</td>
<td>CoNLL F1</td>
<td>MUC</td>
</tr>
<tr>
<td><strong>Test Set: ECB+ Gold Mentions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SameLemma</td>
<td>58.3</td>
<td>83.0</td>
<td>75.9</td>
<td>72.4</td>
<td>84.2</td>
</tr>
<tr>
<td>FFNN+AGG</td>
<td>59.9</td>
<td>85.6</td>
<td>78.4</td>
<td>74.6</td>
<td>77.7</td>
</tr>
<tr>
<td>FFNN+NC</td>
<td>60.7</td>
<td>86.7</td>
<td>79.4</td>
<td>75.6</td>
<td>74.9</td>
</tr>
<tr>
<td>CCNN+AGG</td>
<td>70.5</td>
<td>89.1</td>
<td>83.5</td>
<td>81.0</td>
<td>84.1</td>
</tr>
<tr>
<td>CCNN+NC</td>
<td>70.9</td>
<td>88.9</td>
<td>83.6</td>
<td>81.2</td>
<td>86.4</td>
</tr>
</tbody>
</table>
FINDINGS

- State-of-the-art for event coref
- Contextualized representations
- More holistic clustering
- Char + Lemma Embeddings were the only two necessary features
Total # of Mention-Pairs to test: 8,669

# False Positives: 86
# False Negatives: 569

**False Positives**

- semantics — 82%
  - context-dependent (30%)
  - similar meanings (38%)
  - wide-reading (14%)
- unclear — 13%
- syntax — 3%
- too difficult for me — 2%

**False Negatives**

- semantics — 42%
- unclear — 20%
- slang — 16%
- longer names — 14%
- pronouns — 8%
CCNN + Clustering

False Positive

Sony announced today ...

Friday, Obama announced ...

False Negatives

The casting of Smith ...

Smith stepped into the role ...

Smith was handed the keys to play ...

False Negative

Two of the bombs fell within the Yida Camp, including ...

The UN Refugee Agency on Friday strongly condemned the aerial bombing of ...
Takeaway #3  The community needs a **better corpus**.

Takeaway #4  Event coref is especially hard, but using deep learning w/ **contextualized representations** works well.
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
End-to-end neural systems

Entity Coreference

Mention Detection → Mention Pair Model → Clustering

End-to-end Neural Coreference Resolution. Lee et al. EMNLP 2017.
End-to-end Neural Coreference Resolution. Lee et al. EMNLP 2017.
Entity Coreference

Uses several important features

<table>
<thead>
<tr>
<th>Features</th>
<th>Avg. F1</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our model (ensemble)</td>
<td>69.0</td>
<td>+1.3</td>
</tr>
<tr>
<td>Our model (single)</td>
<td>67.7</td>
<td></td>
</tr>
<tr>
<td>- distance and width features</td>
<td>63.9</td>
<td>-3.8</td>
</tr>
<tr>
<td>- GloVe embeddings</td>
<td>65.3</td>
<td>-2.4</td>
</tr>
<tr>
<td>- speaker and genre metadata</td>
<td>66.3</td>
<td>-1.4</td>
</tr>
<tr>
<td>- head-finding attention</td>
<td>66.4</td>
<td>-1.3</td>
</tr>
<tr>
<td>- character CNN</td>
<td>66.8</td>
<td>-0.9</td>
</tr>
<tr>
<td>- Turian embeddings</td>
<td>66.9</td>
<td>-0.8</td>
</tr>
</tbody>
</table>
The Ever Given has

Encoder #1
r

Encoder #2

Encoder #3

BERT
(encodes rich information very well)

y_1
y_2
y_3
y_4

The
x_1

Ever
x_2

Given
x_3

has
x_4
End-to-End

End-to-End

Entity Coreference

REMAINING ISSUES

- Pronouns (especially in conversation)

- Conflating relatedness with equality
  (e.g., “Flight attendants” with “pilots”)

- World-knowledge
  Also such location devices, *(some ships)* have smoke floats *(they)* can toss out so the man overboard will be able to use smoke signals as a way of trying to, let the rescuer locate *(them)*.

- Mention paraphrasing
  (e.g., “Royals” with “Prince Charles and his wife Camilla”)

Coref is still far from solved

<table>
<thead>
<tr>
<th>Category</th>
<th>Snippet</th>
<th>#base</th>
<th>#large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Related Entities</td>
<td>Watch spectacular performances by dolphins and sea lions at the Ocean Theater... It seems the North Pole and the Marine Life Center will also be renovated.</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Lexical</td>
<td>Over the past 28 years, the Ocean Park has basically.. The entire park has been ...</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>Pronouns</td>
<td>In the meantime, our children need an education. That’s all we’re asking.</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>Mention Paraphrasing</td>
<td>And in case you missed it the Royals are here. Today Britain’s Prince Charles and his wife Camilla...</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>Conversation</td>
<td>(Priscilla:) My mother was Thelma Wahl. She was ninety years old ... (Keith:) Priscilla Scott is mourning. Her mother Thelma Wahl was a resident ...</td>
<td>18</td>
<td>16</td>
</tr>
<tr>
<td>Misc.</td>
<td>He is my, She is my Goddess, ah</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>93</strong></td>
<td><strong>74</strong></td>
</tr>
</tbody>
</table>

Table 3: Qualitative Analysis: #base and #large refers to the number of cluster-level errors on a subset of the OntoNotes English development set. Underlined and bold-faced mentions respectively indicate incorrect and missing assignments to italicized mentions/clusters. The miscellaneous category refers to other errors including (reasonable) predictions that are either missing from the gold data or violate annotation guidelines.
Takeaway #5
Pre-trained LLM (e.g., BERT) capture rich information but miss nuanced cases
However, coref is still far from solved

**Takeaway #5**  
Pre-trained LLM (e.g., BERT) capture rich information but miss nuanced cases

**Takeaway #6**  
Until we have better data, we don’t fully understand the capabilities of our existing systems, nor do we know what is possible.
Takeaway #1

Coreference resolution determines which mentions all refer to the same underlying entity or event, and is ultimately a clustering task.

Takeaway #2

Research has largely relied on ML models w/ many manually-defined features. Strong results but clear limitations.

Takeaway #3

The community needs a better corpus.
Takeaway #4

Event coref is especially hard, but using deep learning w/ contextualized representations works well.

Takeaway #5

Neural pre-trained text encoders (e.g., BERT) capture rich information but miss nuanced cases.

Takeaway #6

Until we have better data, we don’t fully understand the capabilities of our existing systems, or know what’s possible.
Performance is reaching an asymptote.

Instead of hammering away on a problem and throwing complex models at it, pay close attention to:

1. What you’re trying to model (i.e., your data)
2. How you’re framing the problem
   (e.g., a clustering task via pairwise predictions)
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
How can we equip our coreference resolution models with common sense knowledge?

Sahithya Ravi
UBC
PhD Student

Ning Hua
Harvard
Bioinformatics MS

Vered Shwartz
UBC
Assistant Professor
• SOTA coref model uses RoBERTa as a base.

• Before coref training, we fine-tune the RoBERTa base on ConceptNet.
• Also, using graph embeddings and graph alignments to influence coref modelling
Can Bayesian clustering improve joint entity and event coreference?

Joint Entity and Event Coreference

Xin Zeng
IACS MS Thesis
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
Since labelled data is a scarce commodity, can we build a powerful unsupervised model?

Alessandro Stolfo  
ETH-Zurich  
PhD Student

Vikram Gupta  
ETH-Zurich  
Research Affiliate

Mrinmaya Sachan  
ETH-Zurich  
Assistant Professor
We combine the old school, manual rule-based system with the SOTA BERT-based end-to-end model.

Ordered sieves
1. Mention Detection Sieve
2. Discourse Processing Sieve
3. Exact String Match Sieve
4. Relaxed String Match Sieve
5. Precise Constructs Sieve (e.g., appositives)
6-8. Strict Head Matching Sieves A-C
9. Proper Head Word Match Sieve
10. Alias Sieve
11. Relaxed Head Matching Sieve
12. Lexical Chain Sieve
13. Pronouns Sieve

We combine the old school, manual rule-based system with the SOTA BERT-based end-to-end model.

Ordered sieves
1. Mention Detection Sieve
2. Discourse Processing Sieve
3. Exact String Match Sieve
4. Relaxed String Match Sieve
5. Precise Constructs Sieve (e.g., appositives)
6-8. Strict Head Matching Sieves A-C
9. Proper Head Word Match Sieve
10. Alias Sieve
11. Relaxed Head Matching Sieve
12. Lexical Chain Sieve
13. Pronouns Sieve

Unsupervised (doesn’t need training data)

We combine the old school, manual rule-based system with the SOTA BERT-based end-to-end model.

Ordered sieves
1. Mention Detection Sieve
2. Discourse Processing Sieve
3. Exact String Match Sieve
4. Relaxed String Match Sieve
5. Precise Constructs Sieve (e.g., appositives)
6-8. Strict Head Matching Sieves A-C
9. Proper Head Word Match Sieve
10. Alias Sieve
11. Relaxed Head Matching Sieve
12. Lexical Chain Sieve
13. Pronouns Sieve

Let’s use this as synthetic “gold” labels for BERT

Unsupervised (doesn’t need training data)

Supervised (needs training data)

Training with **noisy (imperfect) rule-based labels** would limit our BERT model to perform no better than the rule-based system.
CONCERN

Training with noisy (imperfect) rule-based labels would limit our BERT model to perform no better than the rule-based system.

FINDINGS

Our combined BERT model successfully uses distant-supervision to outperform the rule-based system.
<table>
<thead>
<tr>
<th></th>
<th>MUC</th>
<th></th>
<th></th>
<th>B^3</th>
<th></th>
<th></th>
<th></th>
<th>CEAF_\phi_4</th>
<th></th>
<th></th>
<th>CoNLL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F_1</td>
<td>P</td>
<td>R</td>
<td>F_1</td>
<td>P</td>
<td>R</td>
<td>F_1</td>
<td>P</td>
<td>R</td>
<td>F_1</td>
</tr>
<tr>
<td>Stanford (Lee et al., 2011)</td>
<td>64.3</td>
<td>65.2</td>
<td>64.7</td>
<td>49.2</td>
<td>56.8</td>
<td>52.7</td>
<td>52.5</td>
<td>46.6</td>
<td>49.4</td>
<td>55.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multigraph (Martschat, 2013)</td>
<td>-</td>
<td>-</td>
<td>65.4</td>
<td>-</td>
<td>-</td>
<td>54.4</td>
<td>-</td>
<td>-</td>
<td>50.2</td>
<td>56.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsup. Ranking (Ma et al., 2016)</td>
<td>-</td>
<td>-</td>
<td>67.7</td>
<td>-</td>
<td>-</td>
<td>55.9</td>
<td>-</td>
<td>-</td>
<td>51.8</td>
<td>58.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c2f-coref</td>
<td>65.7</td>
<td>68.0</td>
<td>66.9</td>
<td>50.9</td>
<td>59.4</td>
<td>54.8</td>
<td>52.9</td>
<td>49.1</td>
<td>50.9</td>
<td>57.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-base + c2f-coref</td>
<td>66.8</td>
<td>69.2</td>
<td>68.0</td>
<td>51.5</td>
<td>60.6</td>
<td>55.7</td>
<td>53.1</td>
<td>50.3</td>
<td>51.7</td>
<td>58.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SpanBERT-base + c2f-coref</td>
<td>67.6</td>
<td>68.5</td>
<td>68.1</td>
<td>53.1</td>
<td>60.1</td>
<td>56.4</td>
<td>54.8</td>
<td>50.4</td>
<td>52.5</td>
<td>59.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-large + c2f-coref</td>
<td>67.2</td>
<td>69.7</td>
<td>68.5</td>
<td>52.3</td>
<td>61.2</td>
<td>56.4</td>
<td>54.0</td>
<td>51.0</td>
<td>52.5</td>
<td>59.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SpanBERT-large + c2f-coref</td>
<td>67.4</td>
<td>69.8</td>
<td>68.6</td>
<td>52.4</td>
<td>61.8</td>
<td>56.7</td>
<td>54.1</td>
<td>51.4</td>
<td>52.7</td>
<td>59.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Results on the test set of the English CoNLL-2012 shared task<sup>3</sup>. The scores relative to the c2f-coref model are obtained after training on the labels produced by Stanford’s system. Scores for Multigraph and the Unsupervised Ranking model are reported in Ma et al. (2016).
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
Outline

Coreference Resolution
- Conjoined CNN
- Neural Clustering
- Results

Improvements
- Leveraging Data
- No Data
- Better Data

Additional Research
Coreference Resolution has had many exciting advances in the last 10 years, but it’s far from solved and remains one of the most challenging and exciting NLP tasks.