Realistic Evaluation Principles for Cross-document Coreference Resolution

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi and Ido Dagan

*SEM 2021

Cross-document Coreference Resolution

U.S President nominates new surgeon general: MacArthur “genius grant” fellow Regina Benjamin on the July 13, 2009. Obama emphasize his decision in “her extensive and distinguished career in medicine”.

President Obama will name Dr. Regina Benjamin as U.S. Surgeon General on Monday in a Rose Garden announcement, later this morning.

News that Barack Obama may name Dr. Sanjay Gupta of Emory University and CNN as his Surgeon General has caused a spasm of celebrity reporting.

CNN’s management confirmed that Dr. Gupta had been approached by the Obama team on March 2009. The chief medical correspondent has declined comment.

Cross-document Coreference Resolution

U.S President nominates new surgeon general: MacArthur "genius grant" fellow Regina Benjamin on the July 13, 2009. Obama emphasize his decision in "her extensive and distinguished career in medicine".

President Obama will name Dr. Regina Benjamin as U.S. Surgeon General on Monday in a Rose Garden announcement, later this morning.

News that Barack Obama may name Dr. Sanjay Gupta of Emory University and CNN as his Surgeon General has caused a spasm of celebrity reporting.

CNN's management confirmed that Dr. Gupta had been approached by the Obama team on March 2009. The chief medical correspondent has declined comment.

Applications of CD Coreference Resolution

- Multi-document summarization
- Multi-hop question answering
- Knowledge Base Construction
- Entity Linking
- ……

Related Work

- Iterative algorithm for joint event and entity coreference (Barhom et al., 2019)
- Follow-up improvement using a paraphrase corpus (Meged et al., 2020)
- More recent Transformer-based models (Zeng et al, 2020; Caciularu et al., 2021)
- Results > 80 CoNLL F1

Related Work

- Iterative algorithm for joint event and entity coreference (Barhom et al., 2019)
- Follow-up improvement using a paraphrase corpus (Meged et al., 2020)
- More recent Transformer-based models (Zeng et al, 2020; Caciularu et al., 2021)
- Results > 80 CoNLL F1

But downstream applications don't use these models, why??

Unrealistic Evaluation

High reported results don’t reflect performance in real-world scenarios

Unrealistic Evaluation

High reported results don’t reflect performance in real-world scenarios

1. Evaluating only on **gold** mentions
   → coreference resolution involves also mention detection

Unrealistic Evaluation

High reported results don’t reflect performance in real-world scenarios

1. Evaluating only on **gold** mentions
   → coreference resolution involves also mention detection

2. Rewarding **singleton** prediction in coreference metrics

Unrealistic Evaluation

High reported results don’t reflect performance in real-world scenarios

1. Evaluating only on **gold** mentions
   → coreference resolution involves also mention detection

2. Rewarding **singleton** prediction in coreference metrics

3. Sidestepping a major lexical ambiguity challenge

Unrealistic Evaluation

High reported results don’t reflect performance in real-world scenarios

1. Evaluating only on gold mentions
   → coreference resolution involves also mention detection

2. Rewarding **singleton** prediction in coreference metrics

3. Sidestepping a major lexical ambiguity challenge

Downstream Application of Coreference Resolution

Consider this question-answering example from Quoref (Dasigi et al., 2019)

Anna and Declan eventually make their way on foot to a roadside pub, where they discover the three van thieves going through Anna's luggage. Declan fights them, displaying unexpected strength for a man of his size, and retrieves Anna's bag.

Who does Declan get into a fight with?

Downstream Application of Coreference Resolution

Consider this question-answering example from Quoref (Dasigi et al., 2019)

Anna and Declan eventually make their way on foot to a roadside pub, where they discover the **three van thieves** going through Anna's luggage. Declan fights **them**, displaying unexpected strength for a man of his size, and retrieves Anna's bag.

Who does Declan get into a fight with?

Downstream Application of Coreference Resolution

Consider this question-answering example from Quoref (Dasigi et al., 2019)

Anna and Declan eventually make their way on foot to a roadside pub, where they discover the **three van thieves** going through Anna's luggage. Declan fights **them**, displaying unexpected strength for a man of his size, and retrieves Anna's bag.

Who does Declan get into a fight with?

Three van thieves

Downstream Application of Coreference Resolution

Consider this question-answering example from Quoref (Dasigi et al., 2019)

Anna and Declan eventually make their way on foot to a roadside pub, where they discover the **three van thieves** going through Anna's luggage. Declan fights **them**, displaying unexpected strength for a man of his size, and retrieves Anna's bag.

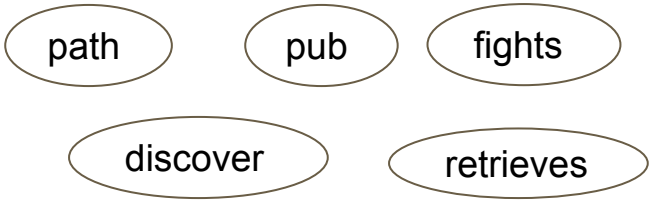
Who does Declan get into a fight with?

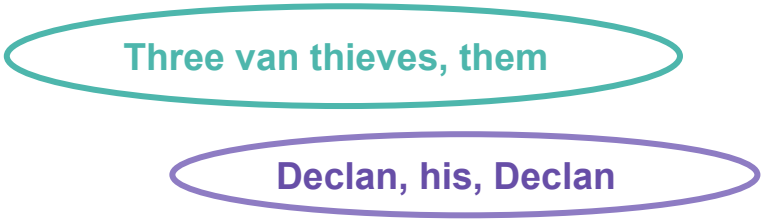
Three van thieves

*Downstream tasks leverage coreference **links** to bridge information across mentions*

Singleton Effect

Gold:





Singleton Effect

Gold:

path pub fights

discover retrieves

Three van thieves, them

Declan, his, Declan

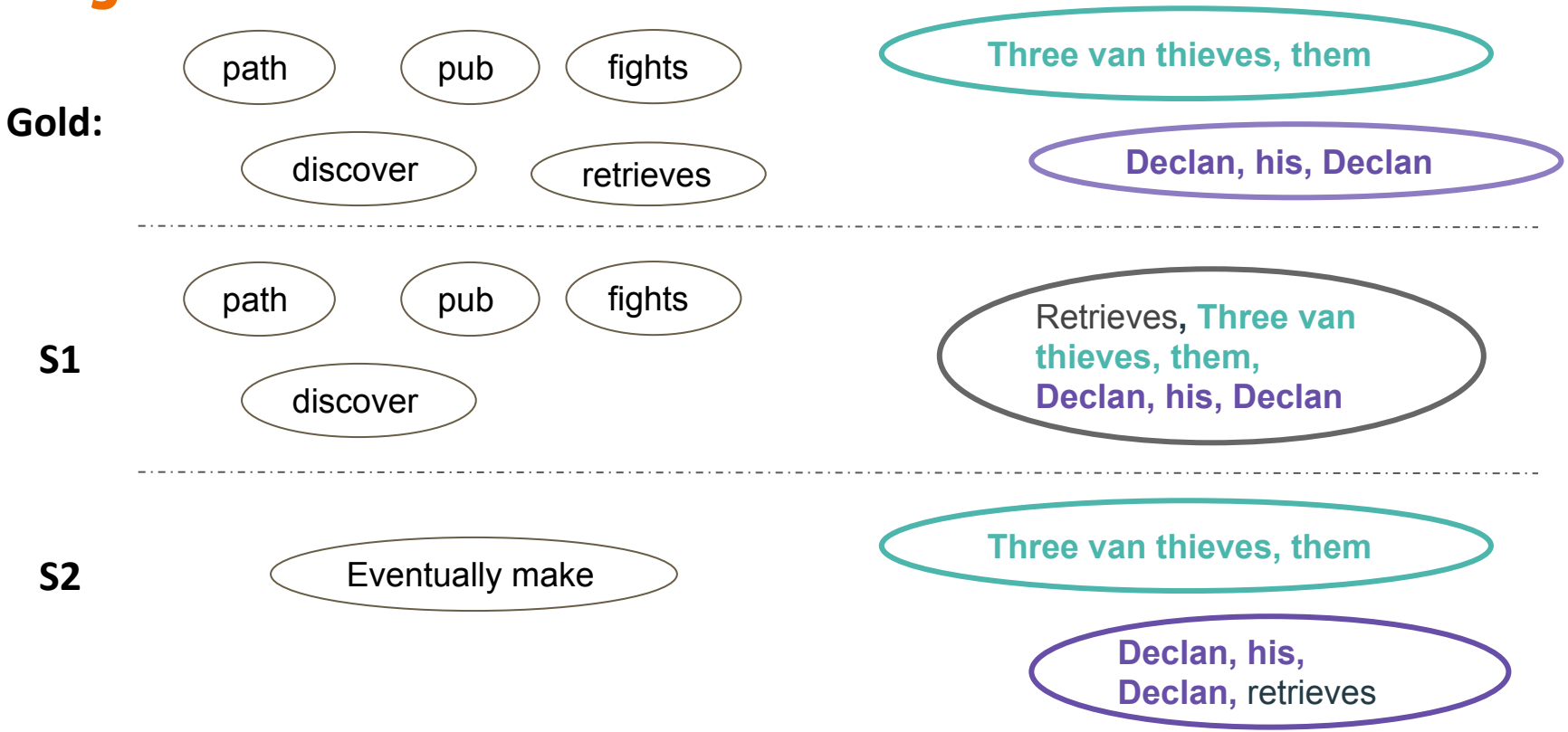
S1

path pub fights

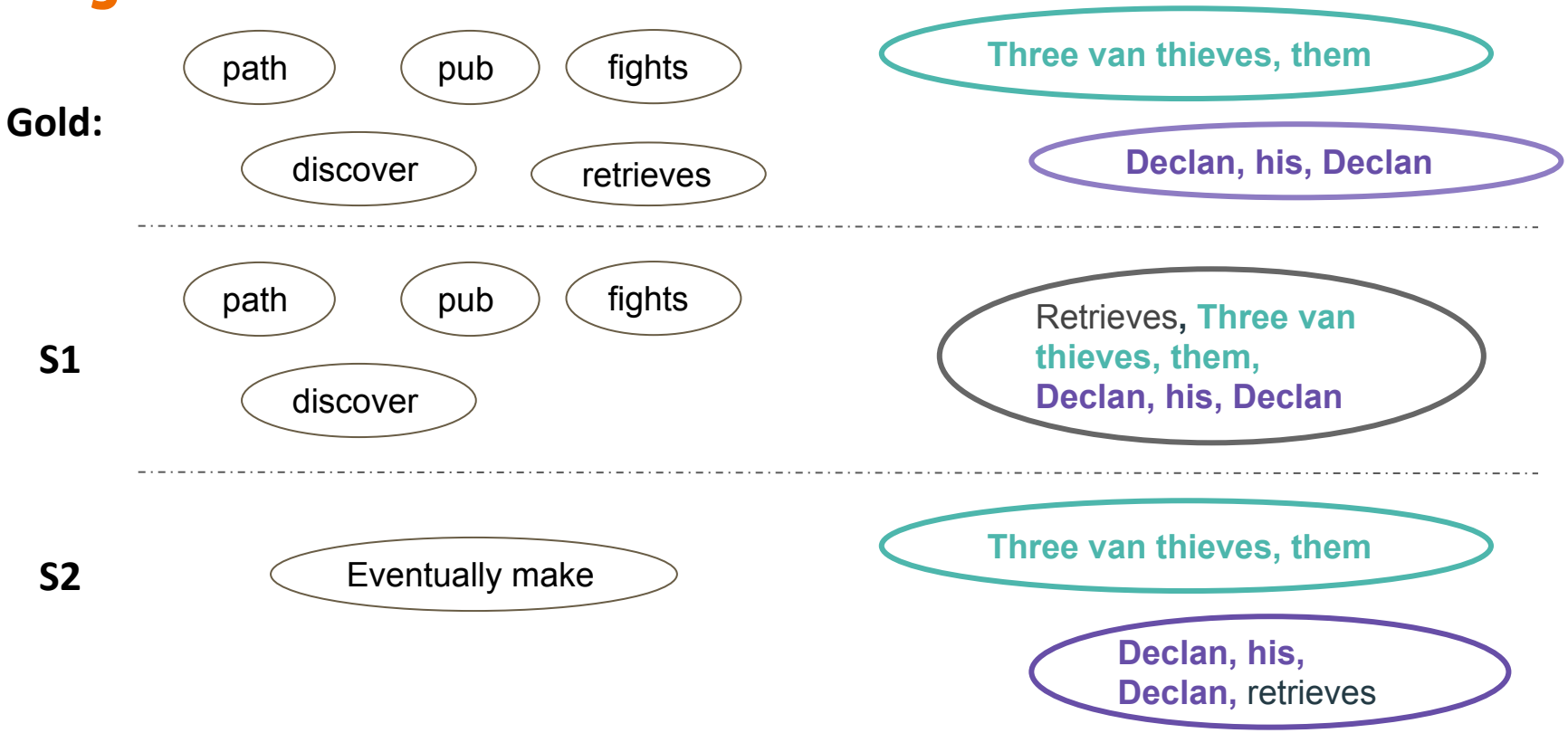
discover

Retrieves, Three van thieves, them, Declan, his, Declan

Singleton Effect



Singleton Effect



From a coreference resolution perspective, S2 is a better model!

Evaluation with or without singletons

		MUC	B^3	CEAFe	LEA	CoNLL
With Singletons	$S1$	75.0	**77.6**	**77.8**	**69.0**	**76.8**
	$S2$	**85.7**	59.2	32.7	50.0	59.2
CoNLL-2012	$S1$	75.0	53.1	44.4	42.1	57.5
	$S2$	**85.7**	**83.9**	**90.0**	**80.0**	**86.5**

Including singletons in coreference metrics may lead to counterproductive results for downstream tasks!

Evaluation with or without singletons

		MUC	B^3	CEAFe	LEA	CoNLL
With Singletons	$S1$	75.0	**77.6**	**77.8**	**69.0**	**76.8**
	$S2$	**85.7**	59.2	32.7	50.0	59.2
CoNLL-2012	$S1$	75.0	53.1	44.4	42.1	57.5
	$S2$	**85.7**	**83.9**	**90.0**	**80.0**	**86.5**

- Focusing the evaluation on **links** → S2 gets higher results than S1 as expected

Evaluation with or without singletons

		MUC	B^3	CEAFe	LEA	CoNLL
With Singletons	$S1$	75.0	**77.6**	**77.8**	**69.0**	**76.8**
	$S2$	**85.7**	59.2	32.7	50.0	59.2
CoNLL-2012	$S1$	75.0	53.1	44.4	42.1	57.5
	$S2$	**85.7**	**83.9**	**90.0**	**80.0**	**86.5**

- Focusing the evaluation on **links** → S2 gets higher results than S1 as expected
- Models are still penalized for wrongly **linking** singletons

Evaluation with or without singletons

		MUC	B^3	CEAFe	LEA	CoNLL
With Singletons	$S1$	75.0	**77.6**	**77.8**	**69.0**	**76.8**
	$S2$	**85.7**	59.2	32.7	50.0	59.2
CoNLL-2012	$S1$	75.0	53.1	44.4	42.1	57.5
	$S2$	**85.7**	**83.9**	**90.0**	**80.0**	**86.5**

- Focusing the evaluation on **links** → S2 gets higher results than S1 as expected
- Models are still penalized for wrongly **linking** singletons

But singletons are still valuable ….

Evaluation with or without singletons

		MUC	B^3	CEAFe	LEA	CoNLL
With Singletons	$S1$	75.0	**77.6**	**77.8**	**69.0**	**76.8**
	$S2$	**85.7**	59.2	32.7	50.0	59.2
CoNLL-2012	$S1$	75.0	53.1	44.4	42.1	57.5
	$S2$	**85.7**	**83.9**	**90.0**	**80.0**	**86.5**

Decouple the evaluation of mention detection from coreference resolution

Evaluation with or without singletons

		MUC	B^3	CEAFe	LEA	CoNLL
With Singletons	$S1$	75.0	**77.6**	**77.8**	**69.0**	**76.8**
	$S2$	**85.7**	59.2	32.7	50.0	59.2
CoNLL-2012	$S1$	75.0	53.1	44.4	42.1	57.5
	$S2$	**85.7**	**83.9**	**90.0**	**80.0**	**86.5**

Decouple the evaluation of mention detection from coreference resolution

Mention detection (including singletons) is a **span detection** task → Span F1

Coreference is a **linking** task → CoNLL-2012 (without singletons)

Evaluation with or without singletons

		MUC	B^3	CEAFe	LEA	CoNLL
With Singletons	*S*1	75.0	77.6	77.8	69.0	76.8
	*S*2	**85.7**	59.2	32.7	50.0	59.2
CoNLL-2012	*S*1	75.0	53.1	44.4	42.1	57.5
	*S*2	**85.7**	**83.9**	**90.0**	**80.0**	**86.5**

Coreference Results

	Recall	Precision	F1
S1	**100**	**100**	**100**
S2	60.0	75.0	66.7

Mention detection Results

Decouple the evaluation of mention detection from coreference resolution

Mention detection (including singletons) is a **span detection** tasks → Span F1

Coreference is a **linking** task → CoNLL-2012 (without singletons)

Unrealistic Evaluation

High reported results don’t reflect performance in real-world scenarios

1. Evaluating only on gold mentions
   → coreference resolution involves also mention detection

2. Rewarding **singleton** prediction in coreference metrics

3. Sidestepping a major lexical ambiguity challenge

History of ECB+ – main benchmark

1. EventCorefBank (ECB) *(Bejan and Harabagiu, 2008)*
   Partial annotation of cross-document **event** coreference on 43 topics

2. Extended ECB (EECB) *(Lee et al., 2012)*
   Exhaustive annotation, and adding **entity** coreference in ECB

3. ECB+ *(Cybulska and Vossen, 2014)*
   Adding **subtopics** (+500 documents!) to challenge models with **lexical ambiguity** to mimic real use cases on a reasonable annotation task

ECB+ – Nomination as Surgeon General topic

U.S President nominates new surgeon general: MacArthur "genius grant" fellow Regina Benjamin on the July 13, 2009. Obama emphasize his decision in "her extensive and distinguished career in medicine".

President Obama will name Dr. Regina Benjamin as U.S. Surgeon General on Monday in a Rose Garden announcement, later this morning.

Subtopic 1

News that Barack Obama may name Dr. Sanjay Gupta of Emory University and CNN as his Surgeon General has caused a spasm of celebrity reporting.

CNN's management confirmed that Dr. Gupta had been approached by the Obama team on March 2009. The chief medical correspondent has declined comment.

Subtopic 2

Confronting Lexical Ambiguity

- Each topic includes exactly two subtopics with short articles.

- Recent work apply a very simple document clustering that reconstruct original subtopics

Confronting Lexical Ambiguity

- Each topic includes exactly two subtopics with short articles.

- Recent work apply a very simple document clustering that reconstruct original subtopics

❌ Bypassing ECB+ goals and sidestepping lexical ambiguity

Confronting Lexical Ambiguity

- Each topic includes exactly two subtopics with short articles.

- Recent work apply a very simple document clustering that reconstruct original subtopics

❌ Bypassing ECB+ goals and sidestepping lexical ambiguity

❌ Models performance on realistic use cases is not assessed

Confronting Lexical Ambiguity

- Each topic includes exactly two subtopics with short articles.

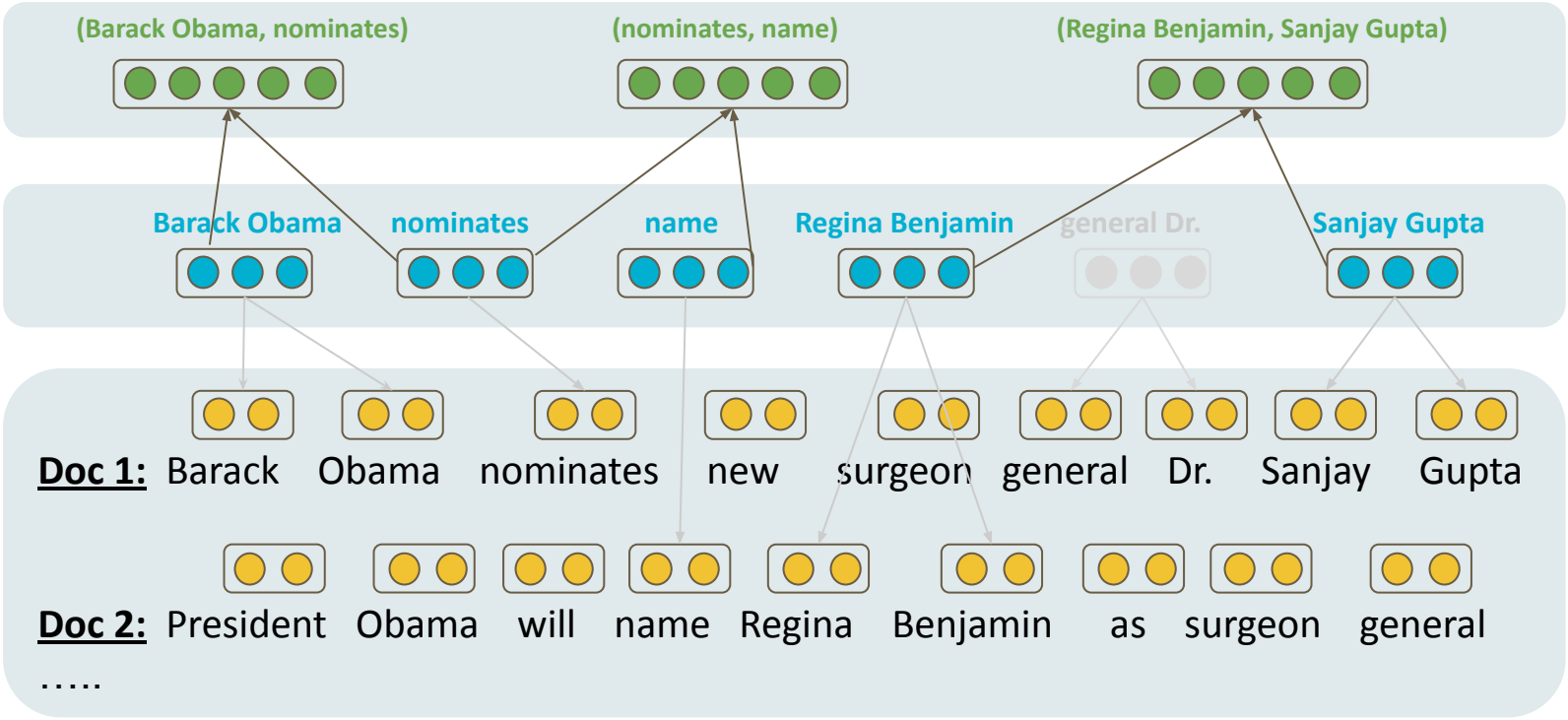
- Recent work apply a very simple document clustering that reconstruct original subtopics

❌ Bypassing ECB+ goals and sidestepping lexical ambiguity

❌ Models performance on realistic use cases is not assessed

✅ **We suggest that models will evaluate at the level of the entire topic without fine-grained subtopic clustering**

Experiments – E2e model (Cattan et al., 2021)



Results – Event coreference on ECB+

	CoNLL F1
Barhom et al. (2019)	79.5
Cattan et al. (2021)	81

Results – Event coreference on ECB+

	CoNLL F1
Barhom et al. (2019)	79.5
Cattan et al. (2021)	81
— singletons	71.1 **(-9.9)**

Results – Event coreference on ECB+

	CoNLL F1
Barhom et al. (2019)	79.5
Cattan et al. (2021)	81
— singletons	71.1 **(-9.9)**
— topic level	62.0 **(-9.1)**

Results – Event coreference on ECB+

	CoNLL F1
Barhom et al. (2019)	79.5
Cattan et al. (2021)	81
— singletons	71.1 **(-9.9)**
— topic level	62.0 **(-9.1)**
— predicted mentions	48.6 **(-13.4)**

Summary

- We propose 3 principles to assess realistic performance
  - Predicted mentions
  - Decouple coreference evaluation
  - Evaluation at the entire topic level

- Applying our evaluation methodology on a SOTA model results on a drop of 32.4 CoNLL F1 (!)

- Large room for improvement under realistic conditions

Thanks!

Questions?

Arie Cattan

github.com/ariecattan/coref

arie.cattan.github.io