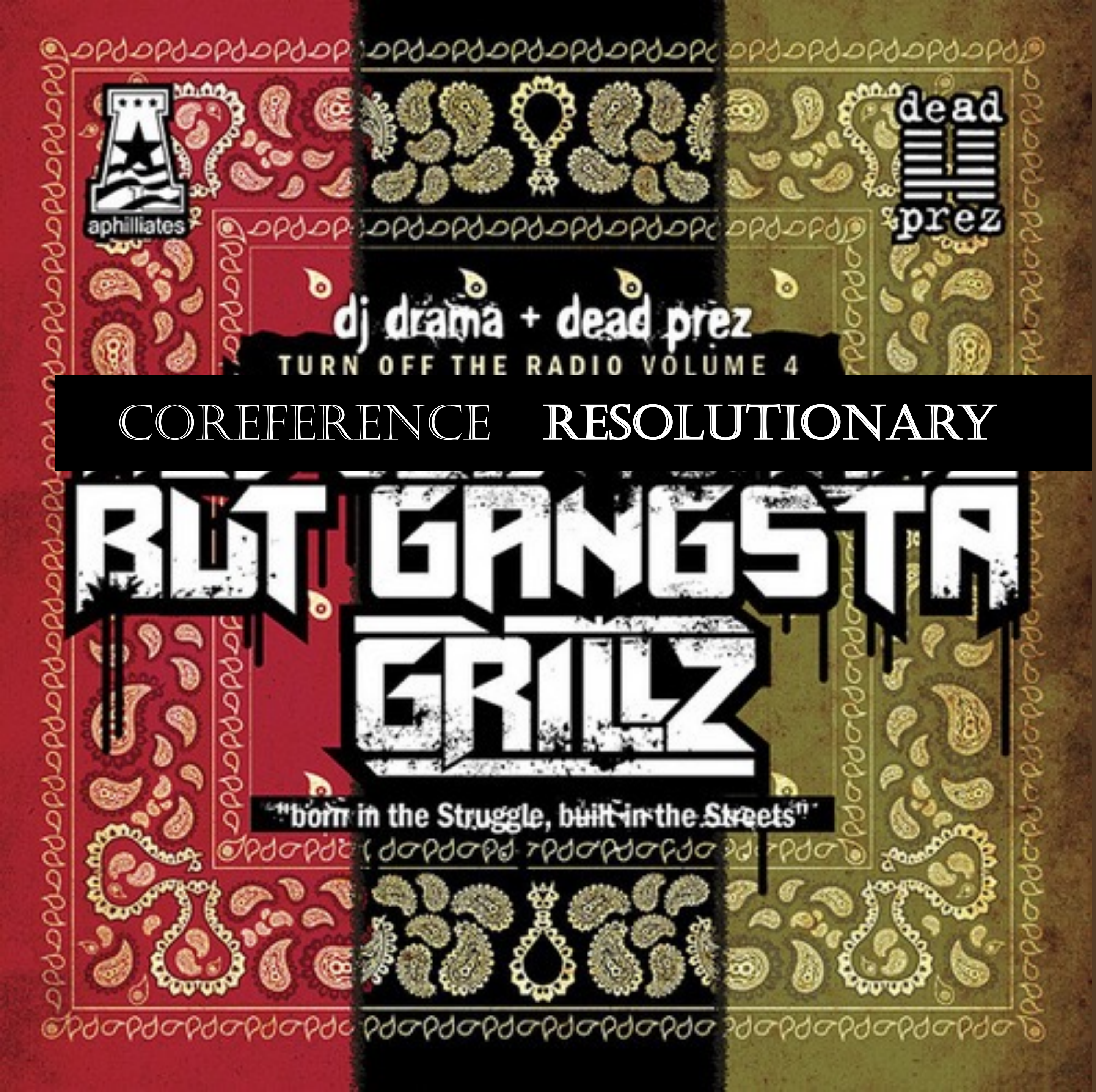
Lecture 16: Coreference Resolution

Determining who is who and what is what

Harvard IACS

Chris Tanner

COREFERENCE RESOLUTIONARY



"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

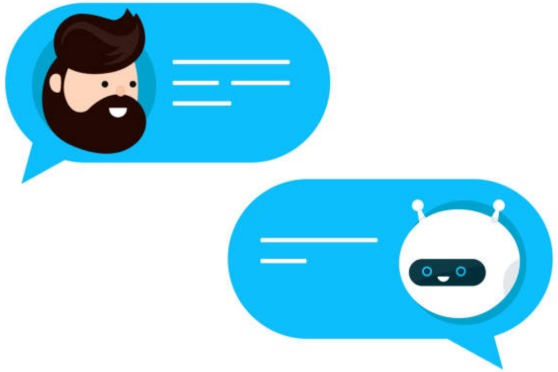
- Coreference Resolution
  - Conjoined CNN
  - Neural Clustering
  - Results
- Improvements
  - Leveraging Data
  - No Data
  - Better Data
- Additional Research

Outline

- Coreference Resolution
  - Conjoined CNN
  - Neural Clustering
  - Results
- Improvements
  - Leveraging Data
  - No Data
  - Better Data
- Additional Research

Discourse

These systems hinge upon understanding *what* you're saying (discourse) and the *meaning* of it (semantics)



Discourse

Also necessary for information retrieval, question-answering, document summarization, etc

"TL;DR crypto stocks are surging"

Google Translate

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

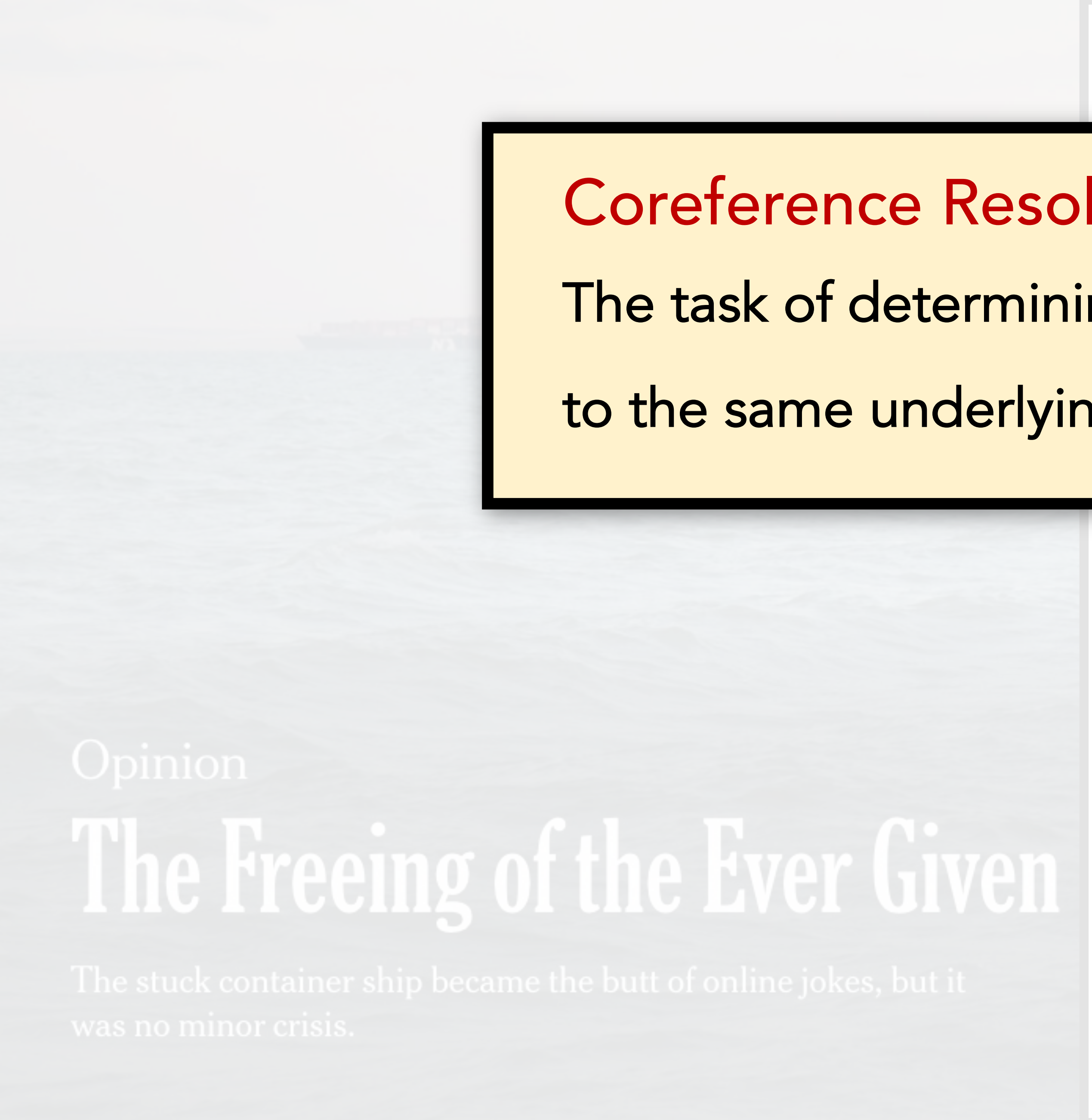
The New York Times

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.

Coreference Resolution

The task of determining which words all refer to the same underlying real-world *thing*



The New York Times

By Serge Schmemann

April 1, 2021

… succeeded … could not, … barge out of … h 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*

EASY FOR HUMANS

State-of-the-art neural model?

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

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.

State-of-the-art neural model?

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

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.

HARD FOR COMPUTERS

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.

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

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.

AP

By SAMY MAGDY and JON GAMBRELL

March 30, 2021

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

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.

AP

By SAMY MAGDY and JON GAMBRELL

March 30, 2021

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?

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.

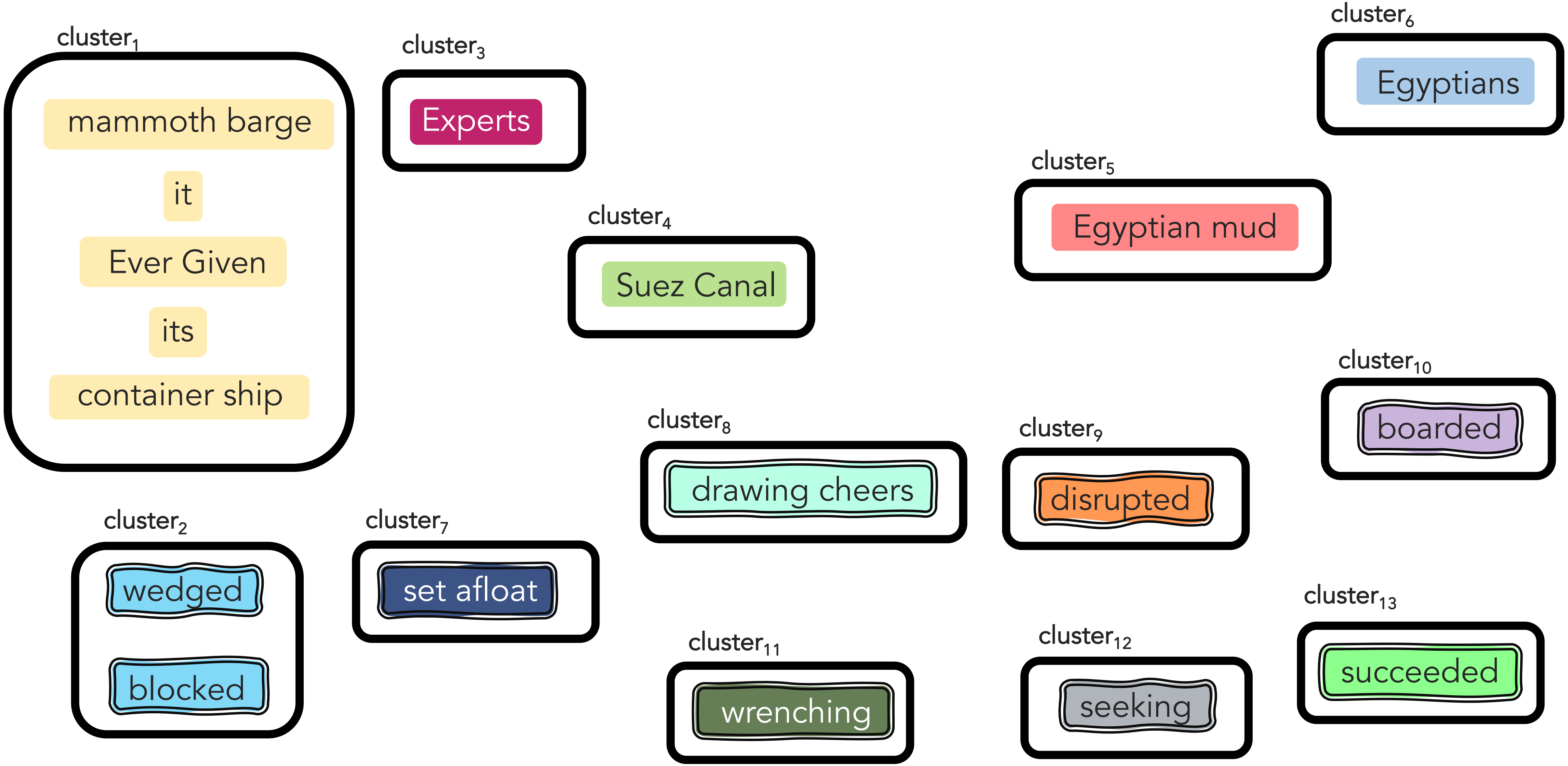
AP

By SAMY MAGDY and JON GAMBRELL

March 30, 2021

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?

The New York Times AP



Takeaway #1

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

Entity Coreference (2010 – present)

Early research demonstrated highly-effective rule-based entity coref systems

CoNLL F1: 58.3

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

Table 1: The sieves in our system; sieves new to this paper are in bold.

A Multi-Pass Sieve for Coreference Resolution. Raghunathan et al. EMNLP 2010

Stanford's Multi-Pass Sieve Coreference Resolution System. Lee et al. CoNLL 2011

Entity Coreference (2010 – present)

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 …

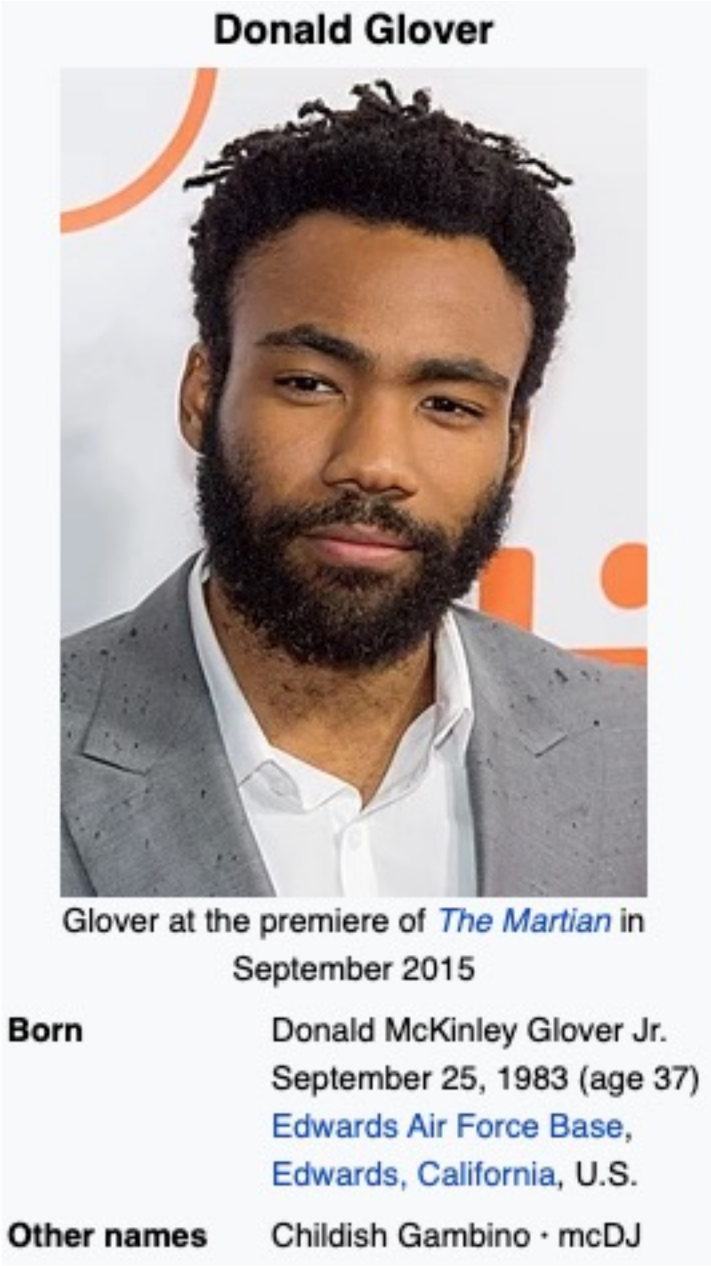
A Multi-Pass Sieve for Coreference Resolution. Raghunathan et al. EMNLP 2010

Stanford's Multi-Pass Sieve Coreference Resolution System. Lee et al. CoNLL 2011

Entity Coreference (2010 – present)

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.



A Multi-Pass Sieve for Coreference Resolution. Raghunathan et al. EMNLP 2010

Stanford's Multi-Pass Sieve Coreference Resolution System. Lee et al. CoNLL 2011

Entity Coreference (2011 – present)

Then, many systems threw **tons** of manually-defined features into their models

CoNLL F1: 65.3

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

Features describing m_j, a candidate antecedent		
1	PRONOUN_1	Y if m_j is a pronoun; else N
2	SUBJECT_1	Y if m_j is a subject; else N
3	NESTED_1	Y if m_j is a nested NP; else N
Features describing m_k, the mention to		
4	NUMBER_2	SINGULAR or PLU
5	GENDER_2	MALE, FEMALE, common first nam
6	PRONOUN_2	Y if m_k is a pron
7	NESTED_2	Y if m_k is a neste
8	SEMCLASS_2	the semantic class NIZATION, DATE, mined using Word nizer (Finkel, Gre
9	ANIMACY_2	Y if m_k is determ recognizer; else N
10	PRO_TYPE_2	the nominative c feature value for
Features describing the relationship be the mention to be resolved		
11	HEAD_MATCH	C if the mentions
12	STR_MATCH	C if the mentions
13	SUBSTR_MATCH	C if one mention
14	PRO_STR_MATCH	C if both mention
15	PN_STR_MATCH	C if both mention
16	NONPRO_STR_MATCH	C if the two men string; else I
17	MODIFIER_MATCH	C if the mentions don't have a mod
18	PRO_TYPE_MATCH	C if both mention or different only not pronominal; e
19	NUMBER	C if the mention number for one o
20	GENDER	C if the mentions for one or both m
21	AGREEMENT	C if the mentions in both number a
22	ANIMACY	C if the mention animacy for one o
23	BOTH_PRONOUNS	C if both mention
24	BOTH_PROPER_NOUNS	C if both mention else NA
25	MAXIMALNP	C if the two ment tion; else I
26	SPAN	C if neither ment
27	INDEFINITE	C if m_k is an inde else I
28	APPOSITIVE	C if the mentions
29	COPULAR	C if the mentions are in a copular construction; else I

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 is 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.

Entity Coreference (2011 – present)

Then, m

tons of

features

Takeaway #2

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

Strong results but clear limitations.

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

Event Coreference (2014 - present)

ECB+ corpus has 982 short documents

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!

Original word	Lemmatization
running	run
ran	run

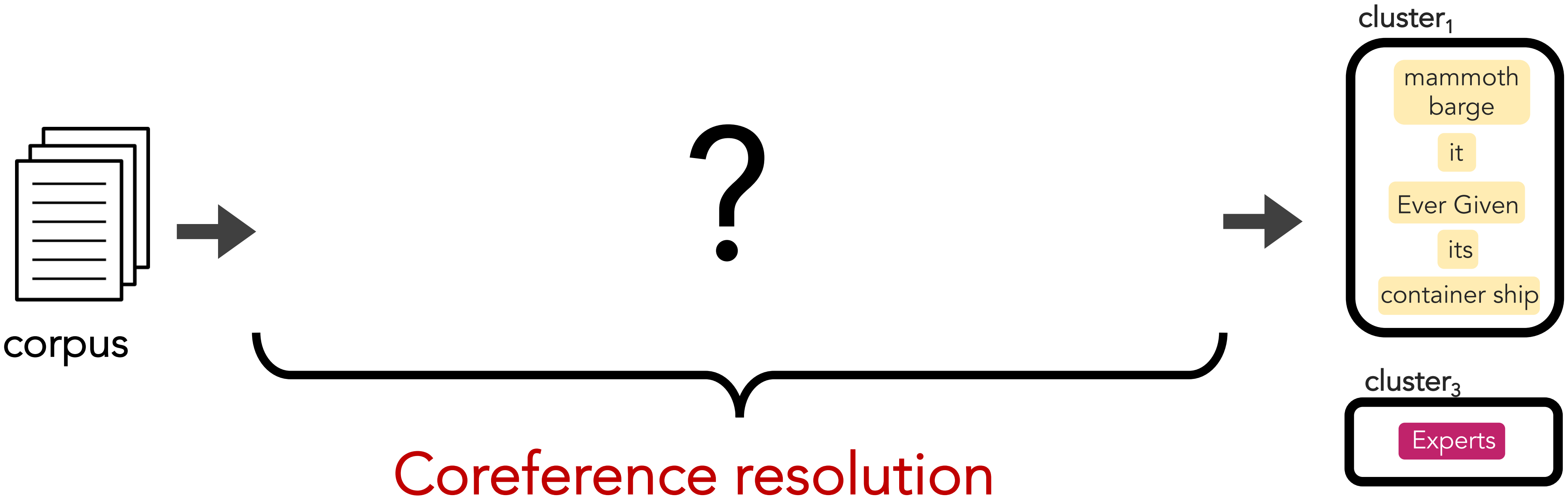
This shouldn't work so well, but it does.

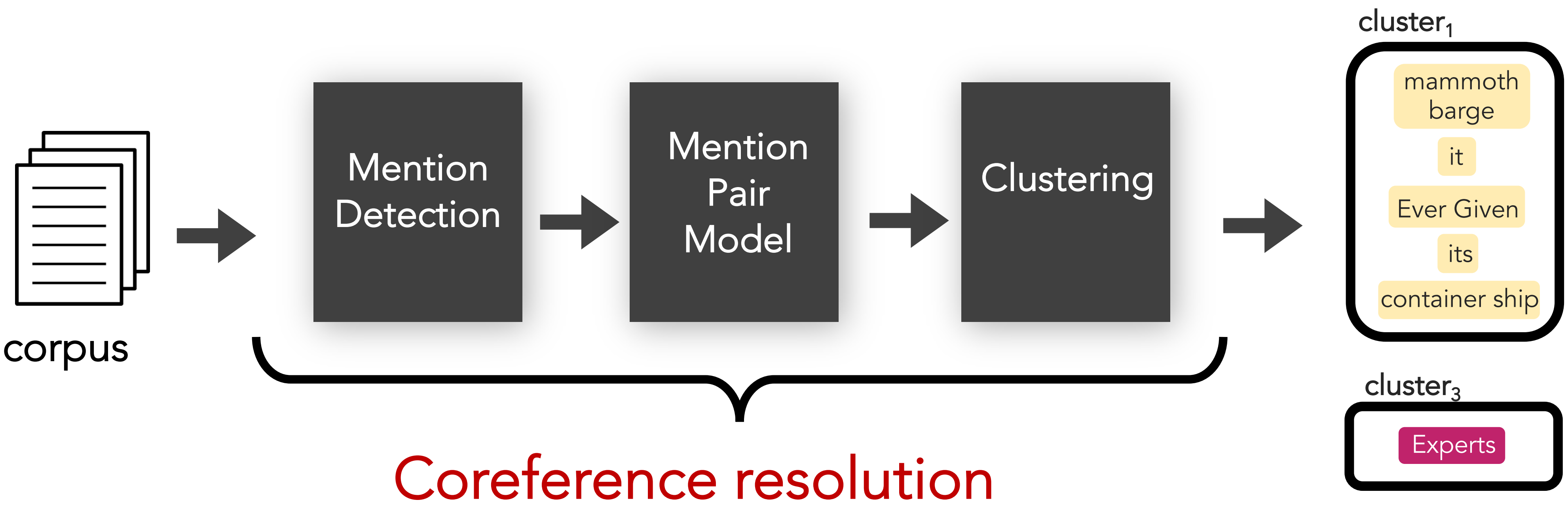
Outline

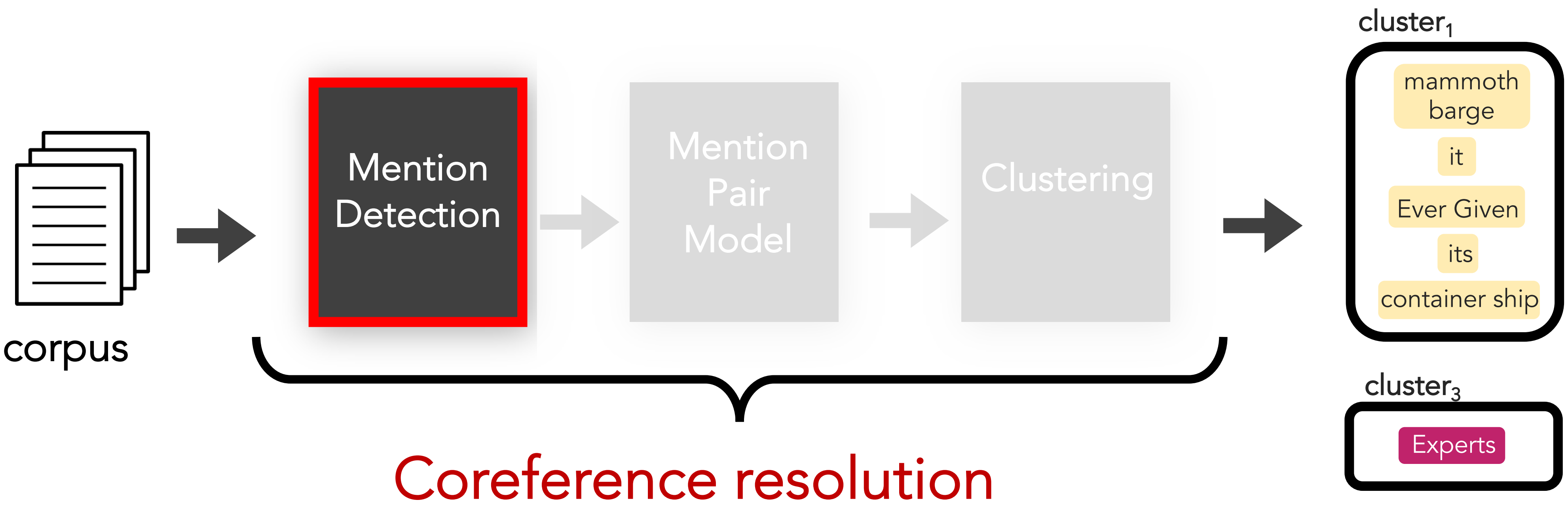
- Coreference Resolution
  - Conjoined CNN
  - Neural Clustering
  - Results
- Improvements
  - Leveraging Data
  - No Data
  - Better Data
- Additional Research

Outline

- Coreference Resolution
  - Conjoined CNN
  - Neural Clustering
  - Results
- Improvements
  - Leveraging Data
  - No Data
  - Better Data
- Additional Research


corpus
?
Coreference resolution
cluster$_1$
mammoth barge
it
Ever Given
its
container ship
cluster$_3$
Experts


corpus
Mention Detection
Mention Pair Model
Clustering
Coreference resolution
cluster$_1$
mammoth barge
it
Ever Given
its
container ship
cluster$_3$
Experts


corpus
Mention Detection
Mention Pair Model
Clustering
Coreference resolution
cluster$_1$
mammoth barge
it
Ever Given
its
container ship
cluster$_3$
Experts

Mention Detection

Determines which spans of words constitute a mention

4.6 Magnitude Quake Recorded in Sonoma County

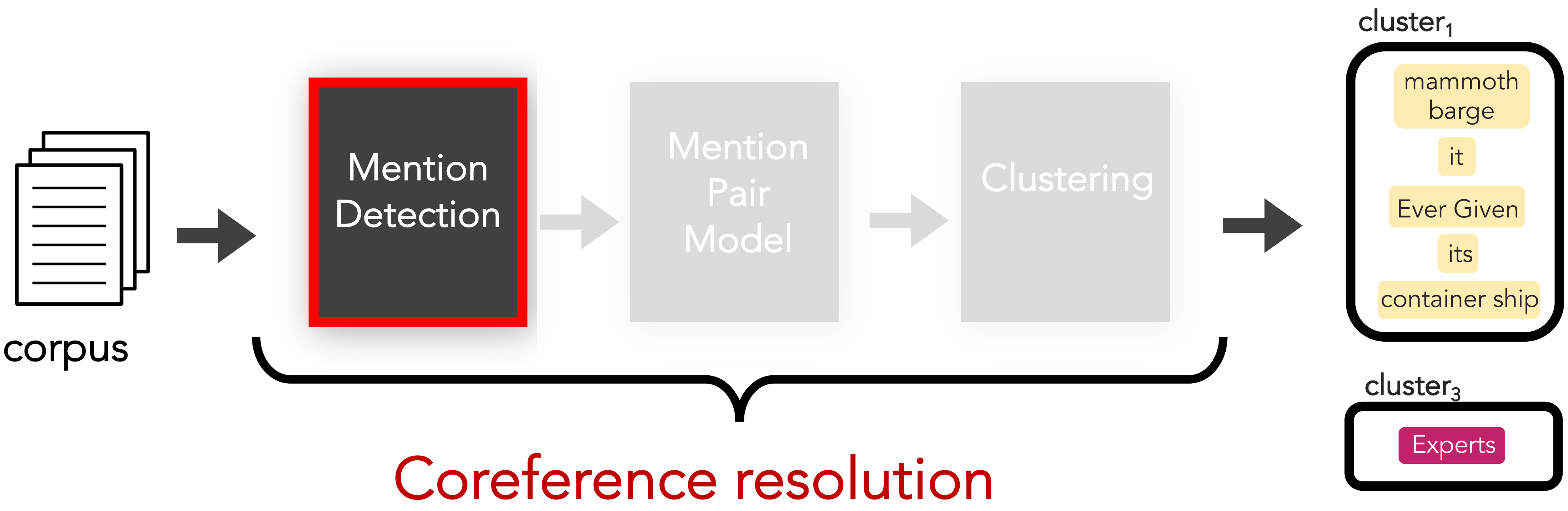
An earthquake with a preliminary magnitude of 4.6 was recorded in the North Bay this morning, according to the U.S. Geological Survey. The quake occurred at 2:09 a.m. about 14 miles north-northeast of Healdsburg and had a depth of 1.2 miles. ~~It was followed by a 2.9 aftershock at 2:12 a.m. and a 2.2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

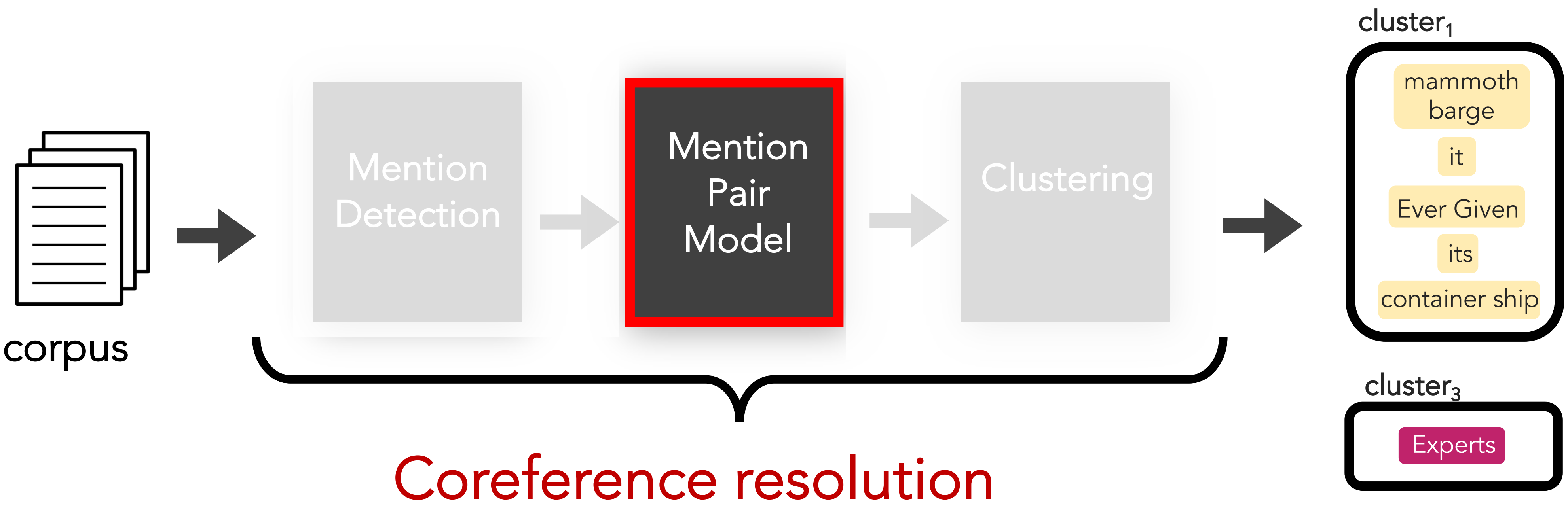
Doc 1

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.~~

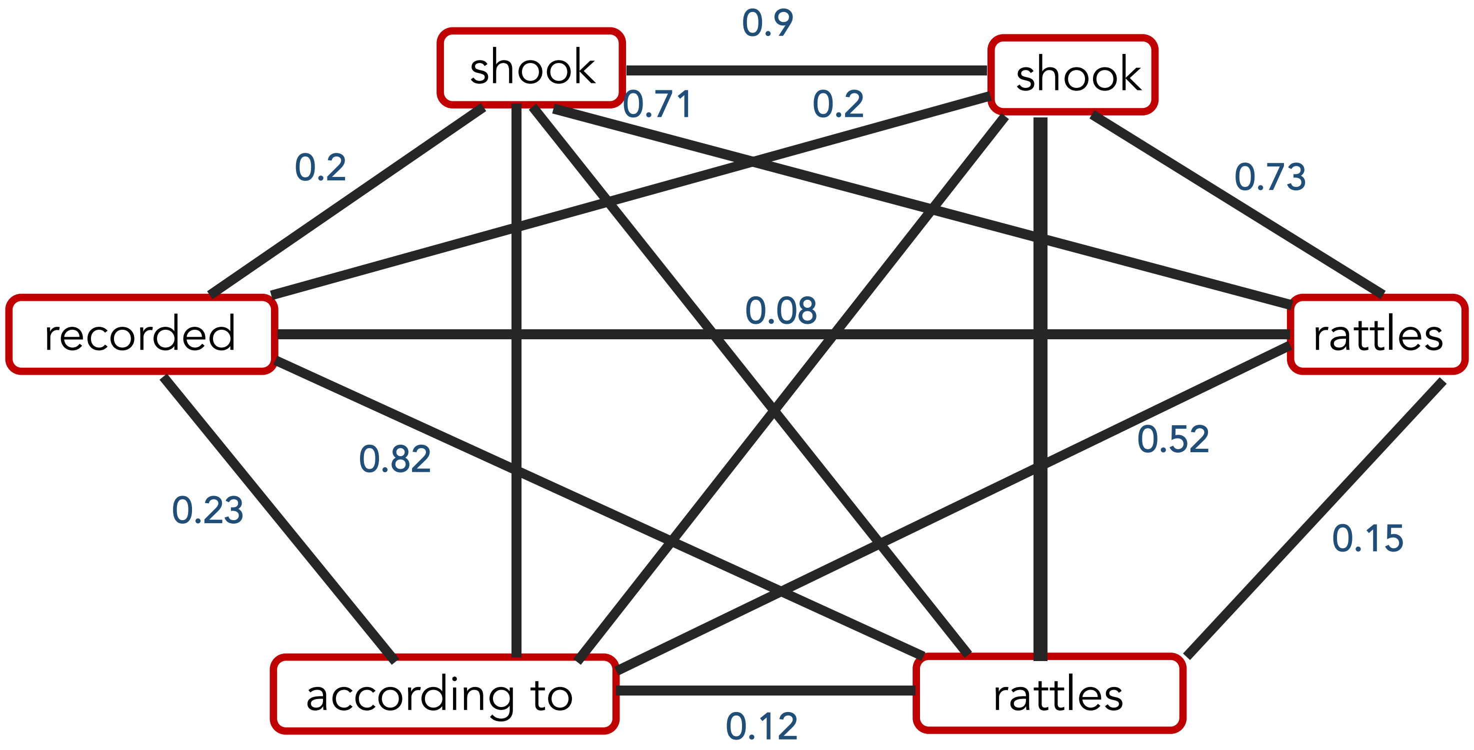
Doc 2


corpus
Mention Detection
Mention Pair Model
Clustering
cluster1
mammoth barge
it
Ever Given
its
container ship
cluster3
Experts
Coreference resolution


corpus
Mention Detection
Mention Pair Model
Clustering
$cluster_1$
mammoth barge
it
Ever Given
its
container ship
$cluster_3$
Experts
Coreference resolution

Mention Pair Model

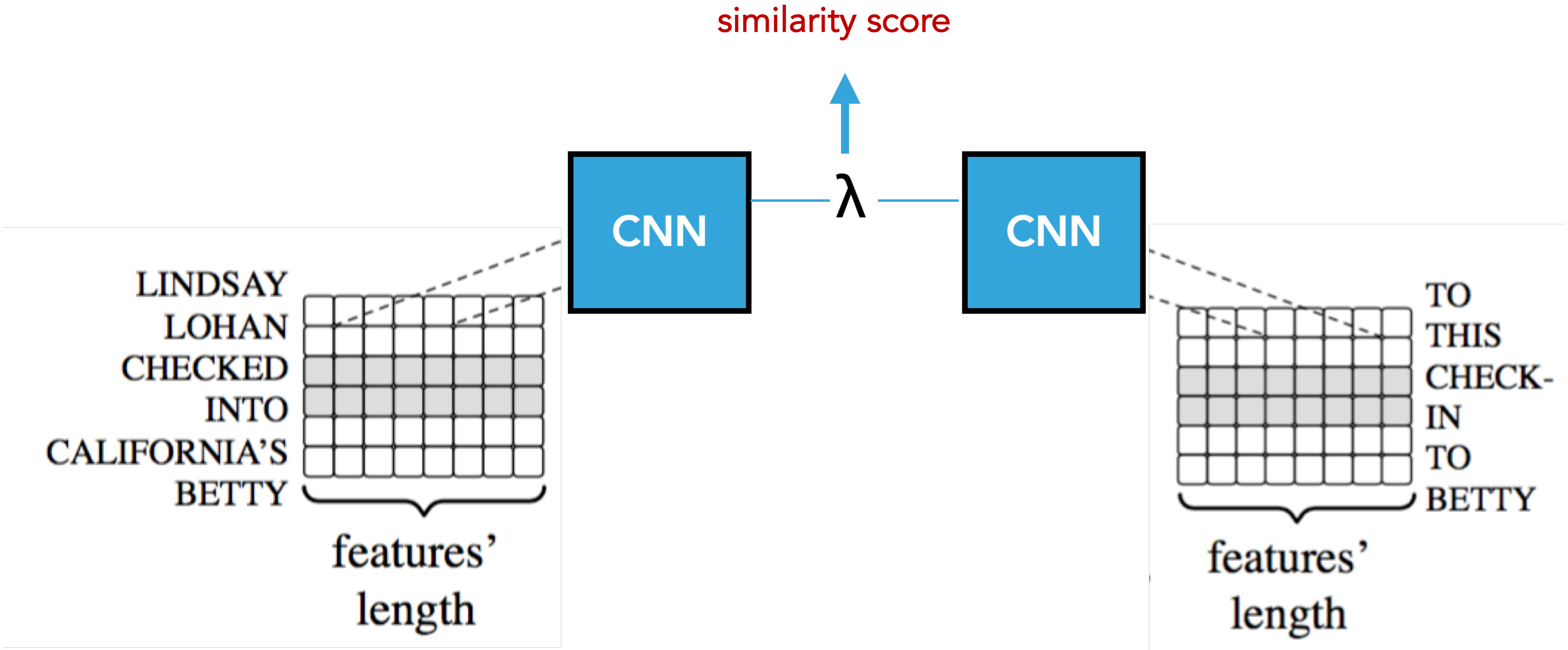
Calculates a coref probability for all pairs of mentions



Conjoined CNN

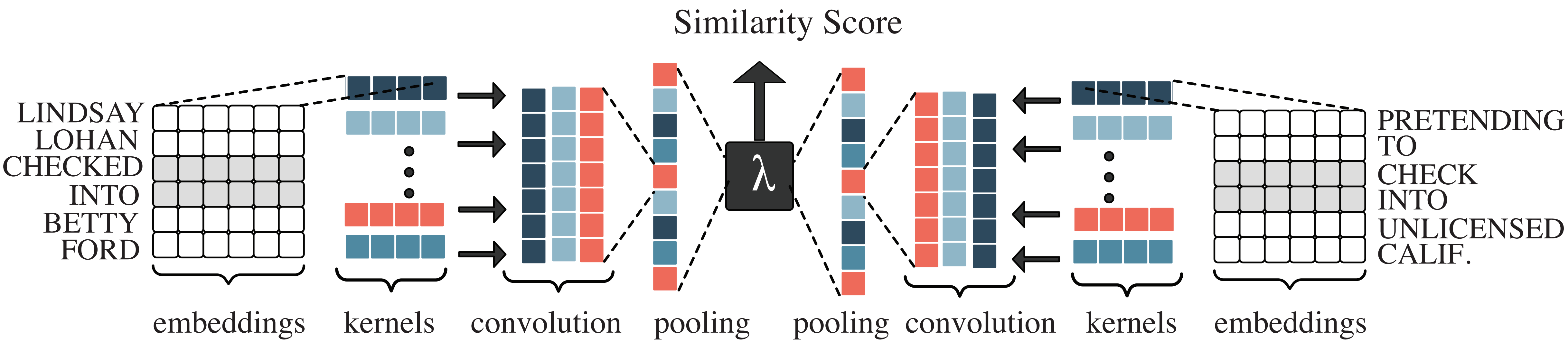
Feature	# dimensions
Word embeddings	300
Lemma embeddings	300
Dependency Parse embeddings	400
Character embeddings	100
Part-of-speech embeddings	300

Conjoined CNN



Cross-Document Coreference Resolution for Entities and Events. Tanner. Brown University Dissertation. 2019

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

Cross-Document Coreference Resolution for Entities and Events. Tanner. Brown University Dissertation. 2019

Conjoined CNN

We predict these should coref as pairs

m_{17}, m_2	0.0	erupted	erupted
m_{17}, m_4	0.0	erupted	erupted
m_5, m_{923}	0.03	announced	announce
m_{78}, m_{57}	0.05	erupt	erupted

0.5 threshold

m_{801}, m_{39}	0.97	revealed	broke into
m_{26}, m_{48}	0.98	handed down	confirmed

We predict these should NOT coref as pairs

accuracy: **92.4**

precision: **55.8**

recall: **71.2**

f1: **62.8**

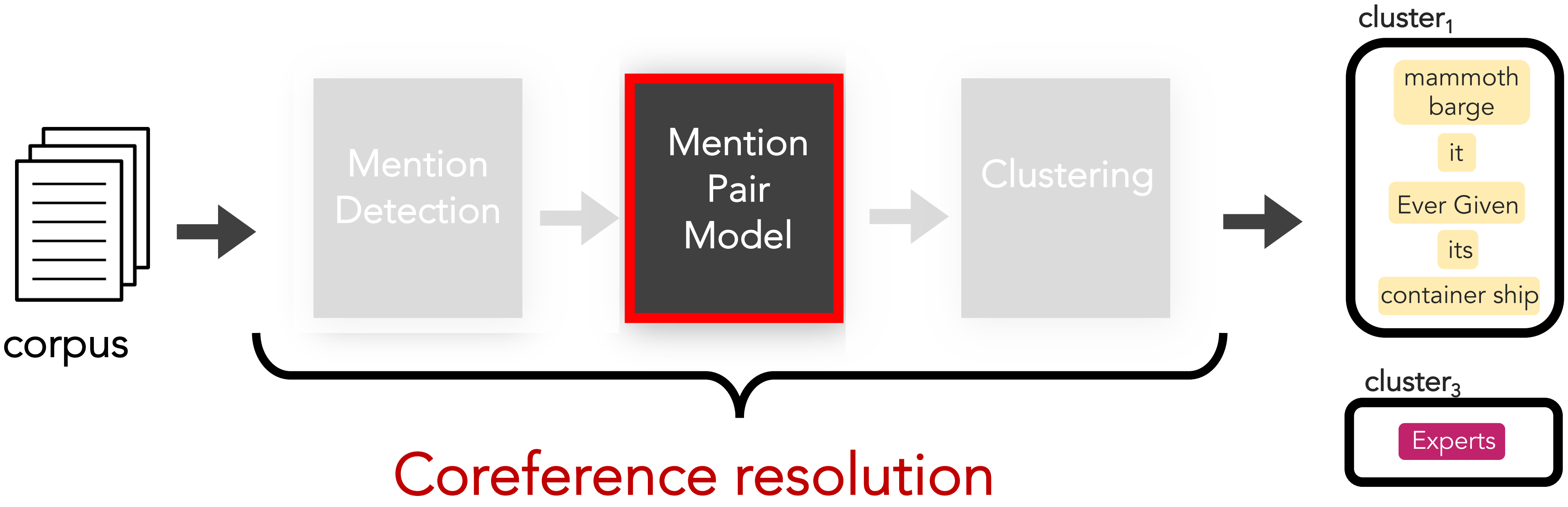
Conjoined CNN

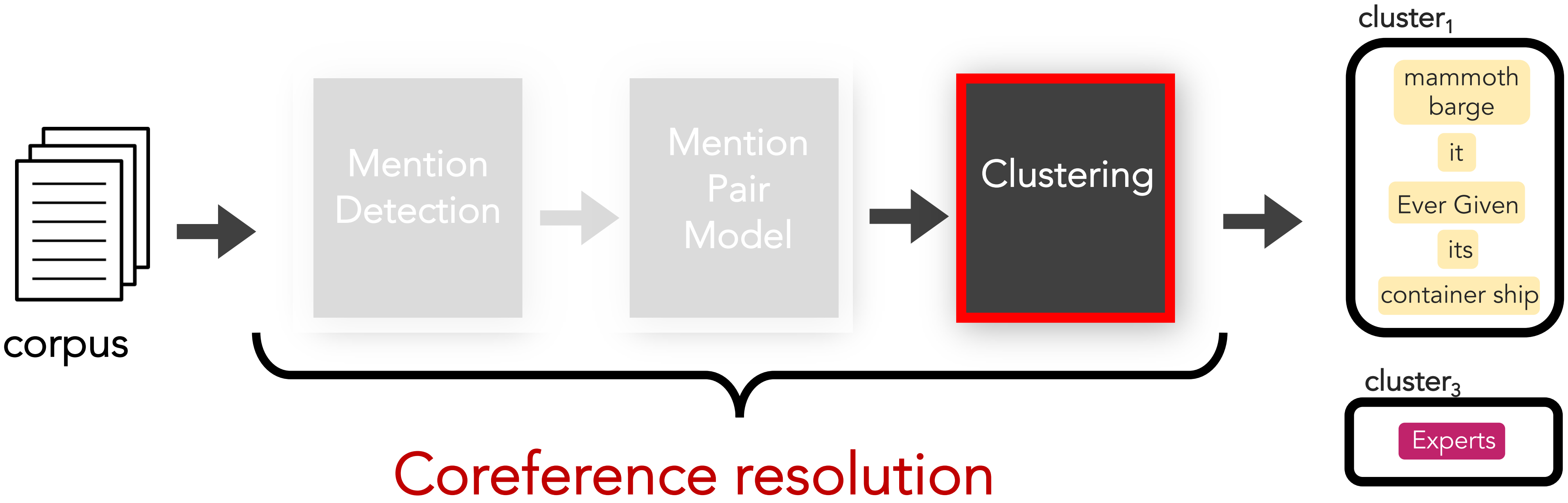
Development Set Results

	Precision	Recall	F1
	Within-Document		
SameLemma	53.9	48.0	50.8
LibSVM	51.2	52.0	51.6 (0.01)
FFNN	50.3	59.8	54.6 (0.5)
CCNN	51.5	68.2	**58.7 (0.8)**
	Cross-Document		
SameLemma	55.6	54.1	54.8
LibSVM	58.6	59.1	58.8 (0.02)
FFNN	55.3	62.0	58.5 (0.6)
CCNN	55.8	71.2	**62.8 (0.6)**

LibSVM and FFNN received same features as CCNN,
plus relational features (e.g., cosine sim., dot-product, WordNet)

Cross-Document Coreference Resolution for Entities and Events. Tanner. Brown University Dissertation. 2019


corpus
Mention Detection
Mention Pair Model
Clustering
cluster$_1$
mammoth barge
it
Ever Given
its
container ship
cluster$_3$
Experts
Coreference resolution


corpus
Mention Detection
Mention Pair Model
Clustering
Coreference resolution
cluster$_1$
mammoth barge
it
Ever Given
its
container ship
cluster$_3$
Experts

Outline

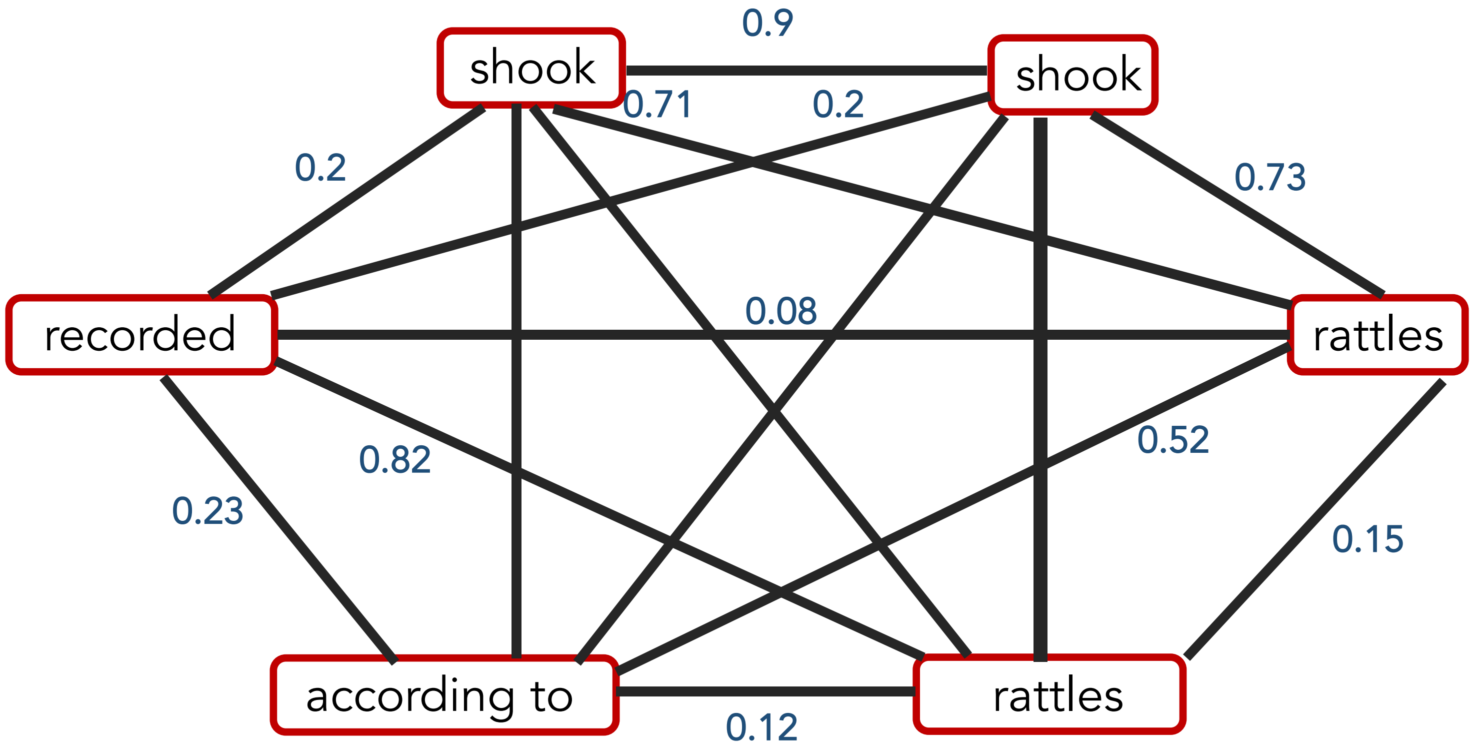
- Coreference Resolution
  - Conjoined CNN
  - Neural Clustering
  - Results
- Improvements
  - Leveraging Data
  - No Data
  - Better Data
- Additional Research

Outline

- Coreference Resolution
  - Conjoined CNN
  - Neural Clustering
  - Results
- Improvements
  - Leveraging Data
  - No Data
  - Better Data
- Additional Research

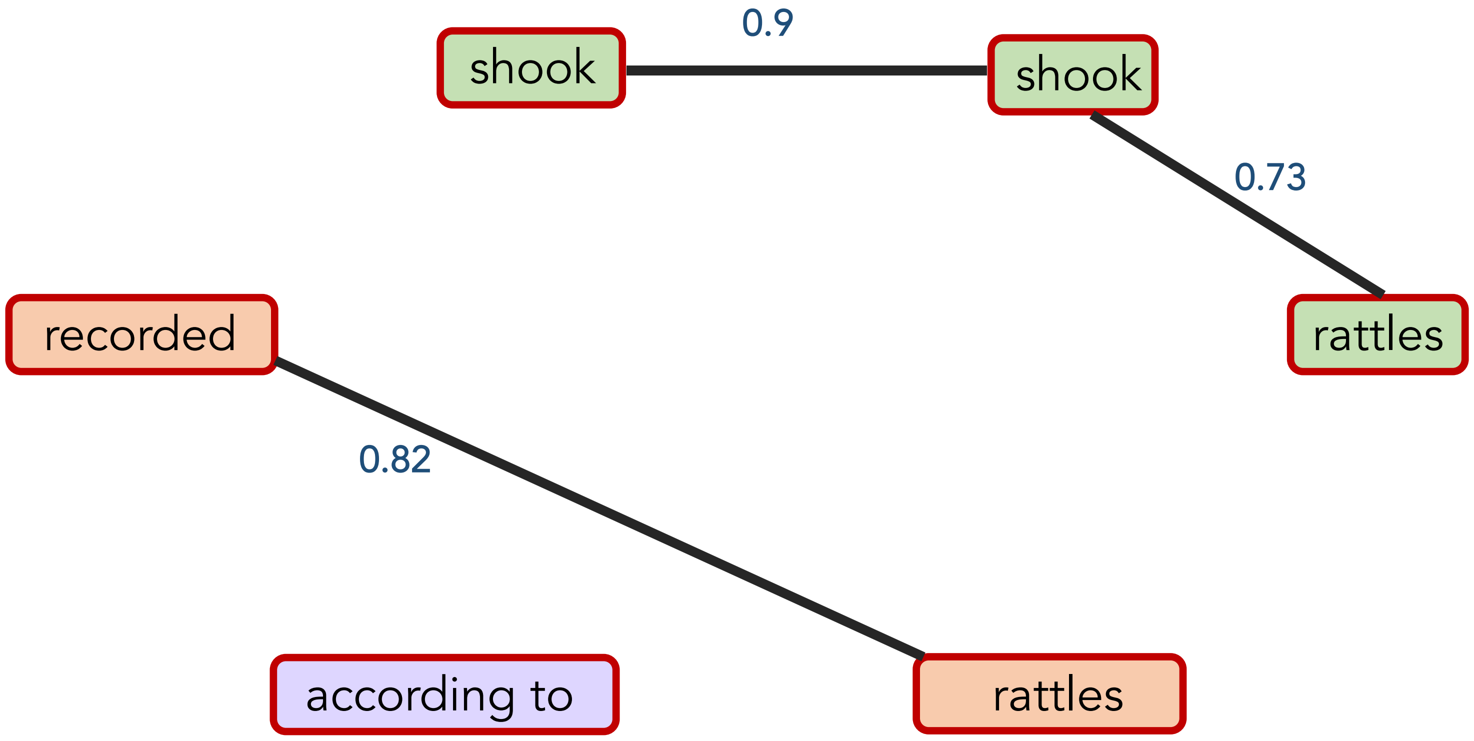
Clustering

Goal is to return clusters from the fully-connected graph



Clustering

Goal is to return clusters from the fully-connected graph



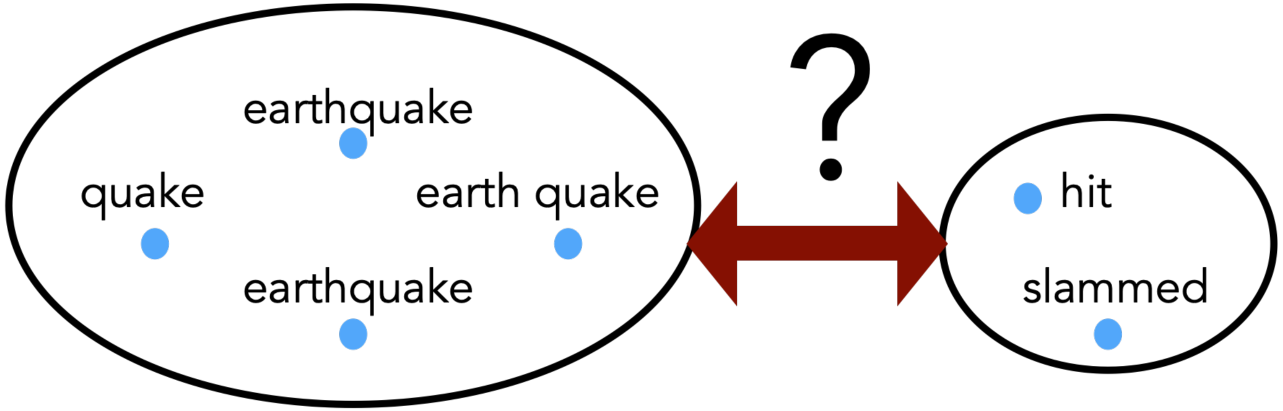
Clustering

Nearly 100% of past systems simply performed agglomerative clustering.

We want:

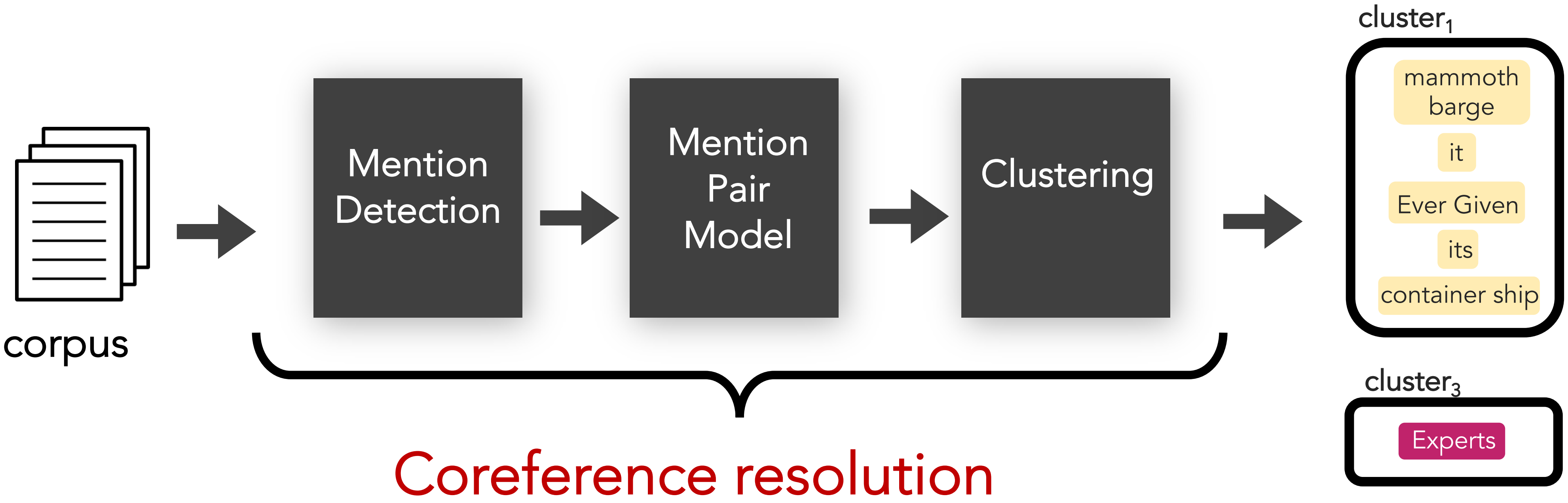
- More holistic, **cluster-to-cluster** predictions
- Less sensitivity to non-uniformity across topics
- No additional stopping parameter
- Prevention against an all-subsuming cluster

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: $\frac{\|C_x\| + \|C_y\|}{\sum_z \|C_z\|}$

Cross-Document Coreference Resolution for Entities and Events. Tanner. Brown University Dissertation. 2019


corpus
Mention Detection
Mention Pair Model
Clustering
Coreference resolution
cluster1
mammoth barge
it
Ever Given
its
container ship
cluster3
Experts

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

Mention Detection

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)

ECB+ Development Performance

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
POS	■		■	■	■	■	■	■	■	■	■	■				
Lemma	■	■		■	■	■	■	■	■				■	■	■	
Dependency	■	■	■		■	■	■			■	■		■	■		■
Character	■	■	■	■		■		■		■		■	■		■	■
Word	■	■	■	■	■				■		■	■		■	■	■
Within-Doc	79.9	79.6	78.2	79.9	79.1	78.9	78.6	**83.2**	81.0	76.6	77.1	78.5	80.7	80.6	**82.6**	77.8
Cross-Doc	72.0	73.8	73.5	77.5	73.2	73.7	73.2	**78.0**	76.6	72.0	73.2	76.5	72.6	75.5	**78.5**	71.5

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
POS	■	■	■	■							■				
Lemma	■				■	■	■					■			
Dependency		■			■			■	■				■		
Character			■			■		■		■				■	
Word				■			■		■	■					■
Within-Doc	78.5	79.3	81.0	78.1	82.3	**84.9**	82.8	79.6	77.8	**83.7**	77.2	**84.5**	80.6	**83.8**	81.9
Cross-Doc	75.7	73.4	78.4	76.6	74.2	**79.5**	75.0	74.1	71.4	**78.4**	74.7	**79.0**	63.2	**75.6**	74.9

■ = Feature Used   = Best Performances

Lemma + Character Embeddings yields the best performance

Using Predicted Mentions

	Within-Document				Cross-Document			
	MUC	B^3	CEAF	CoNLL F1	MUC	B^3	CEAF	CoNLL F1
SameLemma$_{any}$	40.4	66.4	66.2	57.7	66.7	51.4	46.2	54.8
HDDCRP [108]	53.4	75.4	71.7	66.8	73.1	53.5	49.5	58.7
Choubey [20]	62.6	72.4	71.8	68.9	73.4	61.0	56.5	63.6
FFNN+AGG	61.6	73.6	69.1	68.1 (0.14)	74.8	55.3	60.2	63.4 (0.21)
FFNN+NC	62.5	73.2	70.8	68.8 (0.17)	76.1	56.0	60.4	64.2 (0.18)
CCNN+AGG	65.2	74.2	69.0	69.5 (0.16)	75.8	55.8	62.7	64.8 (0.21)
CCNN+NC	67.3	73.3	69.6	70.1 (0.20)	77.2	56.3	62.0	65.2 (0.22)
CCNN+NC (ensemble)	**67.7**	**73.6**	**69.8**	**70.4 (0.13)**	**78.1**	**56.6**	**62.1**	**65.6 (0.17)**

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 ( ).

Cross-Document Coreference Resolution for Entities and Events. Tanner. Brown University Dissertation. 2019

Using Gold Mentions

	Within-Document				Cross-Document			
	MUC	B^3	CEAF	CoNLL F1	MUC	B^3	CEAF	CoNLL F1
Test Set: ECB+ Gold Mentions								
SameLemma	58.3	83.0	75.9	72.4	84.2	68.2	48.0	66.8
FFNN+AGG	59.9	85.6	78.4	74.6	77.7	69.9	50.1	65.9
FFNN+NC	60.7	86.7	79.4	75.6	74.9	67.8	56.3	67.0
CCNN+AGG	70.5	89.1	83.5	81.0	84.1	70.7	55.5	70.1
CCNN+NC	70.9	88.9	83.6	**81.2**	86.4	71.7	59.1	**72.4**

Cross-Document Coreference Resolution for Entities and Events. Tanner. Brown University Dissertation. 2019

CCNN + Clustering

FINDINGS

- State-of-the-art for event coref
- Contextualized representations
- More holistic clustering
- **Char + Lemma Embeddings** were the only two necessary features

Errors

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 …

CCNN + Clustering

False P

Sony ar

Friday,

Takeaway #3 The community needs a better corpus.

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

…

False N

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

The UN Refugee Agency on Friday strongly condemned the aerial bombing of …

Outline

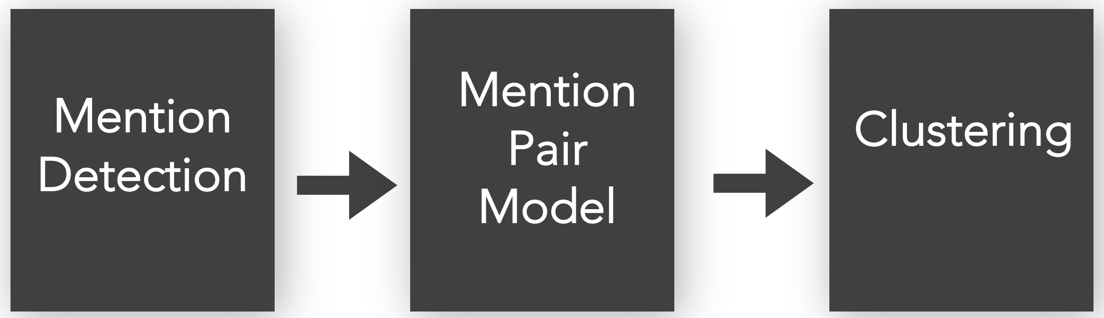
- Coreference Resolution
  - Conjoined CNN
  - Neural Clustering
  - Results
- Improvements
  - Leveraging Data
  - No Data
  - Better Data
- Additional Research

Outline

- Coreference Resolution
  - Conjoined CNN
  - Neural Clustering
  - Results
- Improvements
  - Leveraging Data
  - No Data
  - Better Data
- Additional Research

Entity Coreference

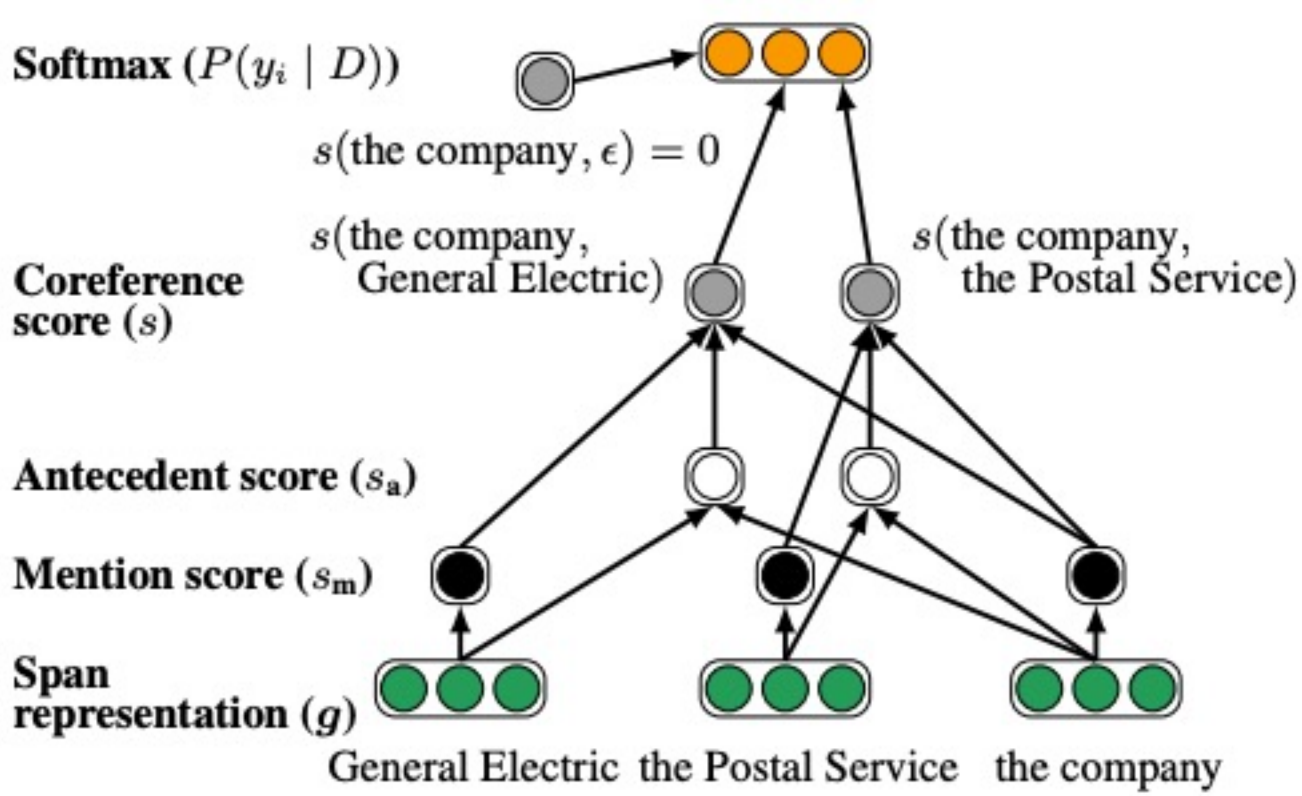
End-to-end neural systems



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

Entity Coreference

End-to-end



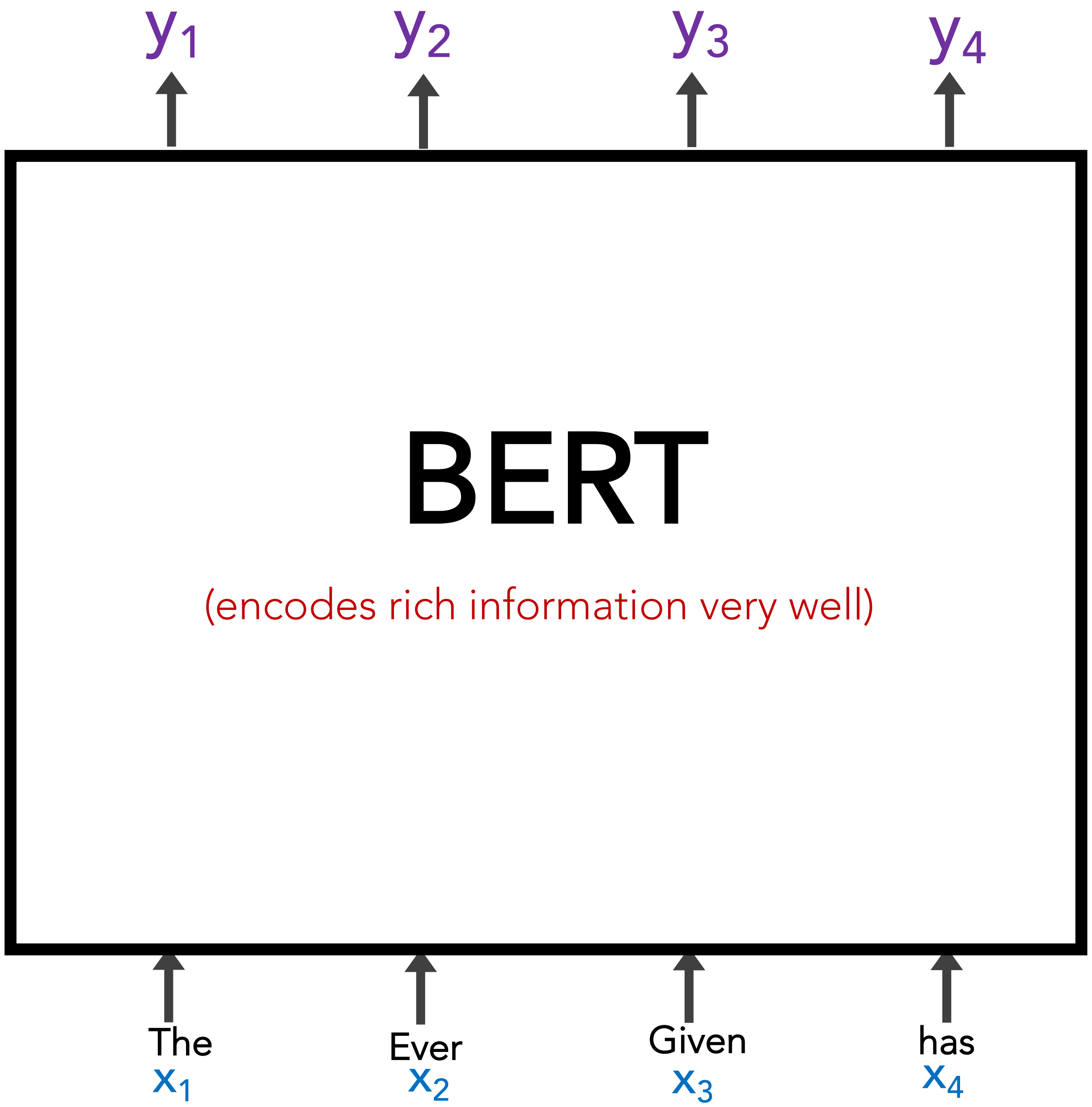
End-to-end Neural Coreference Resolution. Lee et al. EMNLP 2017.

Entity Coreference

Uses several important features

	Avg. F1	Δ
Our model (ensemble)	69.0	+1.3
Our model (single)	67.7	
– distance and width features	63.9	-3.8
– GloVe embeddings	65.3	-2.4
– speaker and genre metadata	66.3	-1.4
– head-finding attention	66.4	-1.3
– character CNN	66.8	-0.9
– Turian embeddings	66.9	-0.8

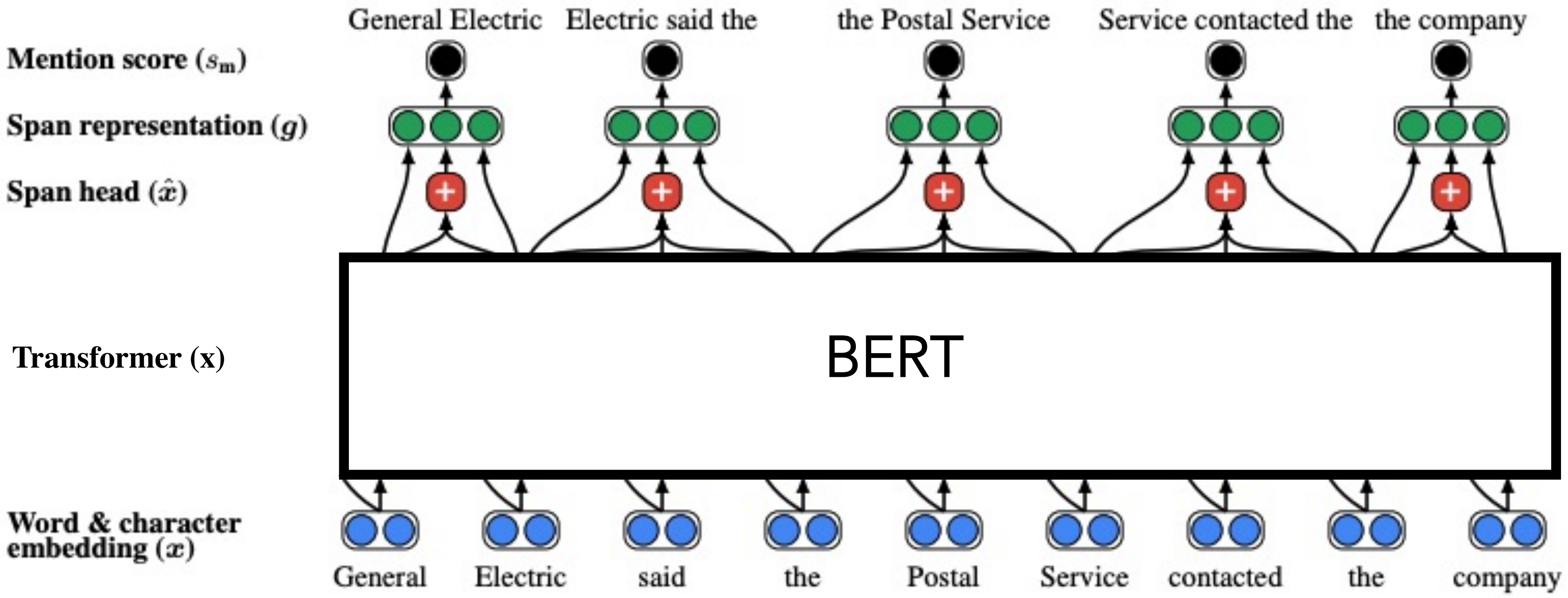
End-to-end Neural Coreference Resolution. Lee et al. EMNLP 2017.


y_1 y_2 y_3 y_4
BERT
(encodes rich information very well)
The Ever Given has
x_1 x_2 x_3 x_4

Entity Coreference

CoNLL F1: 79.6

End-to-End

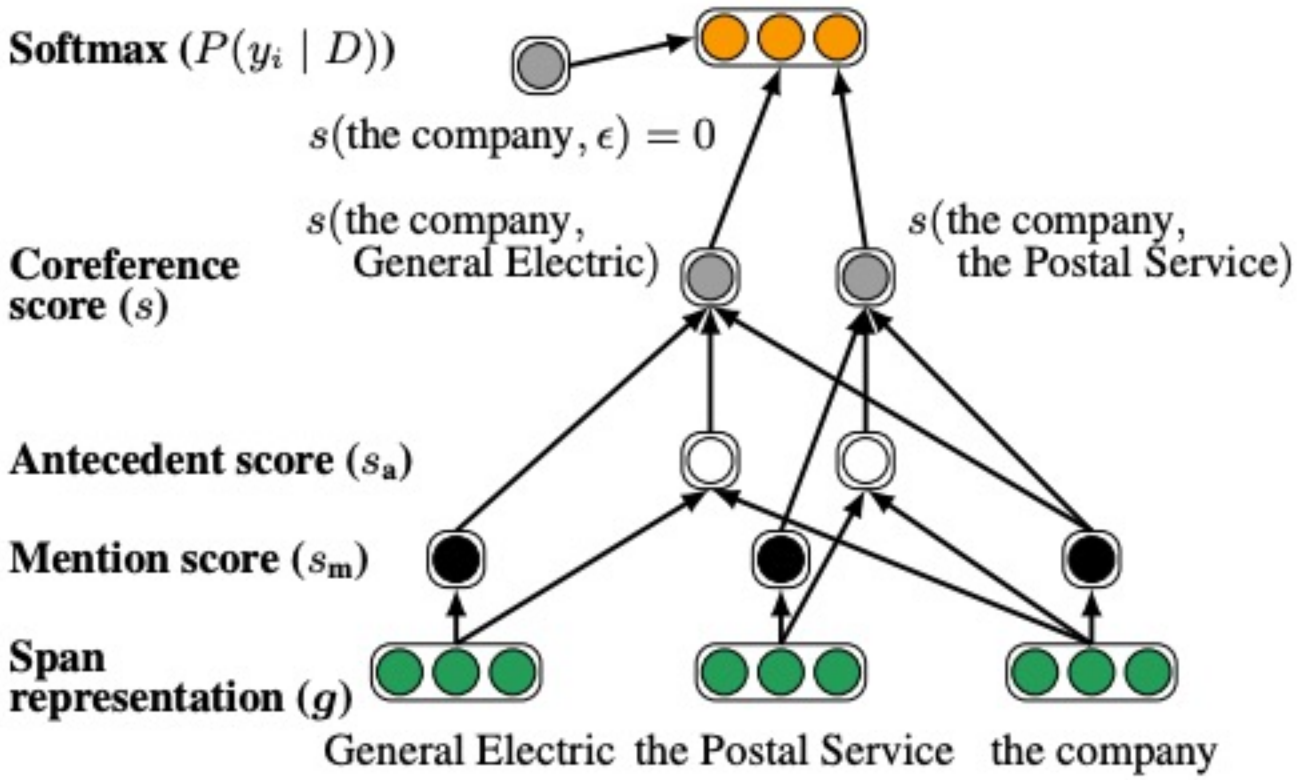


BERT for Coreference Resolution: Baselines and Analysis. Joshi et al. EMNLP 2019.

Entity Coreference

CoNLL F1: 79.6

End-to-End



BERT for Coreference Resolution: Baselines and Analysis. Joshi et al. EMNLP 2019.

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

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

Table 3: Qualitative Analysis: #base and #large refers to the number of cluster-level errors on a subset of the OntoNotes English development set. <u>Underlined</u> 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.

However, coref is still far from solved

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

Category			rge
Related Entities			7
Lexical			9
Pronouns			3
Mention Paraphrasing	And in case you missed it *the Royals* are here. Today Britain's **Prince Charles and his wife Camilla**...	14	12
Conversation	(Priscilla:) **My** mother was Thelma Wahl . She was ninety years old ... (Keith:) *Priscilla Scott* is mourning . *Her* mother Thelma Wahl was a resident ..	18	16
Misc.	He is **my**, She is **my** Goddess , ah	17	17
Total		93	74

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.

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.

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 #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.

INSIGHTS

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

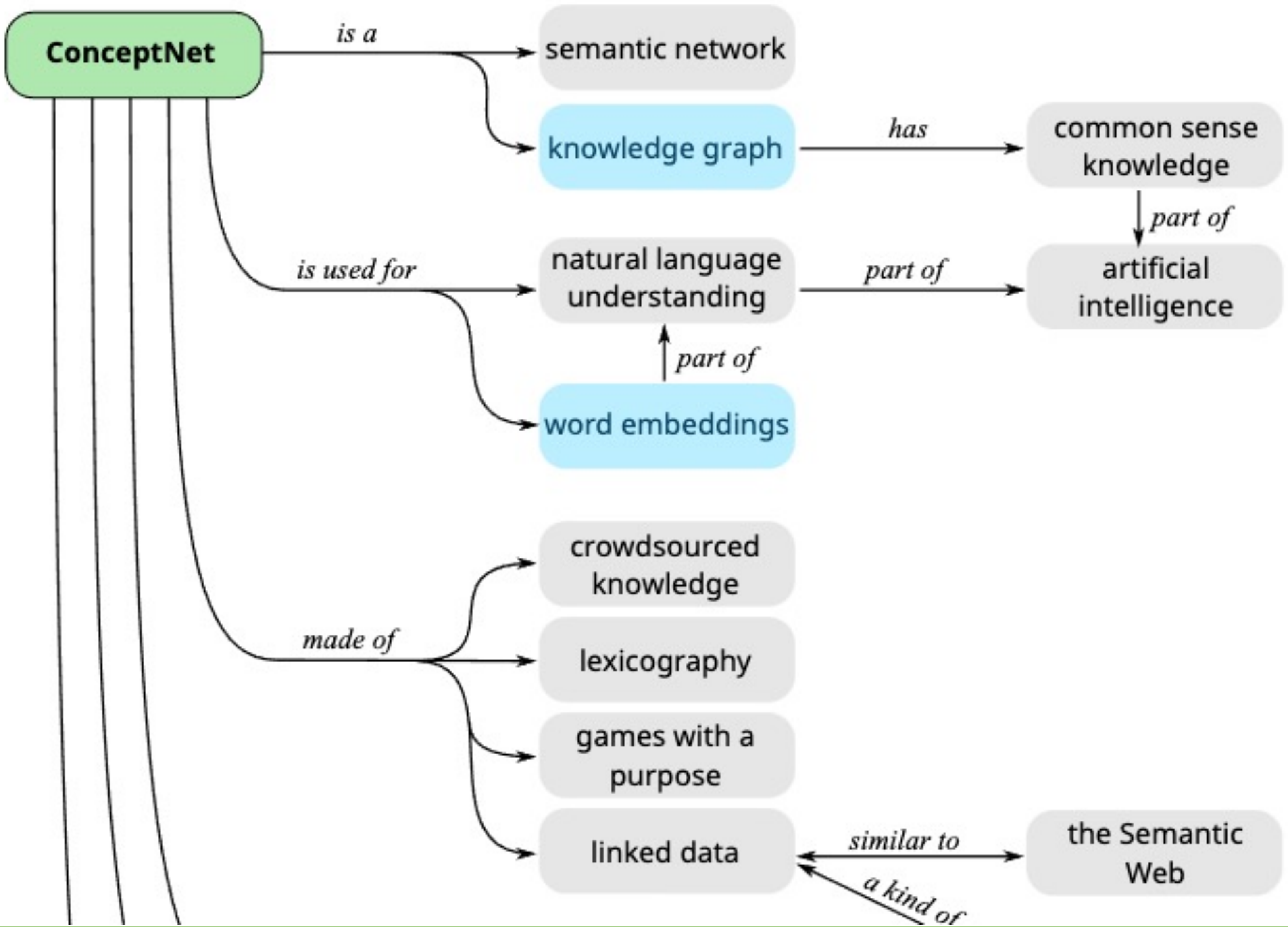
- Coreference Resolution
  - Conjoined CNN
  - Neural Clustering
  - Results
- Improvements
  - Leveraging Data
  - No Data
  - Better Data
- Additional Research

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

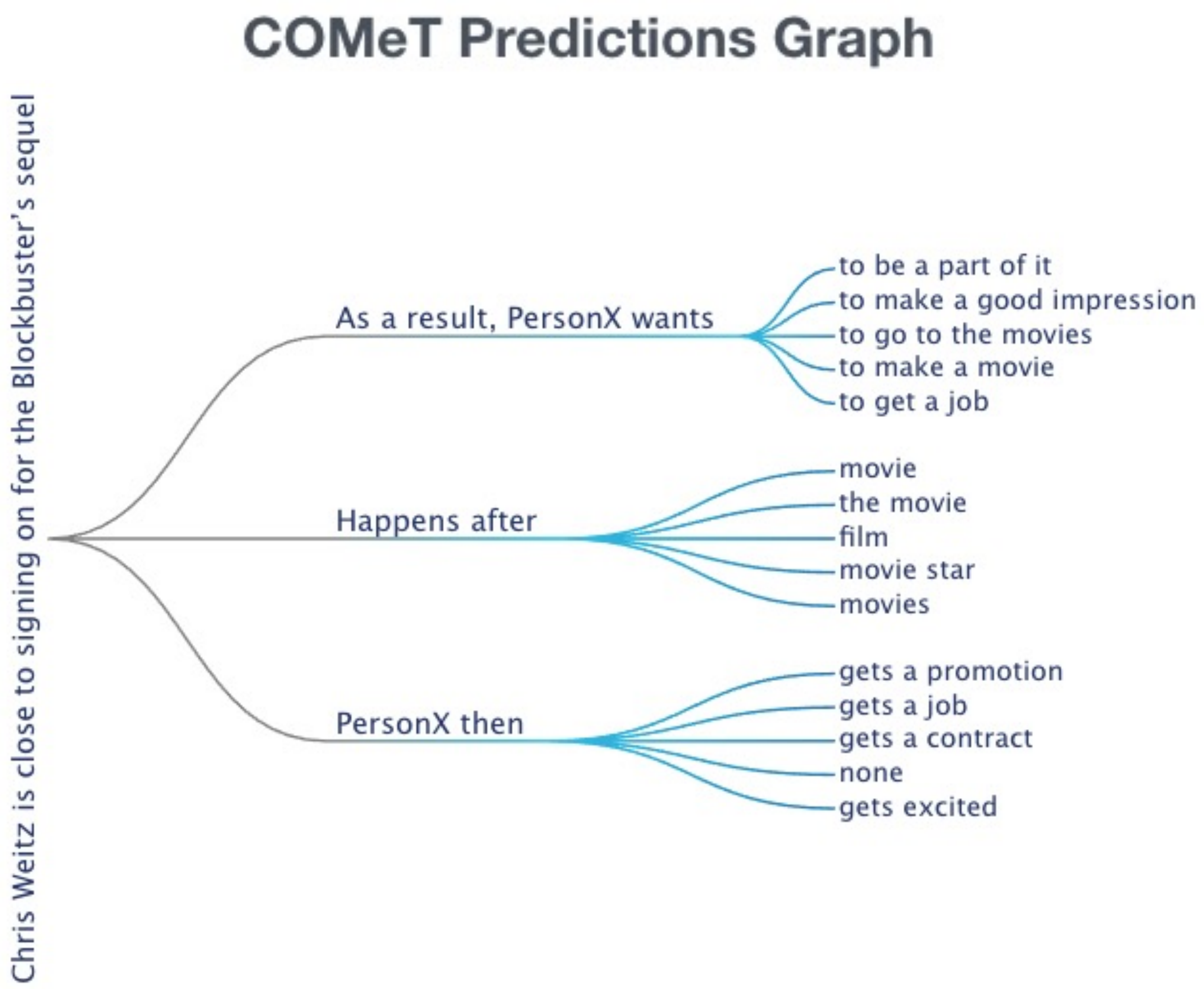
Sahithya Ravi
UBC
PhD Student

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

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

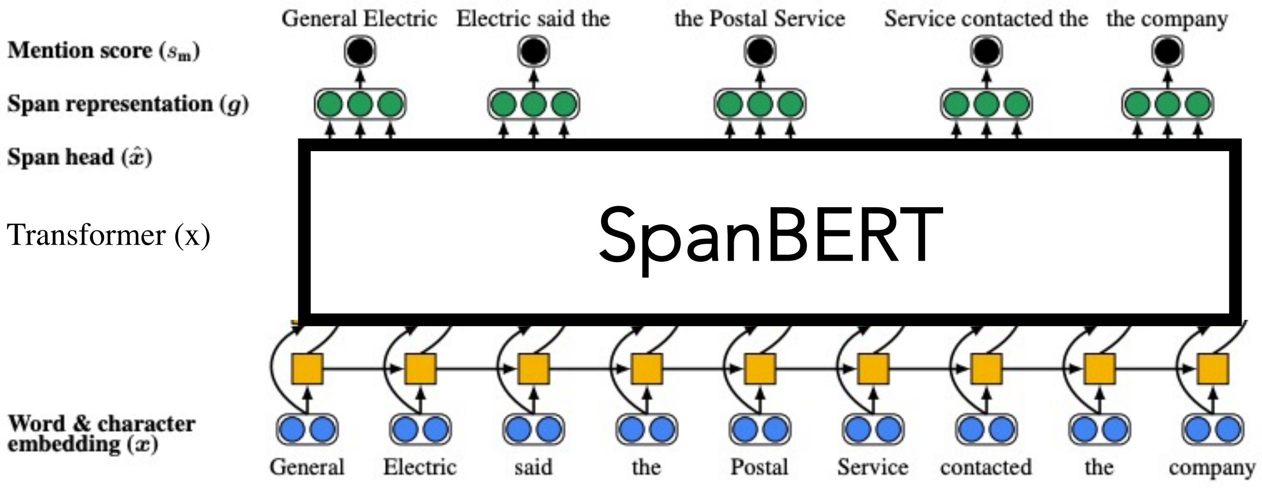
Ordered sieves

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

with the SOTA BERT-based end-to-end model



A Simple and Effective Approach for Unsupervised Coreference using Pretrained Representations and Weak Supervision from Linguistic Rules. Alessandro Stolfo et al. In preparation.

We combine the old school, manual rule-based system

Ordered sieves

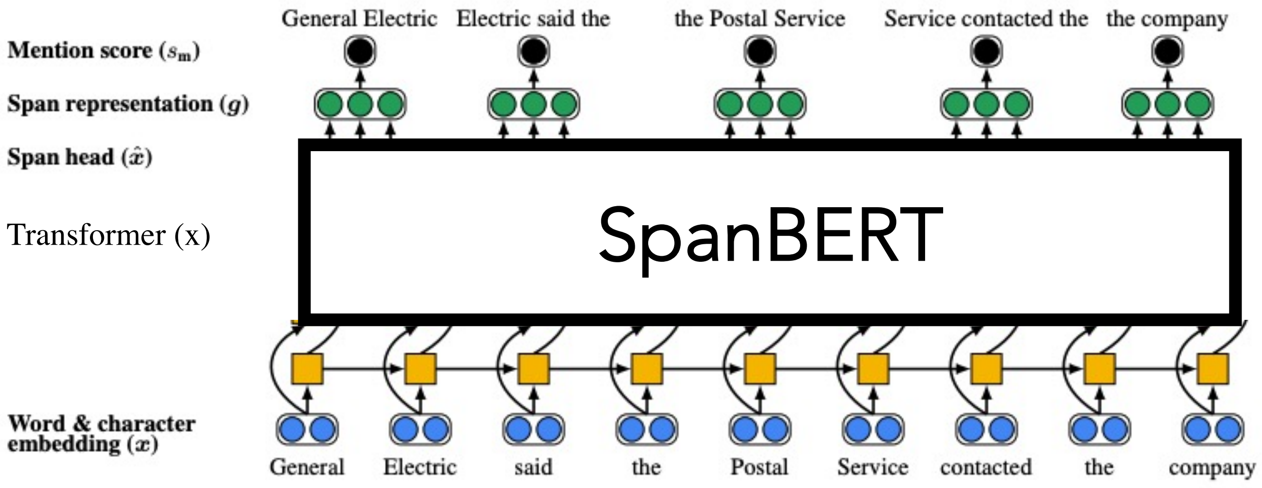
1. **Mention Detection Sieve**
2. **Discourse Processing Sieve**
3. Exact String Match Sieve
4. **Relaxed String Match Sieve**
5. Precise Constructs Sieve (e.g., appositives)

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)

with the SOTA BERT-based end-to-end model



Supervised
(needs training data)

A Simple and Effective Approach for Unsupervised Coreference using Pretrained Representations and Weak Supervision from Linguistic Rules. Alessandro Stolfo et al. In preparation.

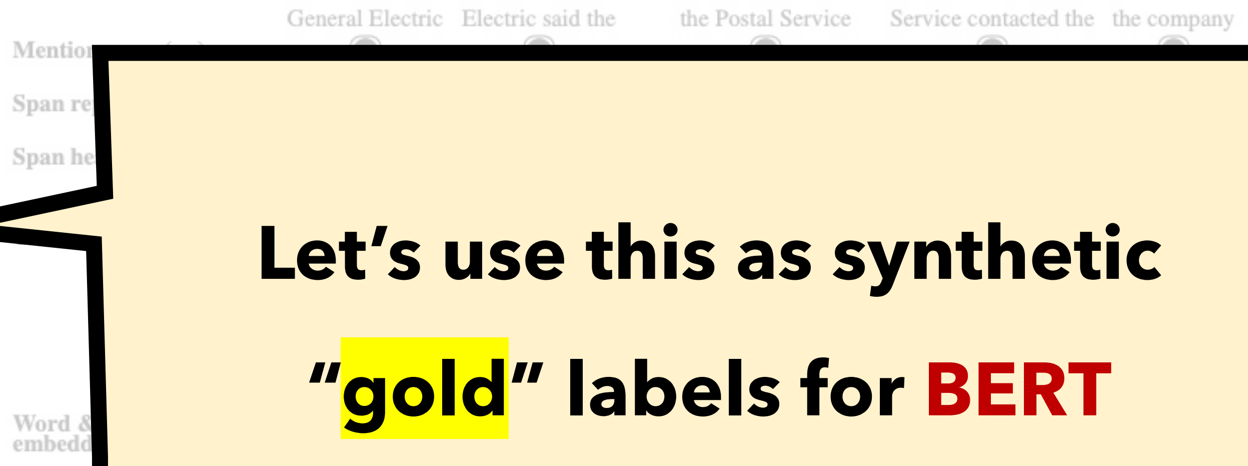
We combine the old school, manual rule-based system

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)



Let's use this as synthetic "gold" labels for BERT

Supervised
(needs training data)

A Simple and Effective Approach for Unsupervised Coreference using Pretrained Representations and Weak Supervision from Linguistic Rules. Alessandro Stolfo et al. In preparation.

CONCERN

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

A Simple and Effective Approach for Unsupervised Coreference using Pretrained Representations and Weak Supervision from Linguistic Rules. Alessandro Stolfo et al. In preparation.

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

A Simple and Effective Approach for Unsupervised Coreference using Pretrained Representations and Weak Supervision from Linguistic Rules. Alessandro Stolfo et al. In preparation.

	MUC			B^3			$CEAF_{\phi_4}$			CoNLL
	P	R	F_1	P	R	F_1	P	R	F_1	F_1
Stanford (Lee et al., 2011)	64.3	65.2	64.7	49.2	56.8	52.7	52.5	46.6	49.4	55.6
Multigraph (Martschat, 2013)	-	-	65.4	-	-	54.4	-	-	50.2	56.7
Unsup. Ranking (Ma et al., 2016)	-	-	67.7	-	-	55.9	-	-	51.8	58.4
c2f-coref	65.7	68.0	66.9	50.9	59.4	54.8	52.9	49.1	50.9	57.5
BERT-base + c2f-coref	66.8	69.2	68.0	51.5	60.6	55.7	53.1	50.3	51.7	58.5
SpanBERT-base + c2f-coref	67.6	68.5	68.1	53.1	60.1	56.4	54.8	50.4	52.5	59.0
BERT-large + c2f-coref	67.2	69.7	68.5	52.3	61.2	56.4	54.0	51.0	52.5	59.1
SpanBERT-large + c2f-coref	67.4	69.8	**68.6**	52.4	61.8	**56.7**	54.1	51.4	**52.7**	**59.3**

Table 1: Results on the test set of the English CoNLL-2012 shared task[3]. 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).

A Simple and Effective Approach for Unsupervised Coreference using Pretrained Representations and Weak Supervision from Linguistic Rules. Alessandro Stolfo et al. In preparation.

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

Conclusions

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.