Hard NLP Tasks

Determining who is who and what is what

Harvard IACS

Lecturer

Chris Tanner

Objectives

- observe a glimpse into current state-of-the-art research
- understand some of the most challenging NLP tasks
- think critically about your ML approaches to solve tasks
- feel inspired to get involved with NLP

Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

Our digital world is inundated with text.

How can we leverage it for useful tasks?

62B pages

500M tweets/day
(6k/sec)

2.6B active users

13M articles

Language is funny

“Red tape holds up new bridges”

“Hospitals are sued by 7 foot doctors”

“Local high school dropouts cut in half”

“Tesla crashed today”

“Obama announced that he will run again”

“Kipchoge announced that he will run again”

“She made him duck”

“Will you visit the bank across from the river bank? You can bank on it”

“Yes” vs “Yes.” vs “YES” vs “YES!” vs “YAS” vs “Yea”

Language is funny (coreference)

“Maria likes May”

“Maria likes May and Joe”

“Maria likes May and June”

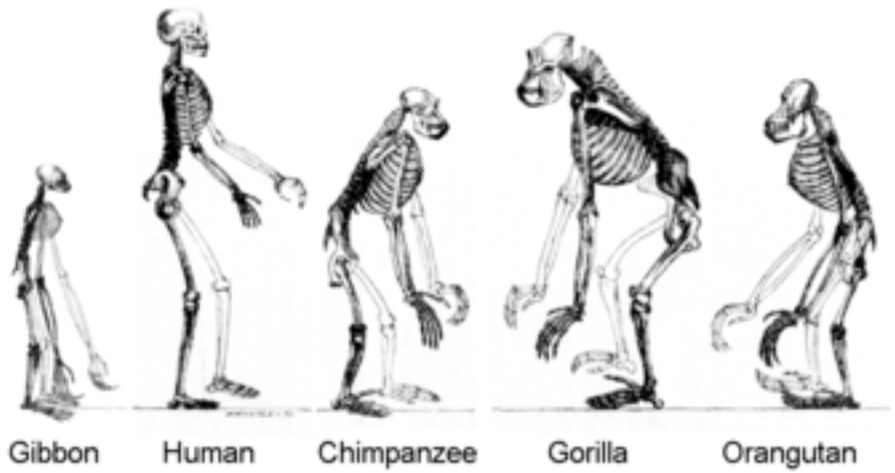
“May likes Maria”

“Maria hit May, then she [fell/ran]”

“Maria and Anqi bullied May, so they got in trouble”

“Maria and Anqi convinced May to prank the teacher, so they got in trouble”

Language is special and complex



- Distinctly human ability
- Paramount to human evolution
- Influenced by many social constructs
- Incredibly nuanced
- Language forms capture multi-dimensions
- Language evolves over time

Slide adapted from or inspired by Chris Manning and Richard Socher

Linguistic Structure

Discourse	what is said; the process underlying language
Pragmatics	how words are used to denote meaning
Semantics	the true meaning
Syntax	rules that govern language structure
Lexemes	basic unit of language (e.g., words)
Morphology	how words are formed
Characters	sub-unit representations

Linguistic Structure

Discourse

Pragmatics

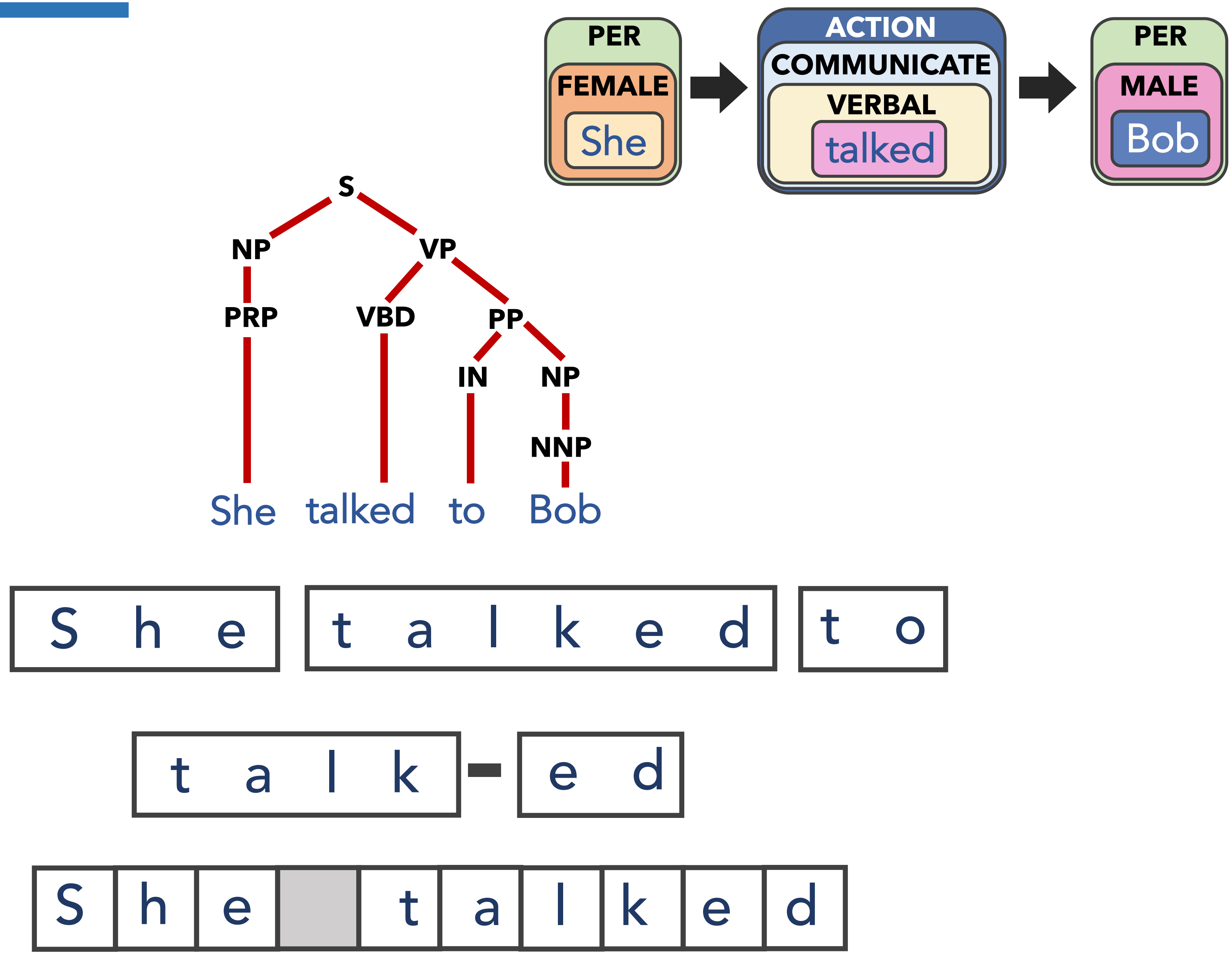
Semantics

Syntax

Lexemes

Morphology

Characters

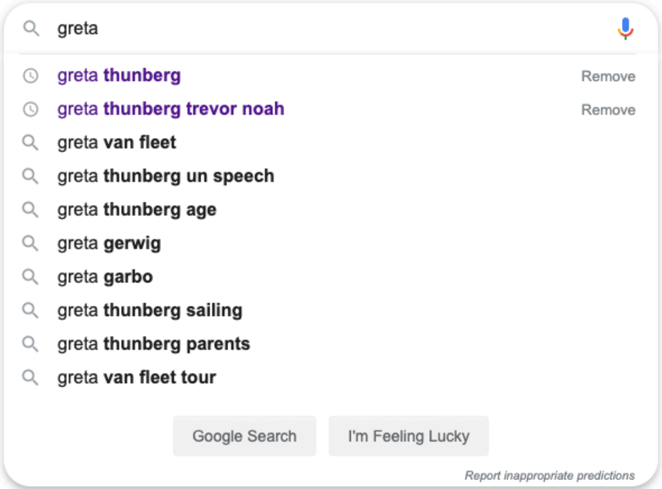


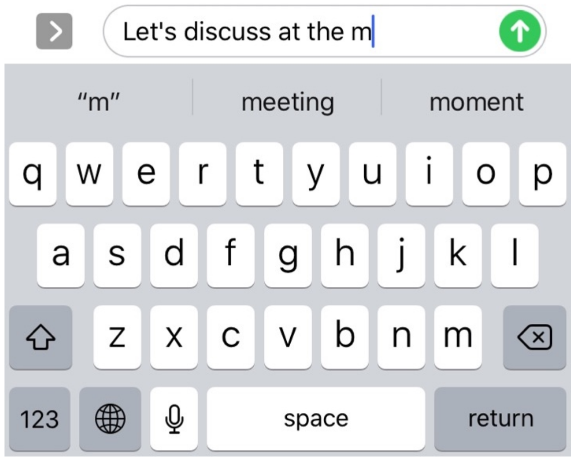
NLP in use

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

Google Translate

Google





Common NLP Tasks (aka problems)

Syntax

Morphology

Word Segmentation

Part-of-Speech Tagging

Parsing

- Constituency
- Dependency

Discourse

Summarization

Coreference Resolution

Semantics

Sentiment Analysis

Topic Modelling

Named Entity Recognition (NER)

Relation Extraction

Word Sense Disambiguation

Natural Language Understanding (NLU)

Natural Language Generation (NLG)

Machine Translation

Entailment

Question Answering

Language Modelling

Common NLP Tasks (aka problems)

Syntax

Morphology

Word Segmentation

Part-of-Speech Tagging

Parsing

- Constituency
- Dependency

Discourse

Summarization

Coreference Resolution

Semantics

Sentiment Analysis

Topic Modelling

Named Entity Recognition (NER)

Relation Extraction

Word Sense Disambiguation

Natural Language Understanding (NLU)

Natural Language Generation (NLG)

Machine Translation

Entailment

Question Answering

Language Modelling

Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions


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 $\boxed{\text{mammoth barge}}$ 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

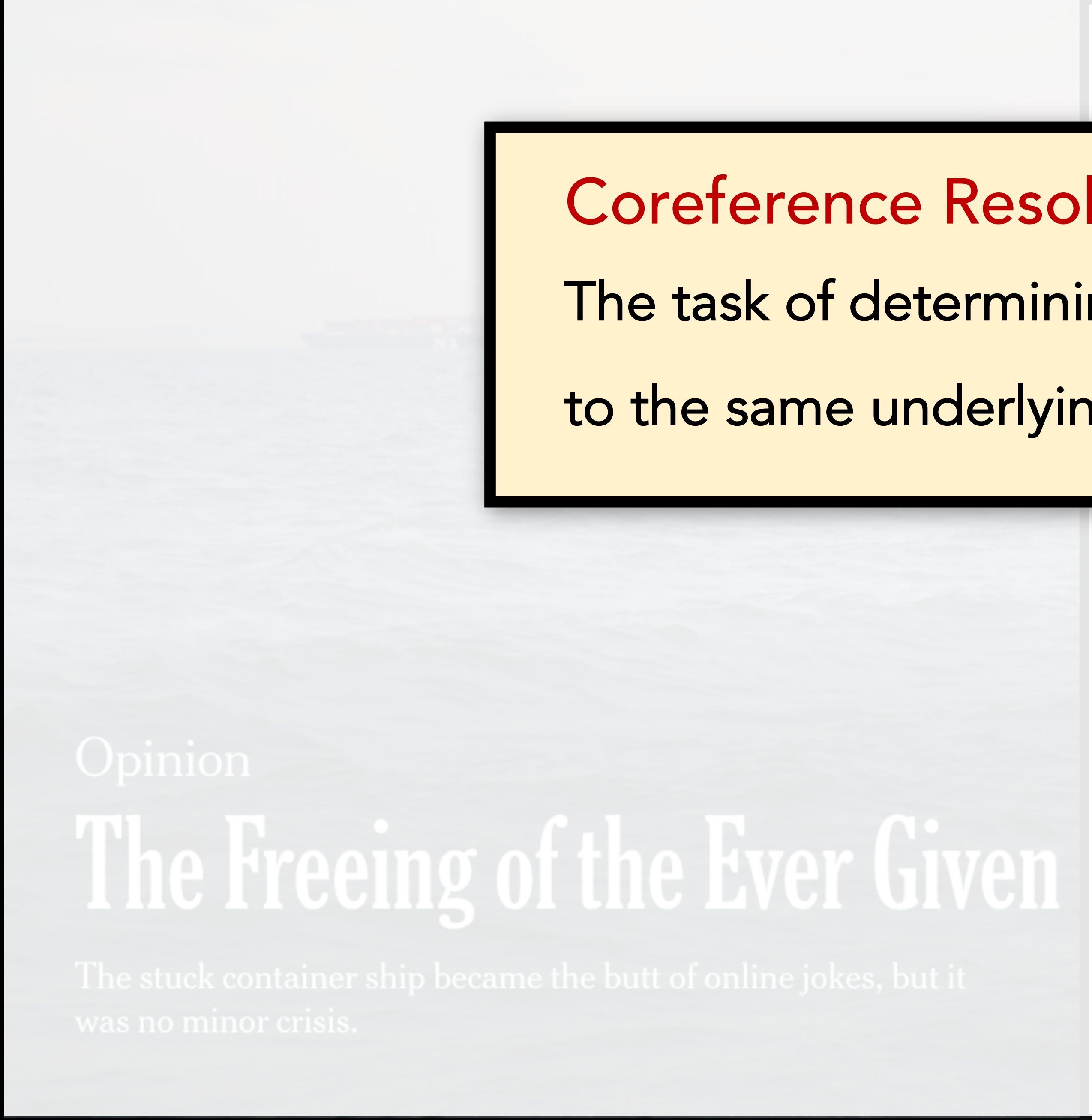
By Serge Schmemann

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*



Coreference Resolution

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

EASY FOR HUMANS

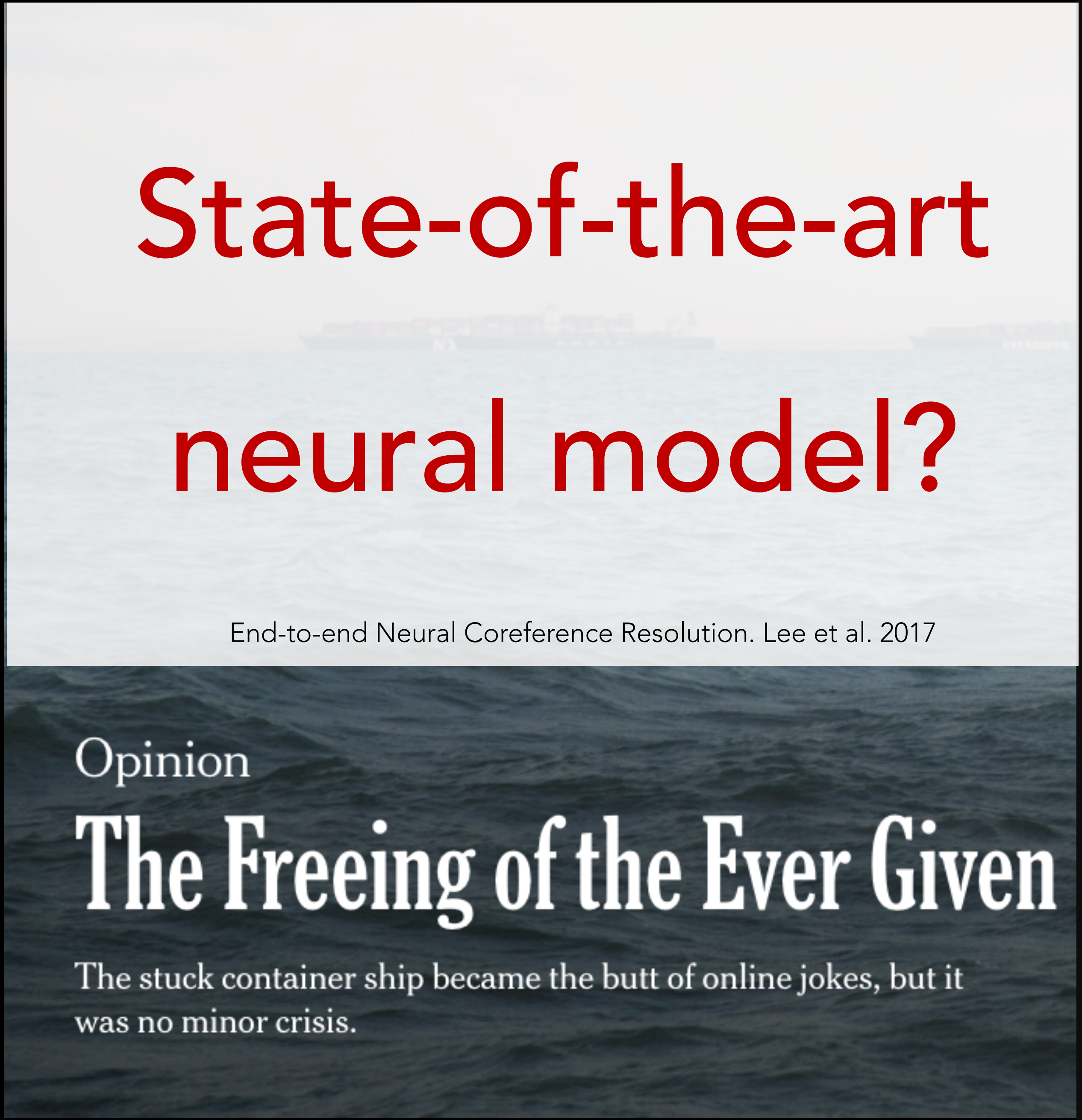
Opinion

The Freeing of the Ever Given

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

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.

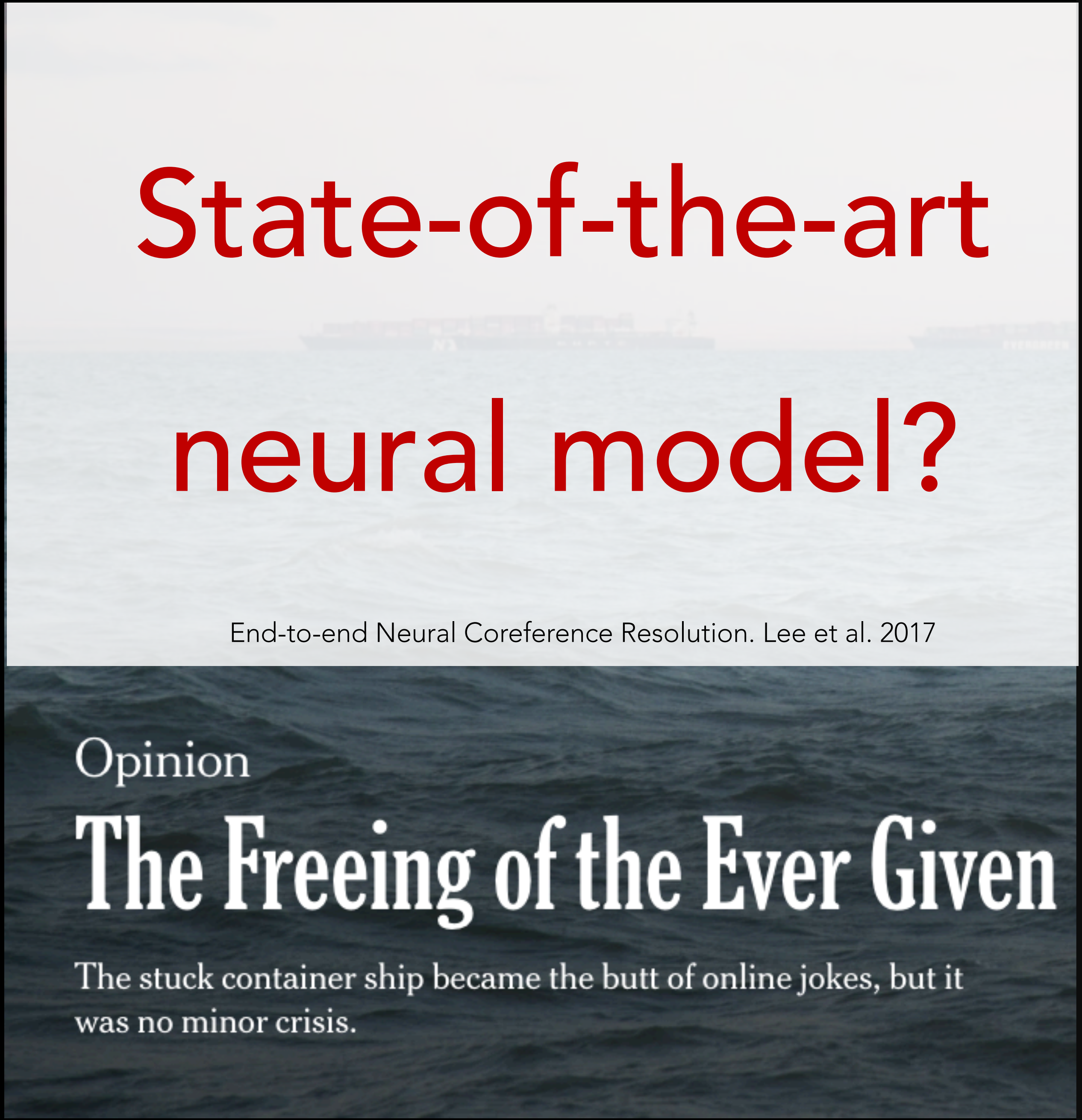
The New York Times

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.

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.

HARD FOR COMPUTERS

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.

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

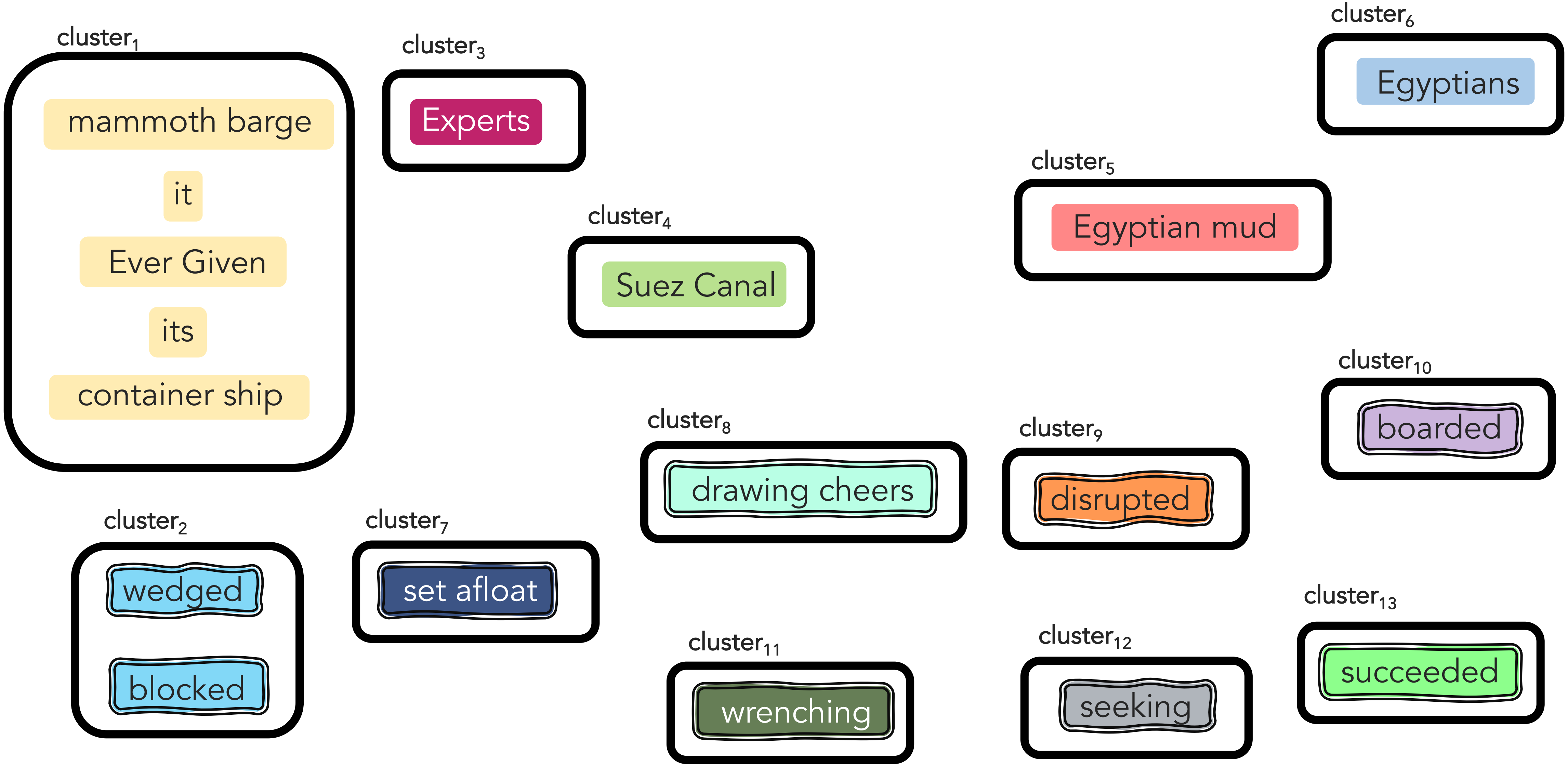
In the end, a full moon succeeded where puny machines could not, wrenching the mammoth barge out of the Egyptian mud in which it became wedged six days earlier. A spring tide finally set the Ever Given and its enormous stack of 18,300 shipping containers afloat again, drawing cheers from Egyptians on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


cluster$_1$
mammoth barge
it
Ever Given
its
container ship
cluster$_2$
wedged
blocked
cluster$_3$
Experts
cluster$_4$
Suez Canal
cluster$_5$
Egyptian mud
cluster$_6$
Egyptians
cluster$_7$
set afloat
cluster$_8$
drawing cheers
cluster$_9$
disrupted
cluster$_{10}$
boarded
cluster$_{11}$
wrenching
cluster$_{12}$
seeking
cluster$_{13}$
succeeded

Takeaway #1

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

Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

Motivation

Coreference resolution allows one to better understand what is going on (i.e., **who is who** and **what is what**)

Helps with other NLP tasks:

- Information extraction/retrieval
- Question Answering
- Document Summarization

"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

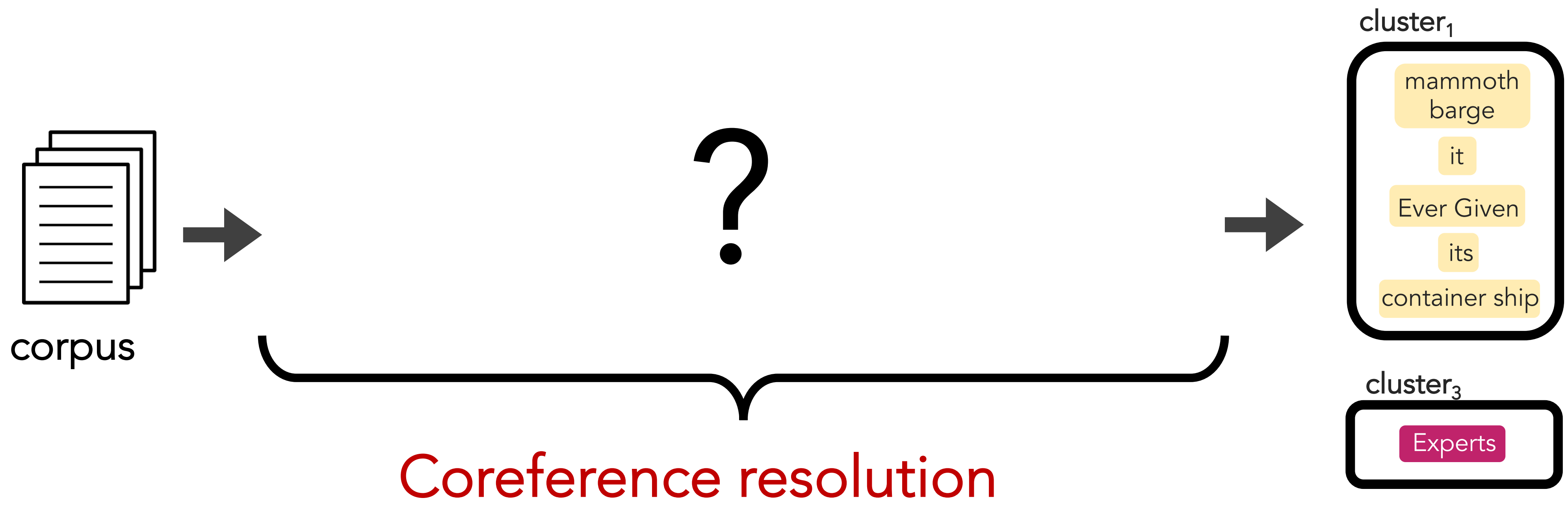
Outline

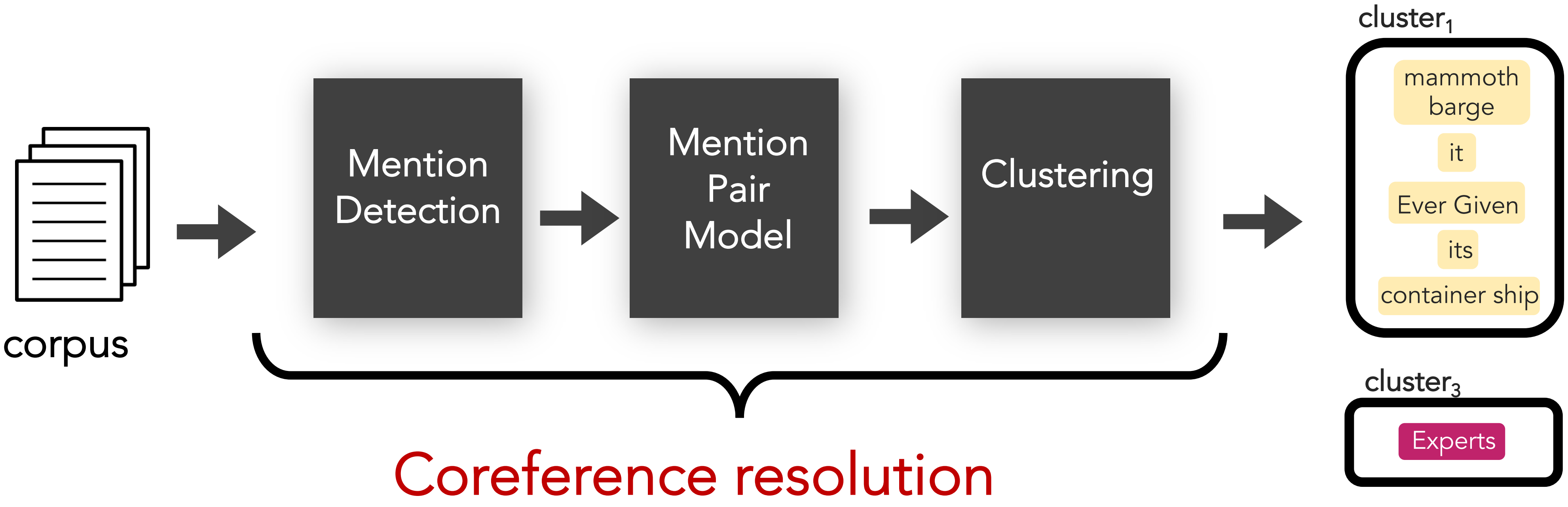
- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

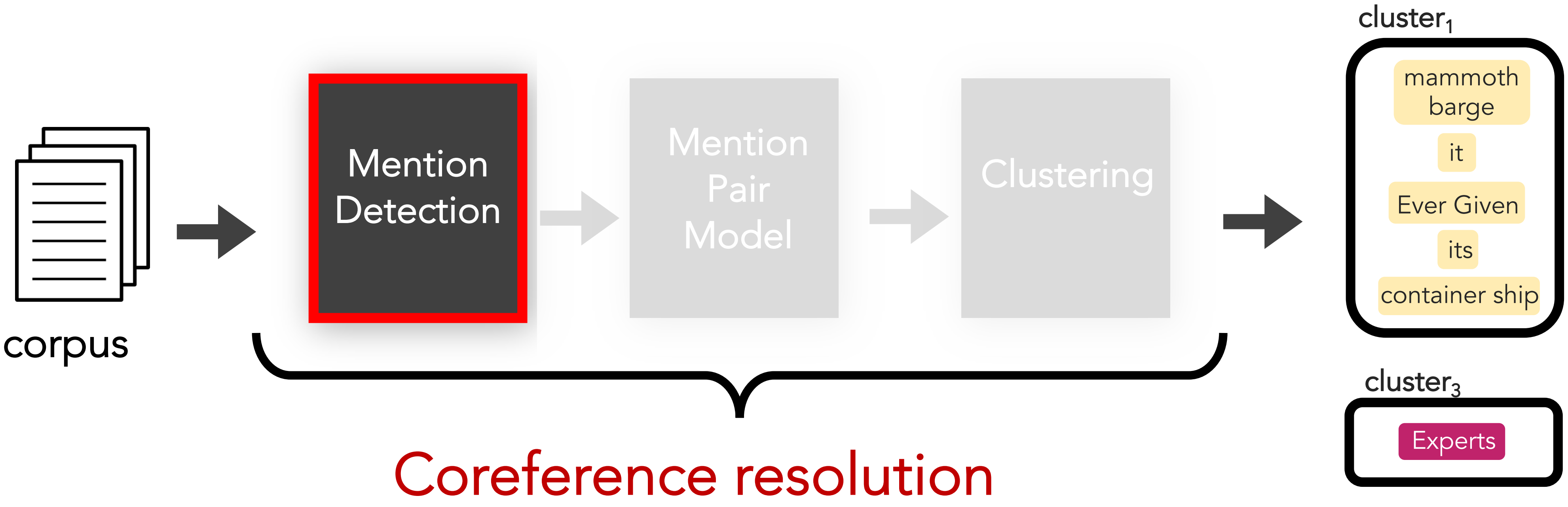
Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

How do all coref systems work?


corpus
?
Coreference resolution
cluster1
mammoth barge
it
Ever Given
its
container ship
cluster3
Experts


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
cluster1
mammoth barge
it
Ever Given
its
container ship
cluster3
Experts

Mention Detection

Determines which spans of words constitute a mention

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

entities

Mention Detection

Determines which spans of words constitute a mention

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

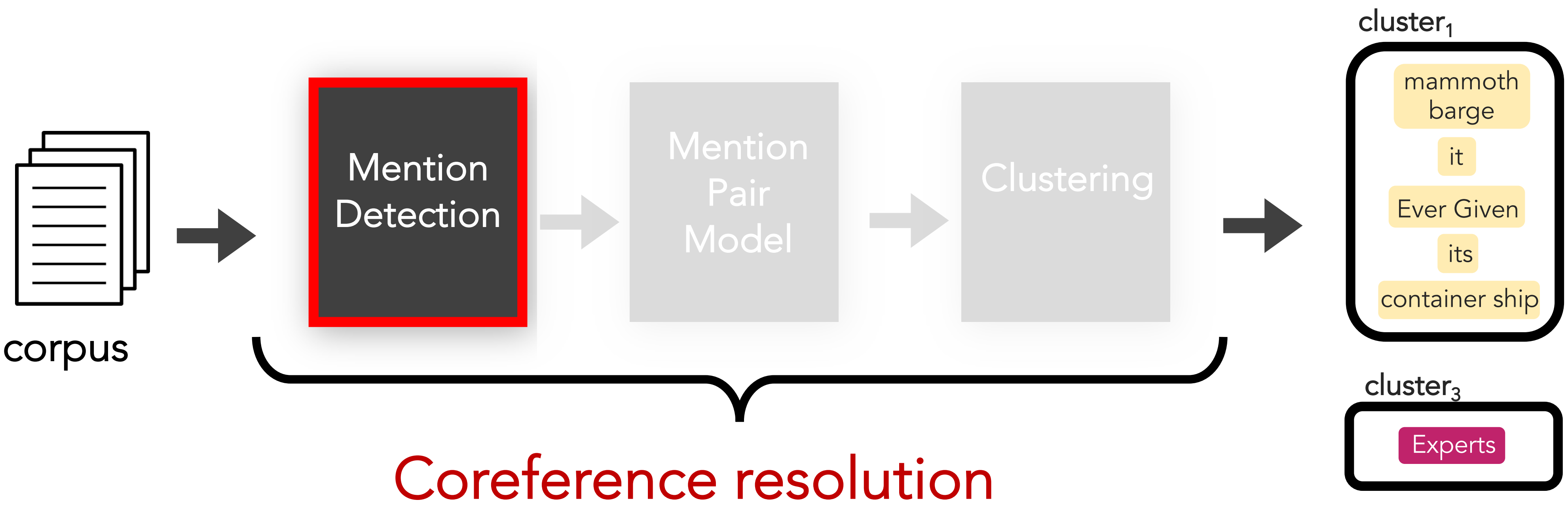
events

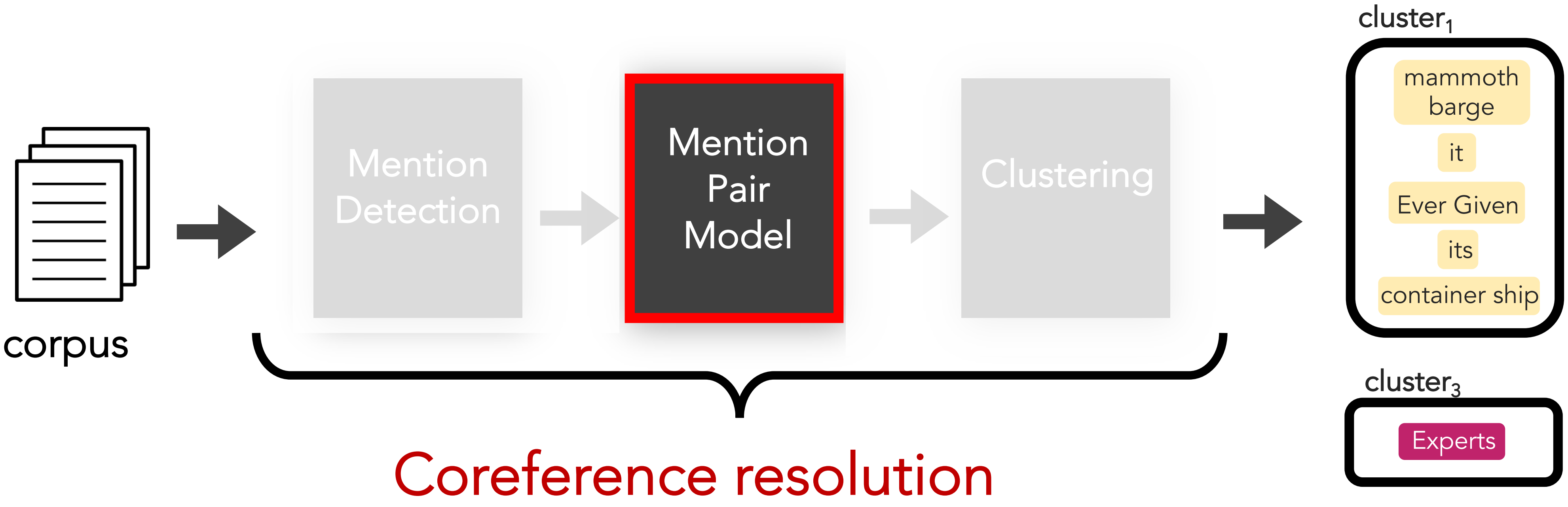
Mention Detection

Determines which spans of words constitute a mention

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

entities + events

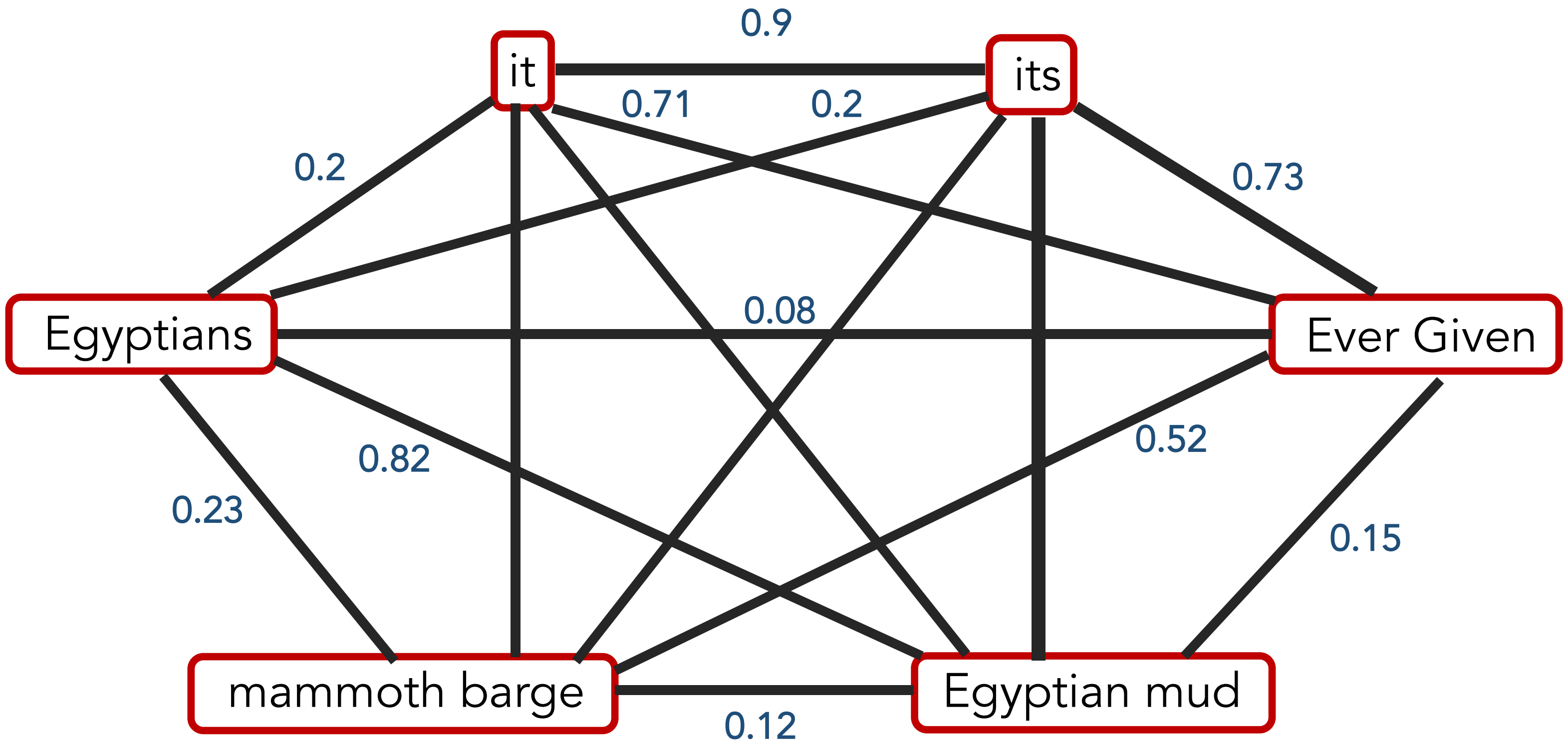

corpus
Mention Detection
Mention Pair Model
Clustering
Coreference resolution
cluster$_1$
mammoth barge
it
Ever Given
its
container ship
cluster$_3$
Experts

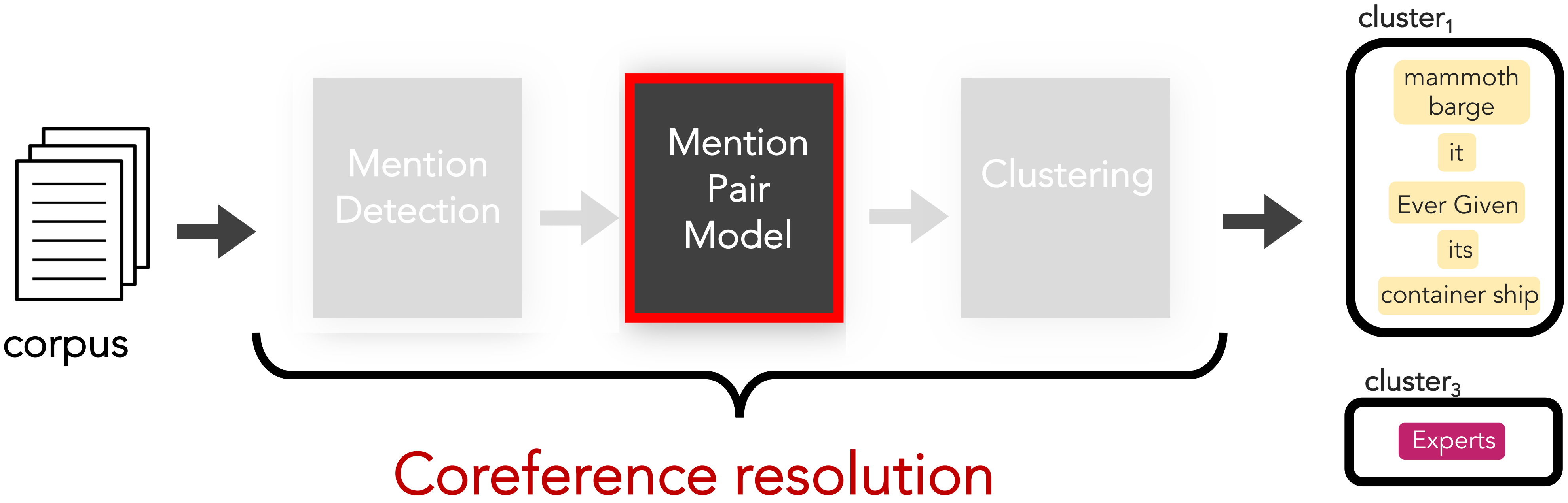


corpus
Mention Detection
Mention Pair Model
Clustering
cluster$_1$
mammoth barge
it
Ever Given
its
container ship
cluster$_3$
Experts
Coreference resolution

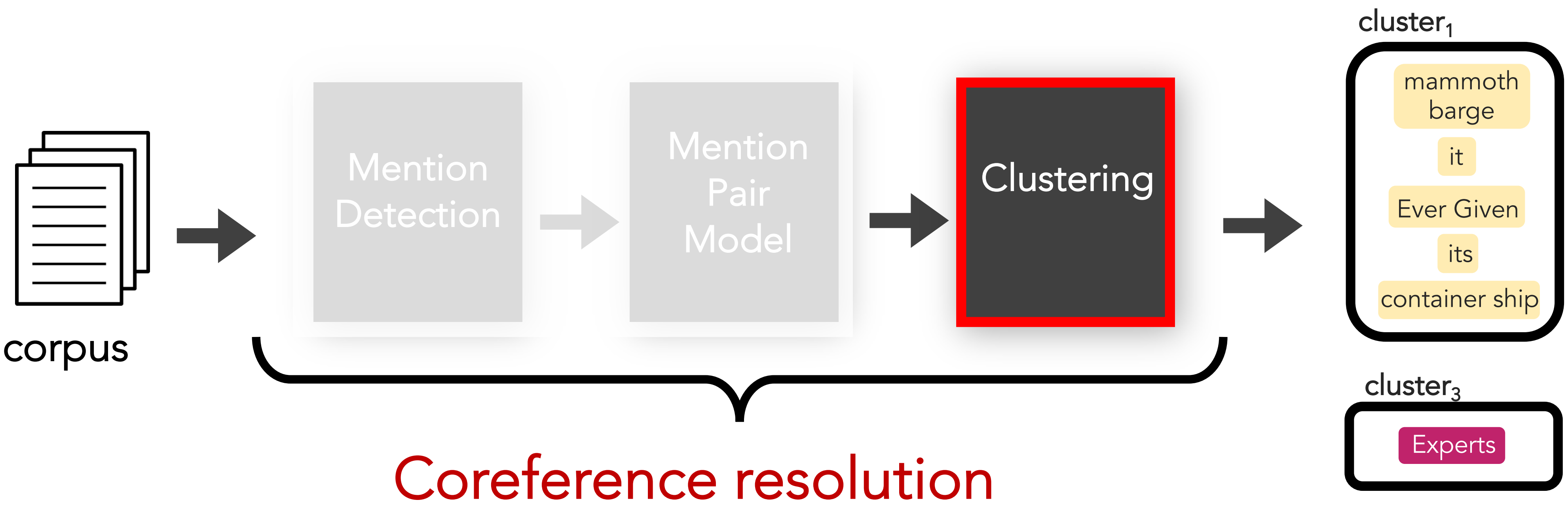
Mention Pair Model

Calculates a coref probability for all pairs of mentions



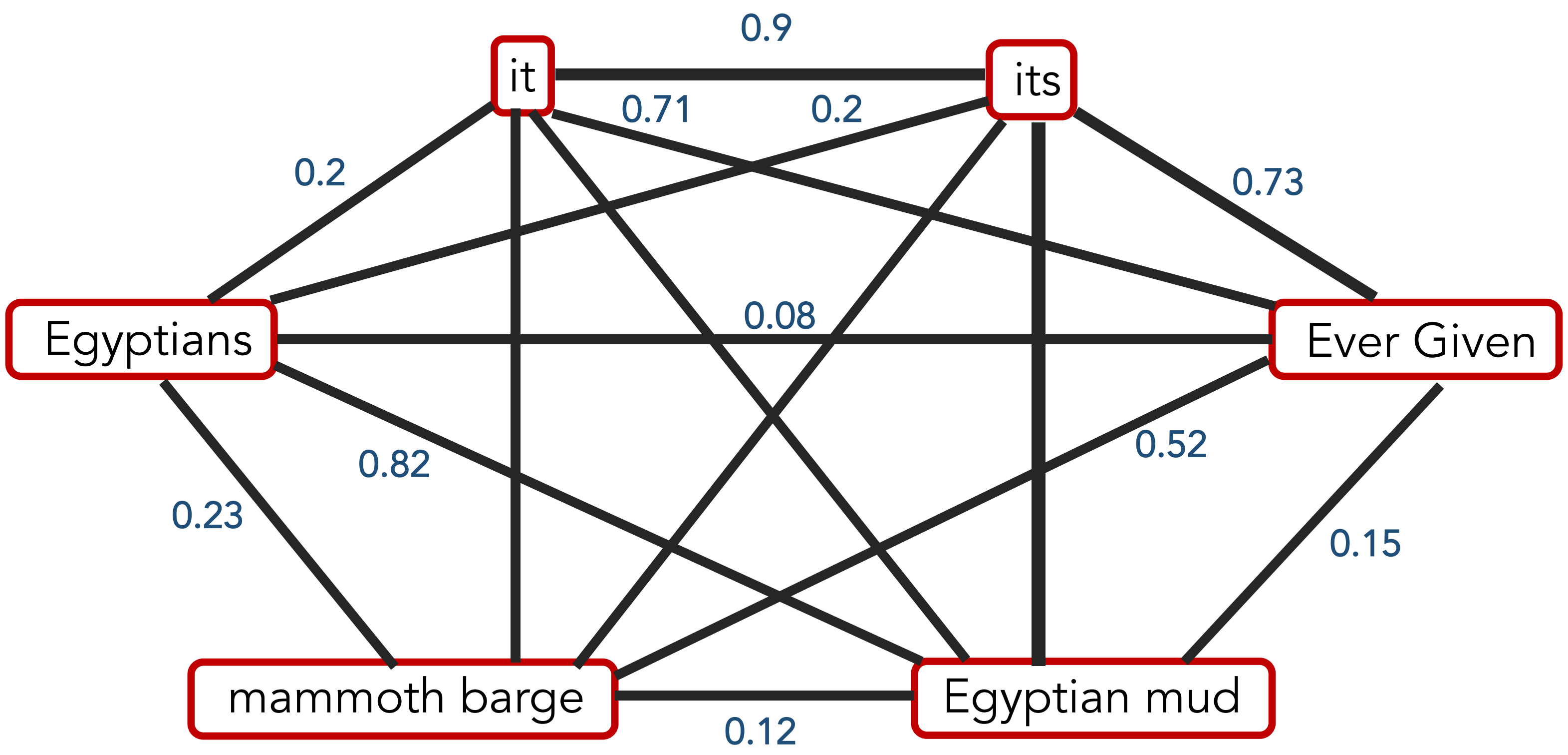


corpus
Mention Detection
Mention Pair Model
Clustering
Coreference resolution
$cluster_1$
mammoth barge
it
Ever Given
its
container ship
$cluster_3$
Experts


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

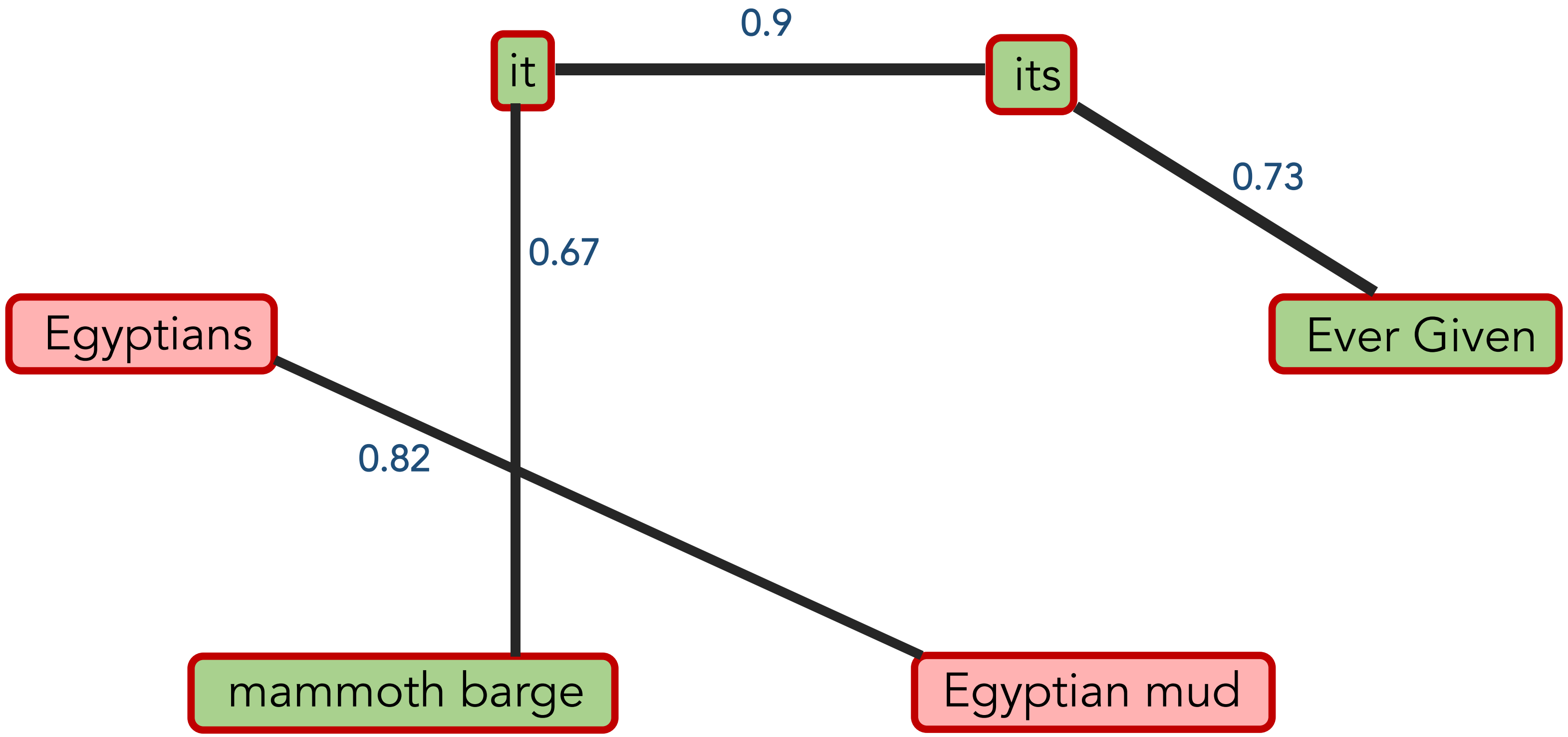
Clustering

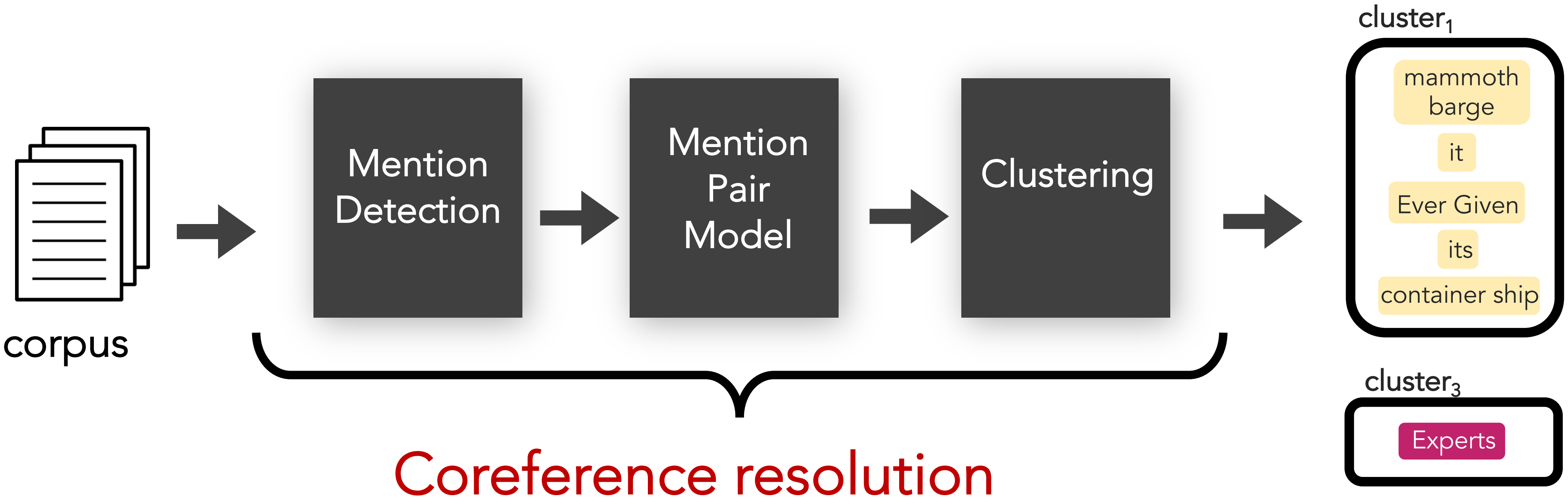
Uses the coref probabilities to determine clusters



Clustering

Uses the coref probabilities to determine clusters




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

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.

Donald Glover

Glover at the premiere of *The Martian* in September 2015

Born	Donald McKinley Glover Jr. September 25, 1983 (age 37) Edwards Air Force Base, Edwards, California, U.S.
Other names	Childish Gambino · mcDJ

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

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 prono
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.

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

Entity Coreference (2011 – present)

Then, m… tons of … features

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

Additional Mention Features: The type of 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.

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.

Event Coreference (2014 - present)

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.

Event Coreference (2017 – 2019)

Novel deep learning approaches

used very few features

Event Coreference Resolution by Iteratively Unfolding Inter-dependencies among Events. Choubey and Huang. EMNLP 2017.

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

Event Coreference (2017 – 2019)

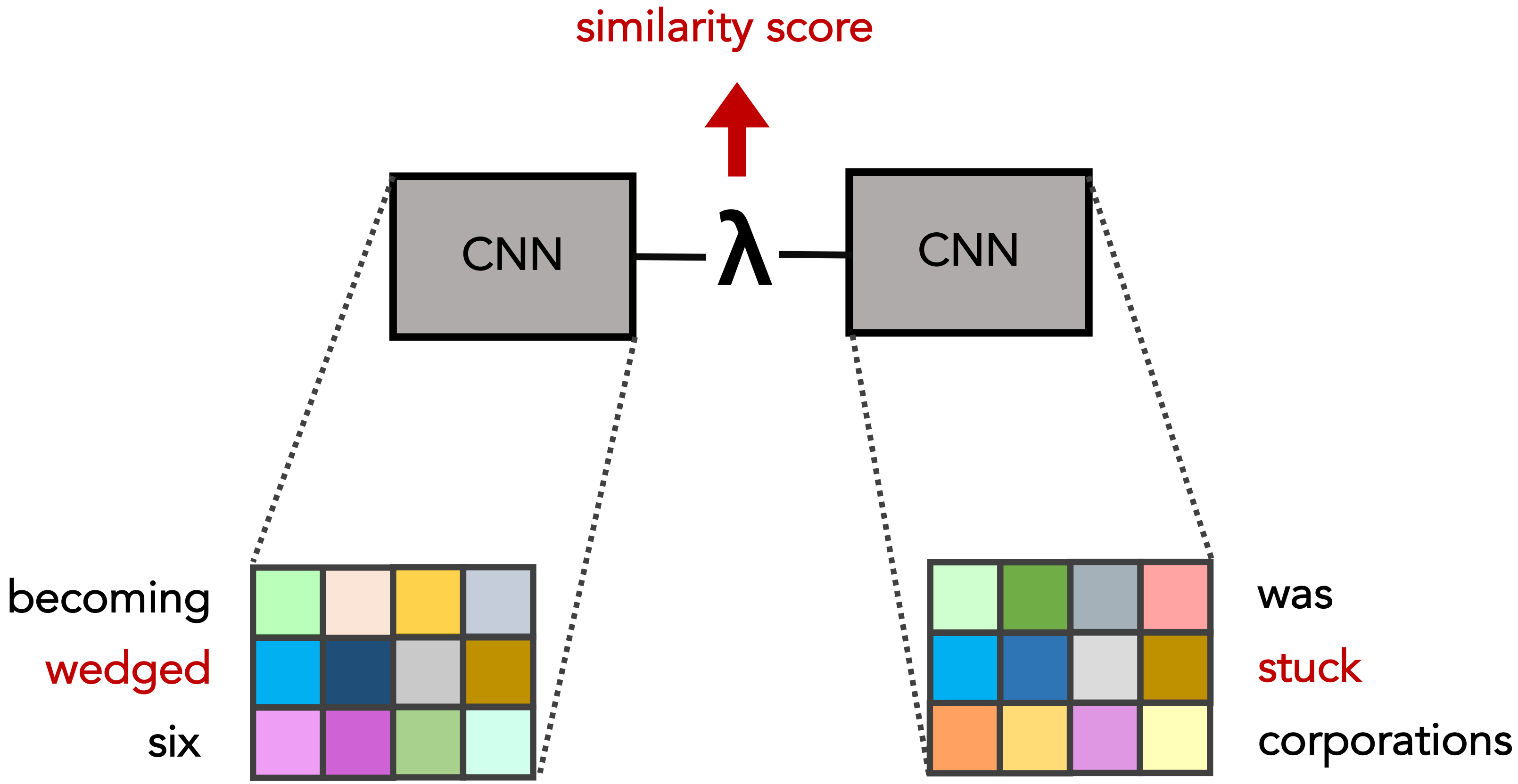
Key insight: use contextualized word embeddings to automatically learn feature representations

The Ever Given was finally freed after becoming wedged six days prior. While it was stuck, corporations lost an estimated $1.2 billion in commerce.

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

Event Coreference (2017 – 2019)

~~Siamese~~ Conjoined CNN w/ Contrastive Loss



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

Event Coreference (2017 – 2019)

	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

Event Coreference (2019)

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 …

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

Event Coreference (2017 – 2019)

FINDINGS

- state-of-the-art for event coref
- **Character Embeddings + Lemma Embeddings** were the only two necessary features

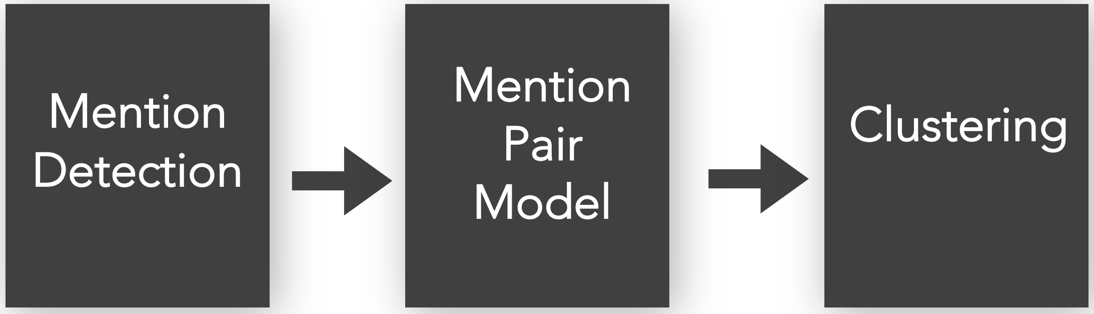
Cross-Document Coreference Resolution for Entities and Events. Tanner. Brown University Dissertation. 2019

Takeaway #3 The community needs a better corpus.

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

Entity Coreference (2017)

First end-to-end neural system

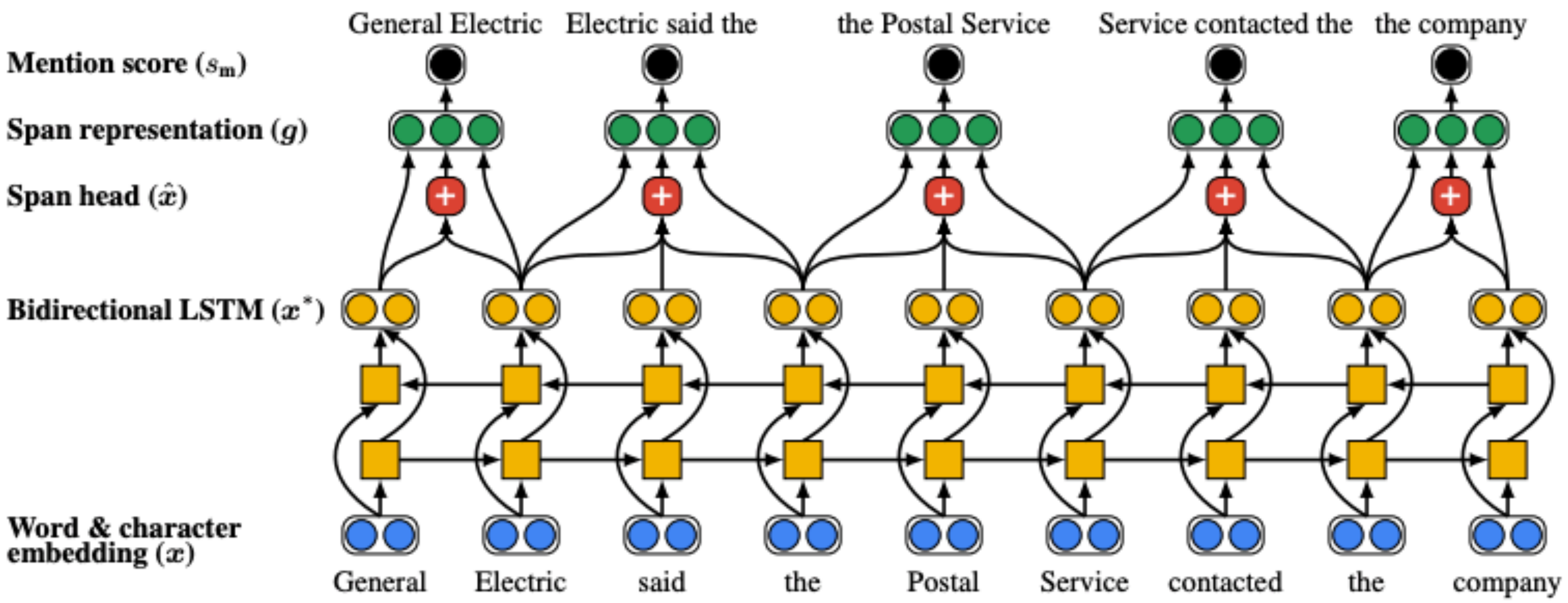


End-to-end Neural Coreference Resolution. Lee et al. EMNLP 2017.

Entity Coreference (2017)

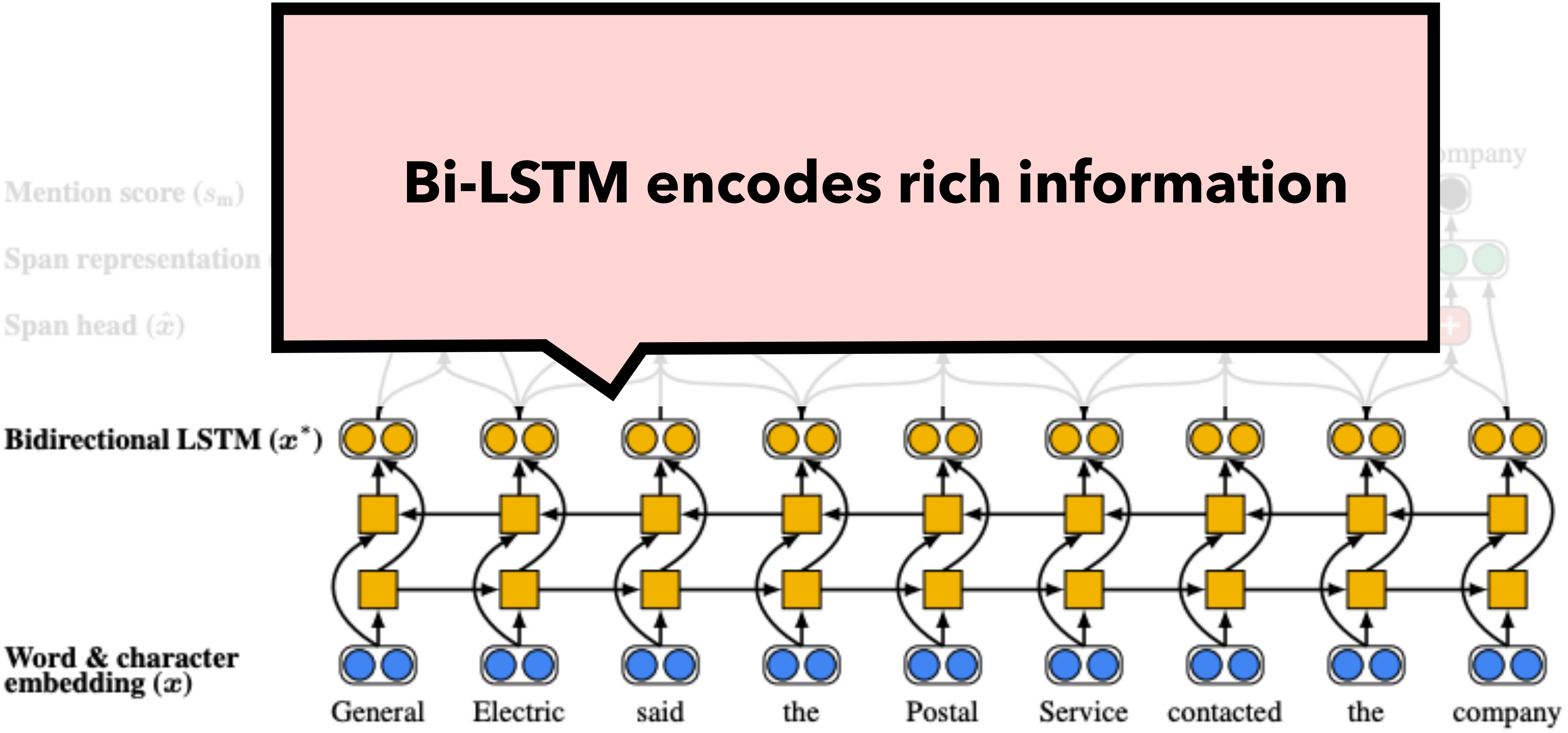
CoNLL F1: 68.8

First end-to-end neural system



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

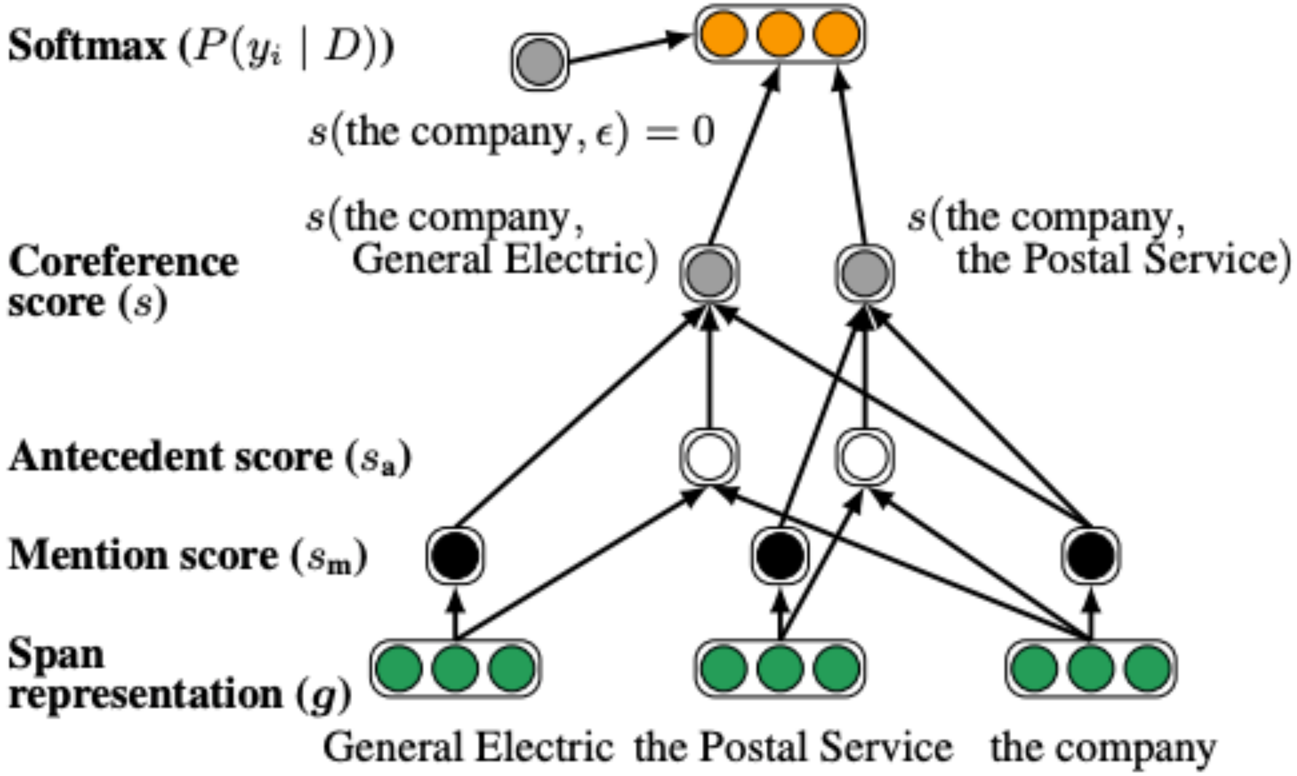
Entity Coreference (2017)



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

Entity Coreference (2017)

First end-to-end neural system



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

Entity Coreference (2017)

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.

Entity Coreference (2019 - present)

REMAINING ISSUES

- Pronouns (especially in conversation)
- Conflating relatedness with equality
  (e.g., "*Flight attendants*" *with* "*pilots*")
- World-knowledge

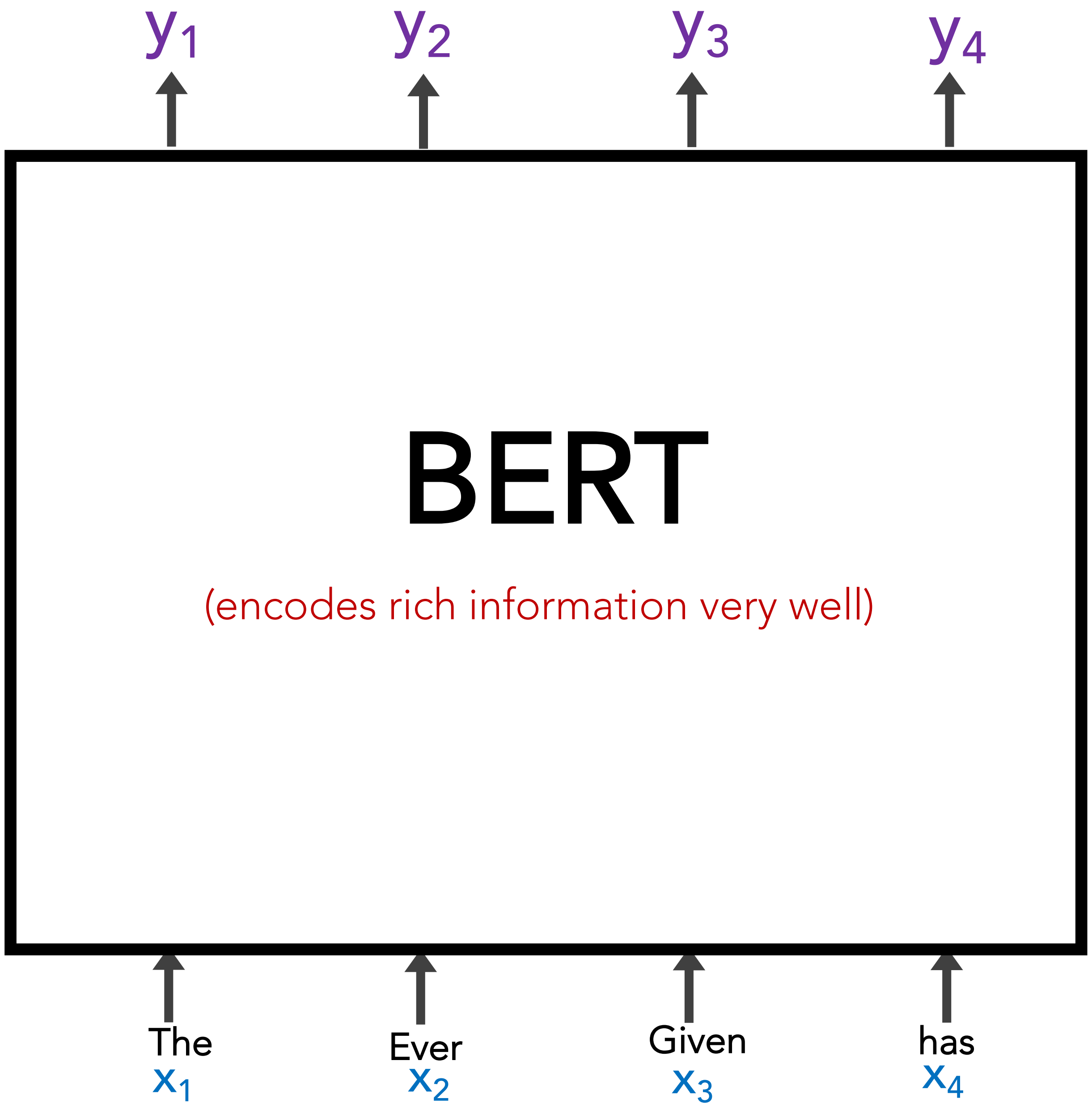
> Also such location devices, (**some ships**) have smoke floats (**they**) can toss out so the man overboard will be able to use smoke signals as a way of trying to, let the rescuer locate (**them**).

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

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

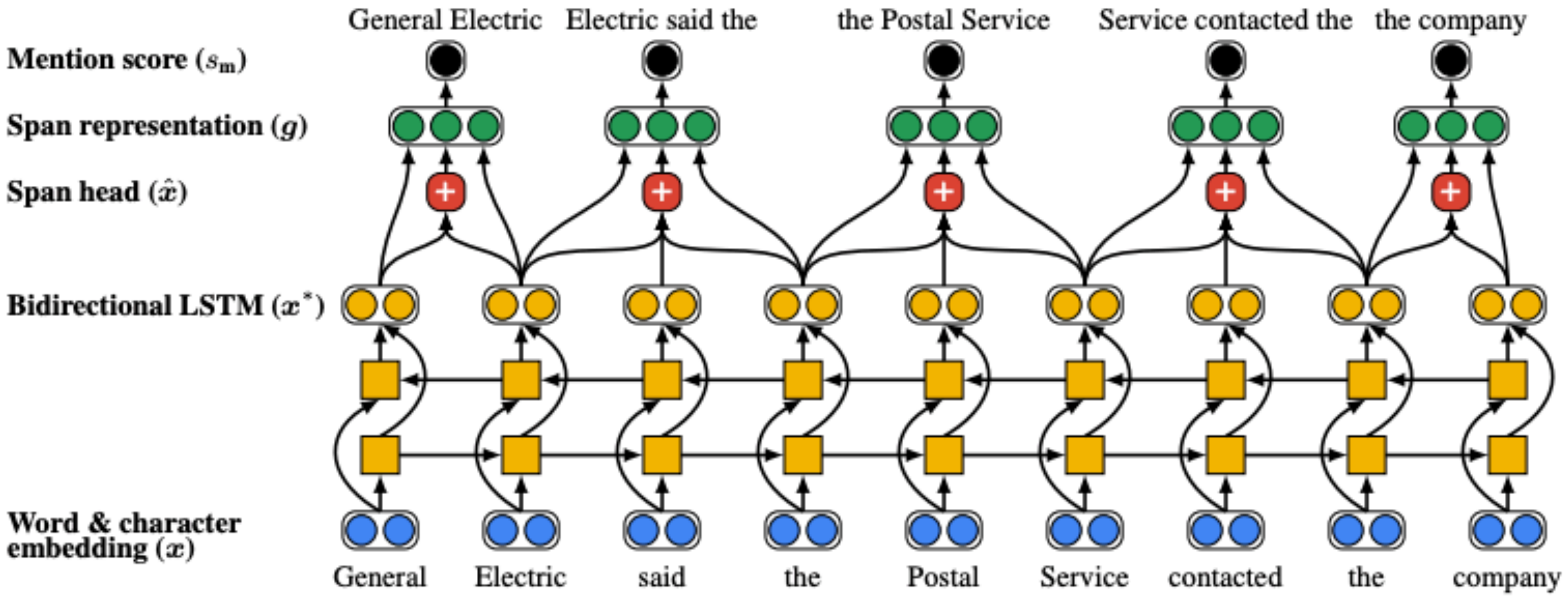
BERT is incredible, can we use it for coreference resolution?

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Devlin et al. NAACL 2019. ~19,000 citations!


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

Entity Coreference (2017)

First end-to-end neural system

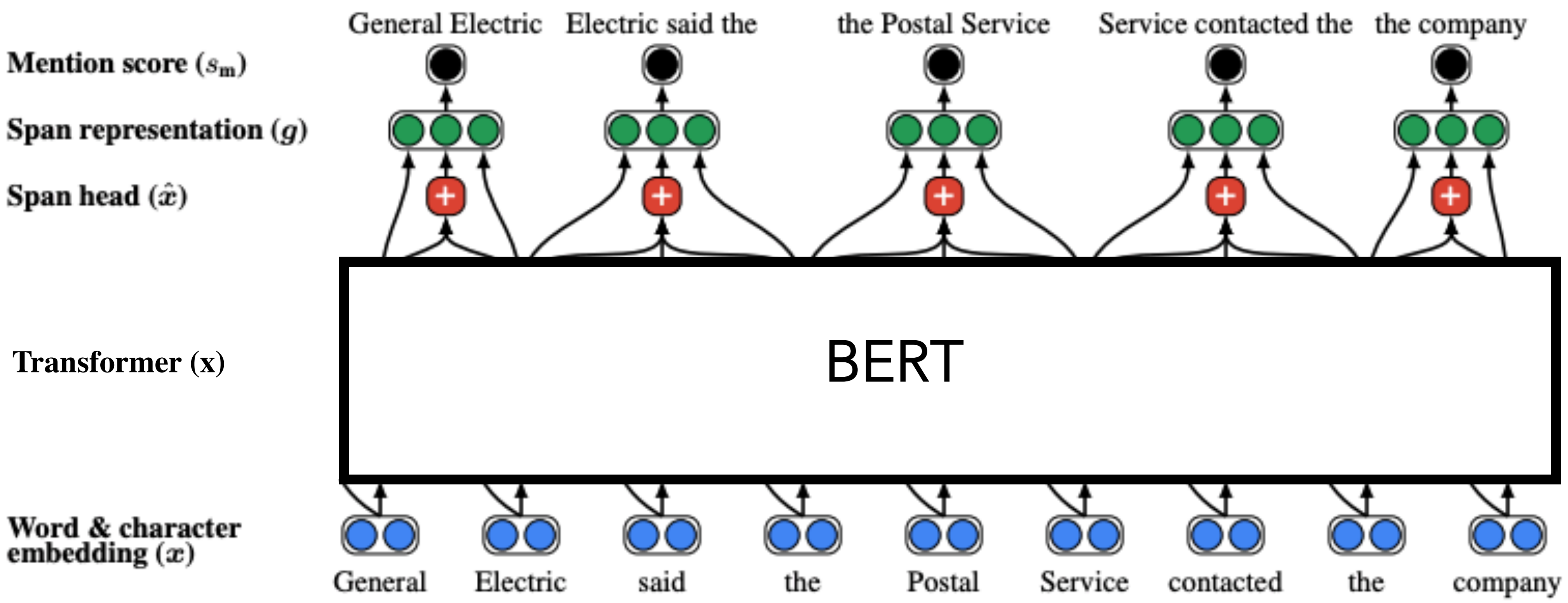


End-to-end Neural Coreference Resolution. Lee et al. EMNLP 2017.

Entity Coreference (2019)

CoNLL F1: 79.6

BERT Improvements



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

Discourse

Pragmatics

Semantics

Syntax

Lexemes

Morphology

Characters

Research has demonstrated that BERT can capture many complex linguistic properties.

However, coref is still far from solved

What Does BERT Look At? An Analysis of BERT's Attention. Clark et al. ACL 2019.

What does BERT learn about the structure of language? Jawahar et al. ACL 2019.

BERT Rediscovers the Classical NLP Pipeline. Tenney et al. ACL 2019

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Devlin et al. NAACL 2019.

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

Category			
Related Entities			
Lexical			
Pronouns			
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 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, nor do we know what is 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

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

Outline

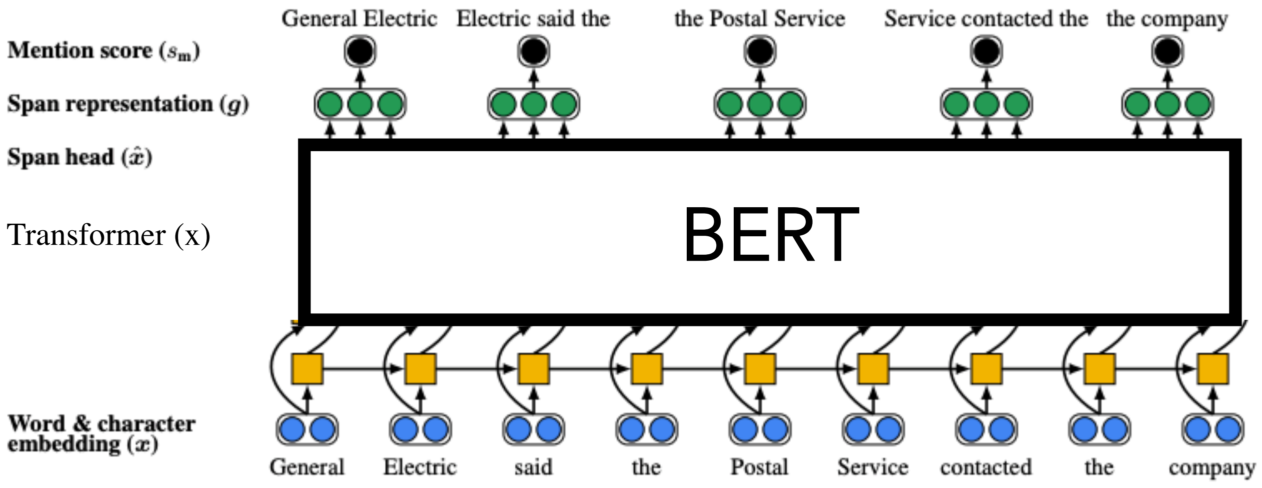
- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

Since labelled data is lacking, can we build a powerful unsupervised model?

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



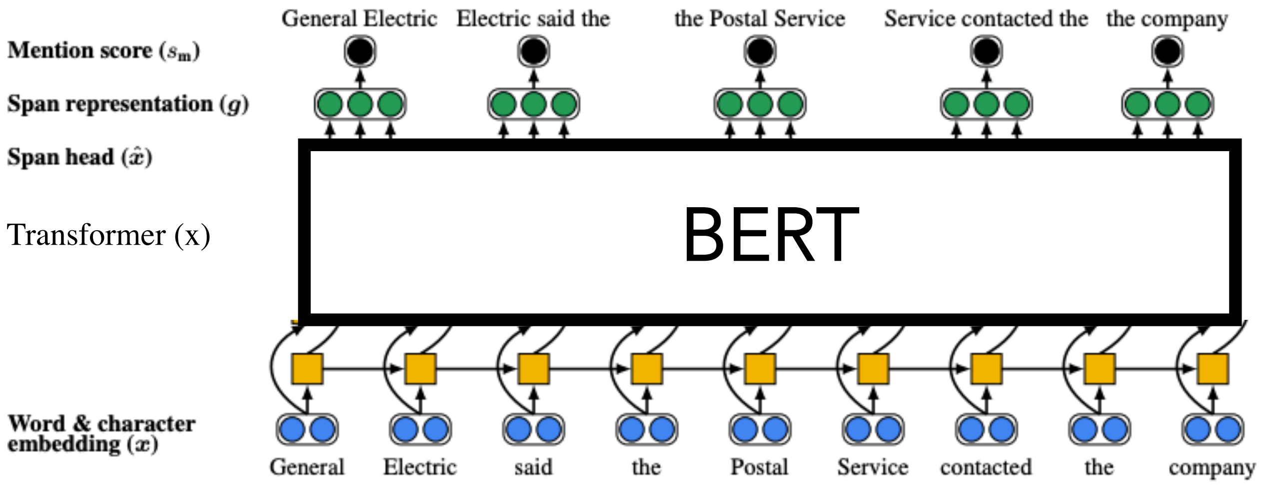
Unsupervised Coreference Resolution with Contextualized Representations and Linguistic Prior Knowledge. Alessandro Stolfo et al. In preparation.

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)

Supervised
(needs training data)

Unsupervised Coreference Resolution with Contextualized Representations and Linguistic Prior Knowledge. Alessandro Stolfo et al. In preparation.

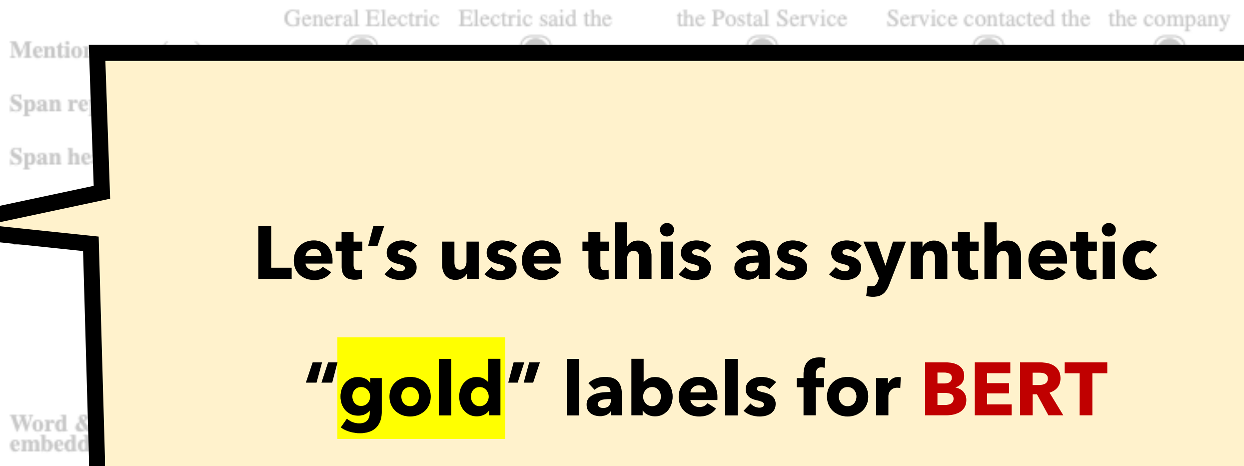
We combine the old school, manual rule-based system

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)

CONCERN

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

Unsupervised Coreference Resolution with Contextualized Representations and Linguistic Prior Knowledge. Alessandro Stolfo et al. In preparation.

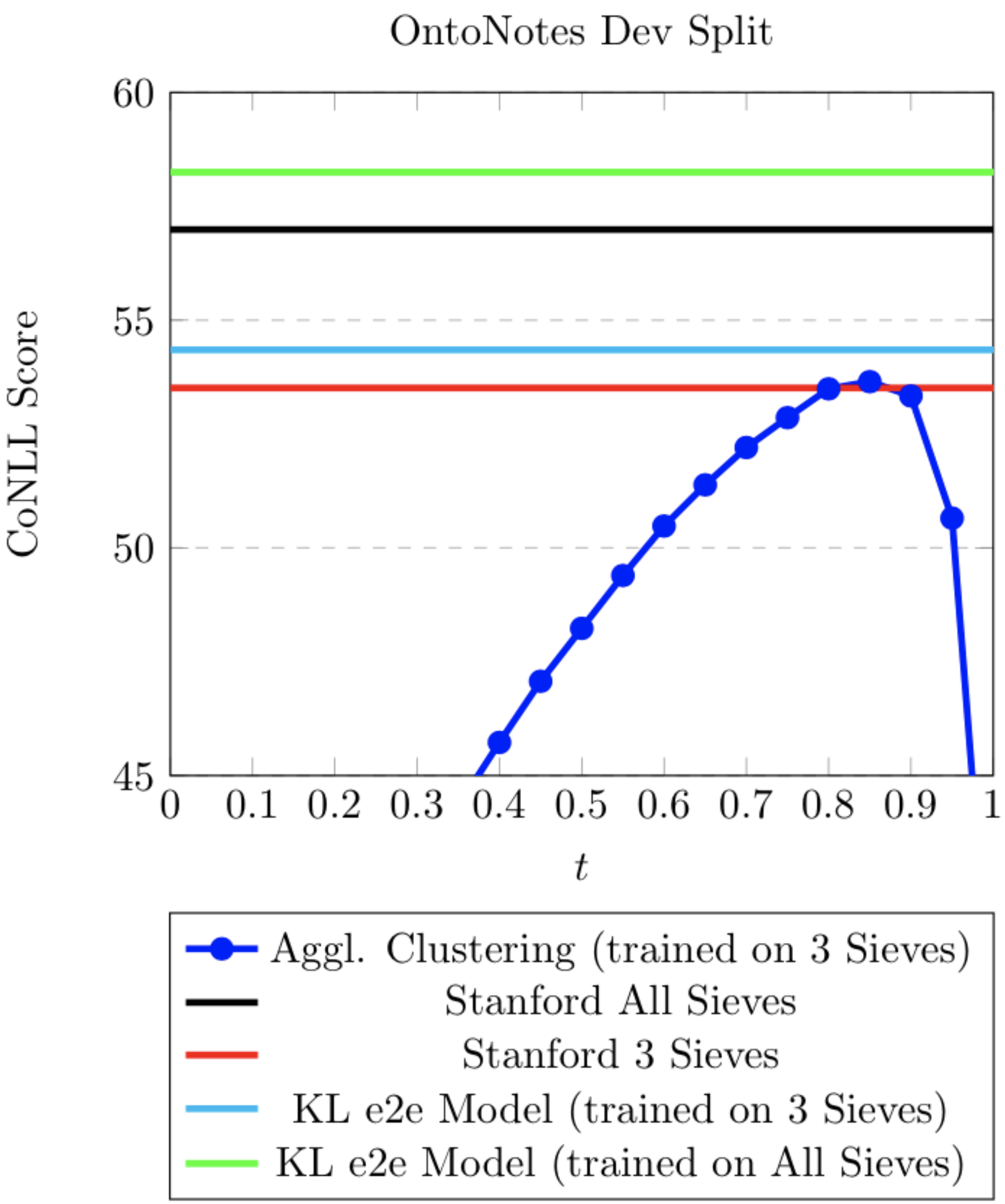
CONCERN

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

Unsupervised Coreference Resolution with Contextualized Representations and Linguistic Prior Knowledge. Alessandro Stolfo et al. In preparation.



Unsupervised Coreference Resolution with Contextualized Representations and Linguistic Prior Knowledge. Alessandro Stolfo et al. In preparation.

Coreference Type	Name	# Docs
events	ECB+	982
entities	OntoNotes	3,493

How can we create the biggest, best coreference dataset for entity and events?

Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

We're building an annotation tool that allows users to:

- collaborate with others
- run remotely on humbleNLP.com (coming soon)
- get started with many state-of-the-art models
- quickly annotate cross-document coref via **entity linking**
- have different permissions (e.g., annotator, approver, admin)

Lindsay Lohan Leaves Betty Ford , Checks Into Malibu Rehab

First Published : June 13 , 2013 4 : 59 PM EDT Lindsay Lohan has left the Betty Ford Center and is moving to a rehab facility in Malibu , Calif . , Access Hollywood has confirmed . A spokesperson for The Los Angeles Superior Court confirmed to Access that a judge signed an order yesterday allowing the transfer to Cliffside

Suggest Mentions

- ● SpanBERT
- ○ SpaCy

Annotate Mentions

- ○ SpanBERT
- ○ SpaCy

Show part-of-speech

- ○ All
- ○ Noun
- ○ Verb
- ○ Pronoun

Lindsay Lohan Leaves Betty Ford , Checks Into Malibu Rehab

First Published : June 13 , 2013 4 : 59 PM EDT Lindsay Lohan has left the Betty Ford Center and is moving to a rehab facility in Malibu , Calif . , Access Hollywood has confirmed . A spokesperson for The Los Angeles Superior Court confirmed to Access that a judge signed an order yesterday allowing the transfer to Cliffside

Suggest Mentions

- ● SpanBERT
- ○ SpaCy

Annotate Mentions

- ○ SpanBERT
- ● SpaCy

Show part-of-speech

- ○ All
- ○ Noun
- ○ Verb
- ○ Pronoun

Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

Outline

- NLP Overview
- Coreference Resolution
  - What
  - Why
  - How
- Improvements
  - No Data
  - Better Data
- Conclusions

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.

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.

Current Students

Efficient Active Learning for Entity-based Annotation

Xin Zeng

IACS MS Thesis

Automated Captioning for Data Visualizations

Anita Mahinpei

IACS MS Thesis

Commonsense Adversarial NLP

Jack Scudder

IACS MS Thesis

Joint Entity and Event Coreference

Xin Zeng

IACS MS Thesis

Joint Entity and Event Coreference

Ning Hua

Smith College x Harvard Independent Study

End-to-End Entity Linking

Mingyue Wei

IACS MS Thesis

Unsupervised Coreference Resolution

Alessandro Stolfo

ETH-Zurich MS Thesis

Co-advised by Mrinmaya Sachan

Grammar Correction and Language Learning

Yoel Zweig

ETH-Zurich MS Thesis

Co-advised by Mrinmaya Sachan

Sign Language Classification for Novice Learners

Ali Hindi

Brunswick High School

Accepted to Stanford

Thomas Fouts

Brunswick High School

Awaiting Decisions

Current Students (continued)

Coreference with Commonsense

Xavier Evans

Harvard
Independent Study

Current Collaborators

Annotation tools for Coreference Resolution

Shivas Jayaram Harvard DCE Graduate

Eduardo Peynetti Harvard DCE Student

Joe Brucker @ self-employed

Data Augmentation for Neural Models

Mingyue Wei IACS MS

Qiang Fei IACS Graduate

Shuyuan Xiao IACS Graduate

Yingsi Jian IACS Graduate

Shahab Asoodeh Harvard Physics Post-doc

Ekin Dogus Cubuk Google

Gene Prediction with Language Modelling

Benjamin Levy IACS Graduate

Zihao Xu IACS Graduate

Liyang Zhao IACS Graduate

Shuying Ni IACS Graduate

Phoebe Wong IACS Graduate

Ross Altman Inari

Karl Kremling Inari

Thanks!

Questions?

BACK-UP SLIDES

More examples of why Event coref is hard

Wide-reading

The attack took place yesterday

The bombing killed four people

Paraphrase

She gave him the book

He was given the book by her

Lexical Identity

It was destroyed

The destruction of the town …

Events are Not Simple: Identity, Non-Identity, and Quasi-Identity. Hovy et al. EVENTS-workshop 2013.

Event Coreference (2019)

False Negatives

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%

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

Event Coreference (2019)

False Positives

False Positives	
Context-Dependent (30%)	
Example 1	The 55-year-old Scottish actor will replace Matt Smith, who **announced** in June that he was leaving the sci-fi show later this year.
	Peter Capaldi has been **announced** as the new Doctor Who, the 12th actor to take up the coveted TV role.
Similar Meanings (38%)	
Example 2	Frederick C. Larue, a top Nixon campaign official who **passed** money from a secret White House fund, **died** Saturday at a hotel in Biloxi, Miss.
Wide-Reading (14%)	
Example 3	Peyton manning helped inspire the Indianapolis Colts to their eighth straight **win** as they **overcame** Jacksonville this season.
Unclear (13%)	
Example 4	Microsoft today issued an emergency update to **plug** a critical security hole present in all version of its browser, a flaw hackers have **used** to steal data from millions of Windows users.
Syntax (3%)	
Example 5	Creighton **defeats** Drake 65-53 in MVC tournament.
	In Saturday's semi-finals, Creighton will play no. 5 seed Indiana state, which **defeated** Evansville 51-50 on Friday.
Too Difficult for Me (2%)	
Example 6	Submarine cable **problem** disrupts telecom services in Alexandria .
	Vodafone has been affected by a **damage** in one of the fiber cables going from the Ramsis Communication Center all the way to Sadat City.

Table 4.4: Examples of CCNN's False Positives from the ECB+ Development Set, grouped by categories of errors.

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

Event Coreference (2019)

False Negatives

False Negatives	
Semantics (42%)	
Example 1	Hansbrough scored 20 points Thursday night, **breaking** North Carolina's career scoring record, and the tar heels beat visiting Evansville, 91-73 .
	Hansbrough **sets** scoring record in victory .
Unclear (20%)	
Example 2	Hewlett-Packard's **purchase** of electronic data systems could mean tougher competition for IBM and its 10,500 triangle employees.
	The all-cash **deal**, announced Tuesday, represents HP's biggest gamble under the leadership of Mark Hurd.
Colloquial Variations (16%)	
Example 3	Industry experts told The Times that two sub-sea cables **went down** just off Alexandra, causing mass disruption.
	Millions of people across the Middle East and Asia have **lost** access to the Internet after two undersea cables in the Mediterranean suffered severe damage.
Longer Names (14%)	
Example 4	An **earthquake** with a preliminary magnitude of 4.6 was recorded in the North Bay this morning, according to the U.S. Geological Survey.
	A **4.6-magnitude earthquake** was recorded near Healdsburg .
Pronouns (8%)	
Example 5	President Obama announces **nominee** for surgeon general.
	Today, President Barack Obama announced his intent to **nominate** Regina M. Benjamin as surgeon general, department of health and human services.

Table 4.5: Examples of CCNN's False Negatives from the ECB+ Development Set, grouped by categories of errors.

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

Stanford Multi-pass sieves

Heeyoung Lee, Yves Peirsman, Angel Chang,
Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky

Components					MUC			B^3			CEAFE			BLANC			
ER	D	S	GA	GM	R	P	F1	R	P	F1	R	P	F1	R	P	F1	avg F1
√					58.8	56.5	57.6	68.0	68.7	68.4	44.8	47.1	45.9	68.8	73.5	70.9	57.3
	√				59.1	57.5	58.3	69.2	71.0	70.1	46.5	48.1	47.3	72.2	78.1	74.8	58.6
√	√				60.1	59.5	59.8	69.5	71.9	70.7	46.5	47.1	46.8	73.8	78.6	76.0	59.1
√	√	√			60.3	58.5	59.4	69.9	71.1	70.5	45.6	47.3	46.4	73.9	78.2	75.8	58.8
√	√		√		63.8	61.5	62.7	71.4	72.3	71.9	47.1	49.5	48.3	75.6	79.6	77.5	61.0
√	√			√	73.6	90.0	81.0	69.8	89.2	78.3	79.4	52.5	63.2	79.1	89.2	83.2	74.2
√	√		√	√	74.0	90.1	81.3	70.2	89.3	78.6	79.7	53.1	63.7	79.5	89.6	83.6	74.5

Table 3: Comparison between various configurations of our system. ER, D, S stand for External Resources, Discourse, and Semantics sieves. GA and GM stand for Gold Annotations, and Gold Mentions. The top part of the table shows results using only predicted annotations and mentions, whereas the bottom part shows results of experiments with gold information. Avg F1 is the arithmetic mean of MUC, B^3, and CEAFE. We used the development partition for these experiments.

		MUC			B^3			CEAFE			BLANC			
Track	Gold Mention Boundaries	R	P	F1	R	P	F1	R	P	F1	R	P	F1	avg F1
Close	Not Gold	61.8	57.5	59.6	68.4	68.2	68.3	43.4	47.8	45.5	70.6	76.2	73.0	57.8
Open	Not Gold	62.8	59.3	61.0	68.9	69.0	68.9	43.3	46.8	45.0	71.9	76.6	74.0	58.3
Close	Gold	65.9	62.1	63.9	69.5	70.6	70.0	46.3	50.5	48.3	72.0	78.6	74.8	60.7
Open	Gold	66.9	63.9	65.4	70.1	71.5	70.8	46.3	49.6	47.9	73.4	79.0	75.8	61.4

Table 4: Results on the official test set.

Closed

[1] Only the provided data can be used, i.e., WordNet and gender gazetteer.

Open

[2] Any external knowledge source can be used. We used additional animacy, gender, demonym, and country and states gazetteers.

Clark and Manning

Improving Coreference Resolution by Learning Entity-Level Distributed Representations

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Avg. F_1
CoNLL 2012 English Test Data										
Clark and Manning (2015)	76.12	69.38	72.59	65.64	56.01	60.44	59.44	52.98	56.02	63.02
Peng et al. (2015)	–	–	72.22	–	–	60.50	–	–	56.37	63.03
Wiseman et al. (2015)	76.23	69.31	72.60	66.07	55.83	60.52	59.41	54.88	57.05	63.39
Wiseman et al. (2016)	77.49	69.75	73.42	66.83	56.95	61.50	62.14	53.85	57.70	64.21
NN Mention Ranker	79.77	69.10	74.05	69.68	56.37	62.32	63.02	53.59	57.92	64.76
NN Cluster Ranker	78.93	69.75	**74.06**	70.08	56.98	**62.86**	62.48	55.82	**58.96**	**65.29**
CoNLL 2012 Chinese Test Data										
Chen & Ng (2012)	64.69	59.92	62.21	60.26	51.76	55.69	51.61	58.84	54.99	57.63
Björkelund & Kuhn (2014)	69.39	62.57	65.80	61.64	53.87	57.49	59.33	54.65	56.89	60.06
NN Mention Ranker	72.53	65.72	68.96	65.49	56.87	60.88	61.93	57.11	59.42	63.09
NN Cluster Ranker	73.85	65.42	**69.38**	67.53	56.41	**61.47**	62.84	57.62	**60.12**	**63.66**

Table 5: Comparison with the current state-of-the-art approaches on the CoNLL 2012 test sets. NN Mention Ranker and NN Cluster Ranker are contributions of this work.

End-to-end Neural Coreference Resolution

Kenton Lee† , Luheng He† , Mike Lewis‡ , and Luke Zettlemoyer†∗

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Our model (ensemble)	**81.2**	**73.6**	**77.2**	**72.3**	**61.7**	**66.6**	**65.2**	**60.2**	**62.6**	**68.8**
Our model (single)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Durrett and Klein (2014)	72.6	69.9	71.2	61.2	56.4	58.7	56.2	54.2	55.2	61.7
Björkelund and Kuhn (2014)	74.3	67.5	70.7	62.7	55.0	58.6	59.4	52.3	55.6	61.6
Durrett and Klein (2013)	72.9	65.9	69.2	63.6	52.5	57.5	54.3	54.4	54.3	60.3

Table 1: Results on the test set on the English data from the CoNLL-2012 shared task. The final column (Avg. F1) is the main evaluation metric, computed by averaging the F1 of MUC, B^3, and $CEAF_{\phi_4}$. We improve state-of-the-art performance by 1.5 F1 for the single model and by 3.1 F1.

SpanBERT

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy

	MUC			B^3			$CEAF_{\phi_4}$			
	P	R	F1	P	R	F1	P	R	F1	Avg. F1
Prev. SotA: (Lee et al., 2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
Google BERT	84.9	82.5	83.7	76.7	74.2	75.4	74.6	70.1	72.3	77.1
Our BERT	85.1	83.5	84.3	77.3	75.5	76.4	75.0	71.9	73.9	78.3
Our BERT-1seq	85.5	84.1	84.8	77.8	76.7	77.2	75.3	73.5	74.4	78.8
SpanBERT	**85.8**	**84.8**	**85.3**	**78.3**	**77.9**	**78.1**	**76.4**	**74.2**	**75.3**	**79.6**

Table 3: Performance on the OntoNotes coreference resolution benchmark. The main evaluation is the average F1 of three metrics: MUC, B^3, and $CEAF_{\phi_4}$ on the test set.