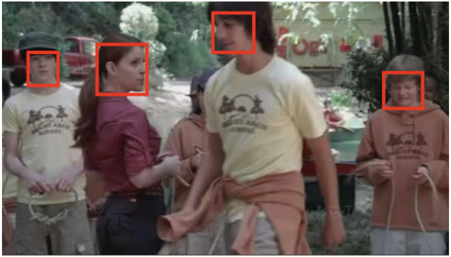
Linking People in Videos with Their Names Using Coreference Resolution

Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei

Stanford University

Images from Ramanathan et al. (2014)

Task



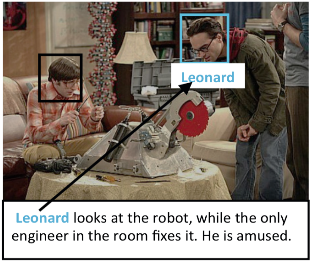
Missy points to the larger kid. The big kid walks off. Other kids jeer.

- No labelled instance. Script is the only source of supervision
- Names include nominal expressions and pronouns

Previous Approach

On person naming:

- Multiple instance learning, using proper names from script
- Treat videos and scripts as bag of face tracks and names
- Unidirectional information flow from text to vision

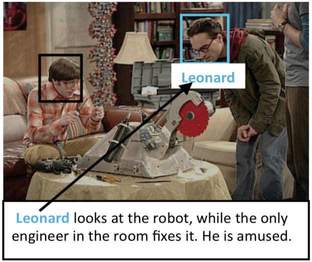


Leonard looks at the robot, while the only engineer in the room fixes it. He is amused.

Previous Approach

On person naming:

- Multiple instance learning, using proper names from script
- Treat videos and scripts as bag of face tracks and names
- Unidirectional information flow from text to vision



Leonard looks at the robot, while the only engineer in the room fixes it. He is amused.

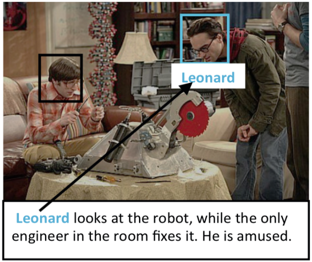
On coreference resolution:

- One of the core task in NLP
- Can operate on language alone
- Not accurate enough

Previous Approach

On person naming:

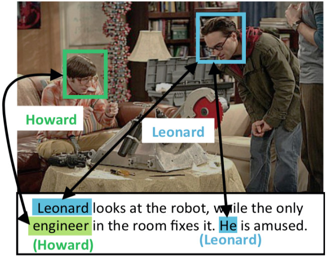
- Multiple instance learning, using proper names from script
- Treat videos and scripts as bag of face tracks and names
- Unidirectional information flow from text to vision



Leonard looks at the robot, while the only engineer in the room fixes it. He is amused.

On coreference resolution:

- One of the core task in NLP
- Can operate on language alone
- Not accurate enough



Leonard looks at the robot, while the only engineer in the room fixes it. He is amused.

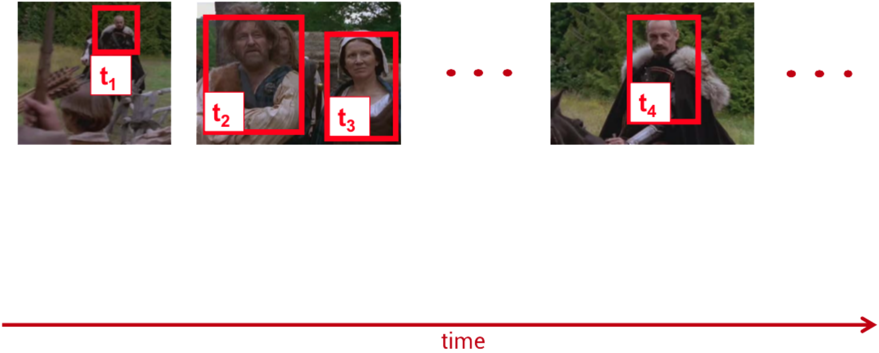
Problem Setup

Input:

Problem Setup

Input:

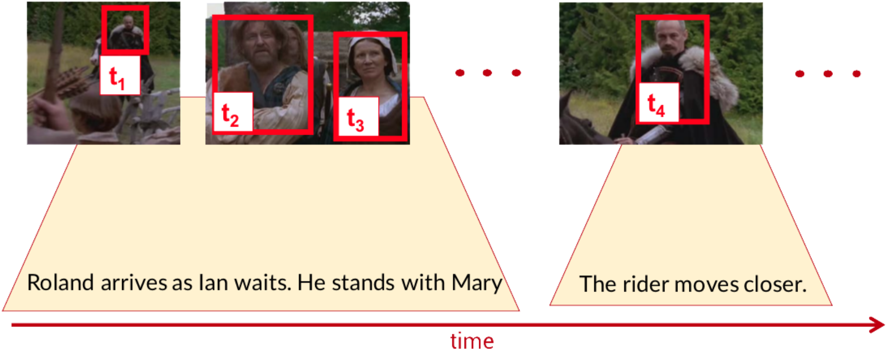
- Videos with detected human tracks



Problem Setup

Input:

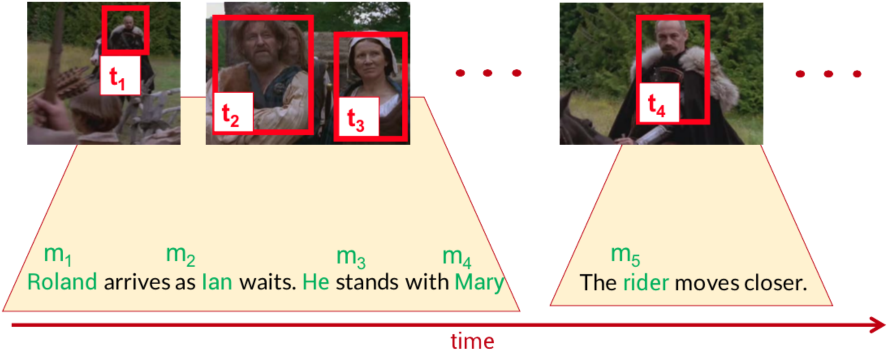
- Videos with detected human tracks
- Script roughly aligned with video segments



Problem Setup

Input:

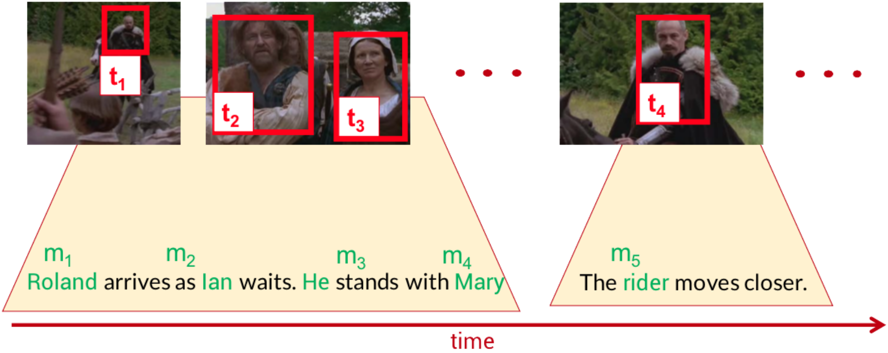
- Videos with detected human tracks
- Script roughly aligned with video segments
- Names (including pronoun/nominals) from script



Problem Setup

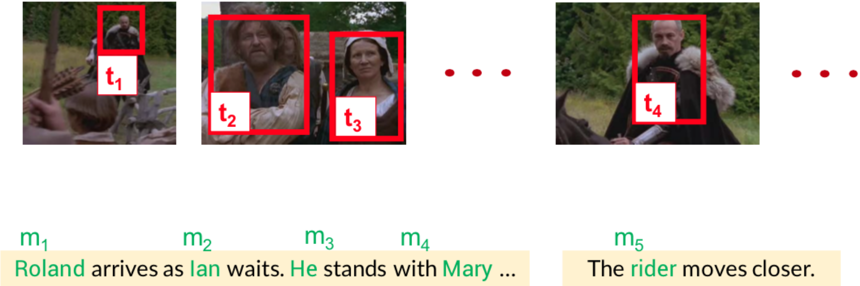
Input:

- Videos with detected human tracks
- Script roughly aligned with video segments
- Names (including pronoun/nominals) from script
- Cast names



Problem Setup

Output:



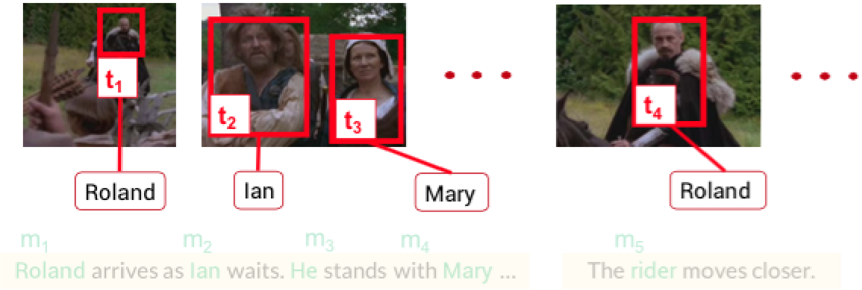
Roland arrives as Ian waits. He stands with Mary ...

The rider moves closer.

Problem Setup

Output:

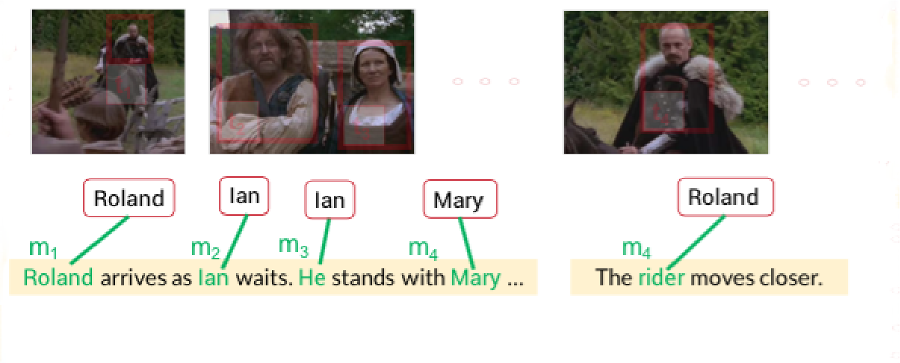
- Name assignment to human tracks in video



Problem Setup

Output:

- Name assignment to human tracks in video
- Name assignment to human mentions in text

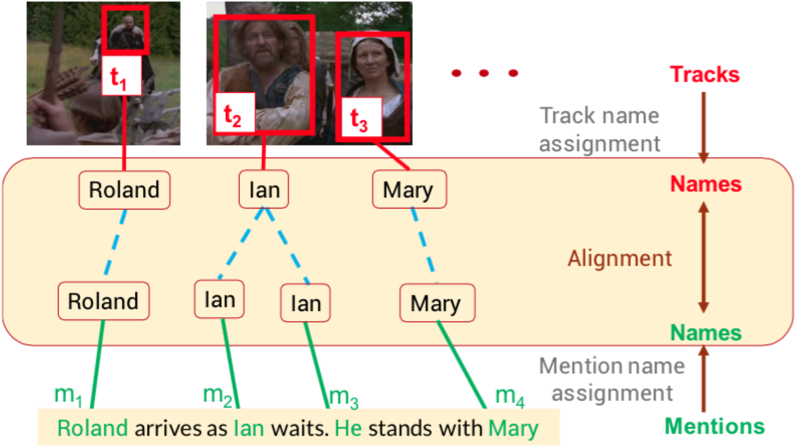


Roland arrives as Ian waits. He stands with Mary ...

The rider moves closer.

Proposed Method

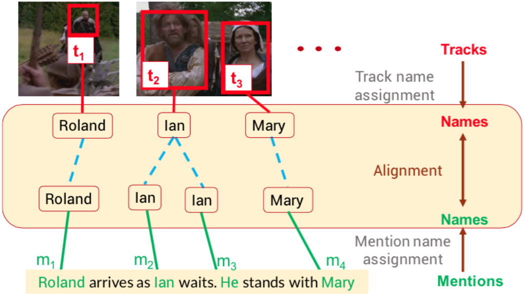
$$C = \gamma_t C_{track} + \gamma_m C_{mention} + C_{align}$$



Proposed Method

$$C = \gamma_t C_{track}(Y) + \gamma_m C_{mention}(Z, R) + C_{align}(A, Y, Z)$$

- Name-Track assignment $Y \in \{0,1\}^{T \times P}$
- Name-Mention assignment $Z \in \{0,1\}^{M \times P}$
- Antecedent matrix $R \in \{0,1\}^{M \times M}$
- Alignment matrix $A \in \{0,1\}^{T \times M}$



$C_{track}(Y)$

- Cost of assigning names to tracks
- Based on video features only
- Formulate cost function of regression based clustering

$$\begin{aligned} C(Y; X, \lambda) &= \arg\min_{W} \sum_{t \in \tau} ||Y - XW||_F^2 + \lambda ||W||_F^2 \\ &= tr(Y^T \Pi(X, \lambda) Y) \end{aligned}$$

Constraints:

- Each track is assigned to exactly one name
- Speaker should be aligned to at least one track
- Name not mentioned in a scene won't be aligned

$C_{mention}(Z, R)$

- Depends on text only
- Proper mentions(68%) are trivial to map
- Pronouns/Nominals alone are not informative
- Apply regression based clustering to predict R

Constraints:

- Each mention has at most one antecedent
- Each mention is assigned to one name
- Gender consistency/no self-association of pronouns
- Connection constraint $R_{m,m'} = 1 \rightarrow Z_m = Z_{m'}$

$C_{align}(A, Y, Z)$

Intuition

- Aligned track/mention should be assigned to the same name
- Tracks and mentions are ordered sequence through time
- Tracks and mentions are roughly aligned in time

Formulation

- Soft connection penalty

$$\min ||A^T Y - Z||_F^2$$

- Monotonic constraint
- Mention mapping constraint

Optimization

$$\begin{aligned} \min \gamma_t C_{track}(Y) + \gamma_m C_{mention}(Z, R) + C_{align}(A, Y, Z) \\ s.t. \quad Y \in C_Y, \qquad Z, R \in C_{Z,R}, \qquad A \in C_A \end{aligned}$$

- Relax Y, R, Z to be $[0, 1]$
- Slack constraints of Y, Z
- Block coordinate descent

Optimization

$$\begin{aligned} &\min \gamma_t C_{track}(Y) + \gamma_m C_{mention}(Z, R) + C_{align}(A, Y, Z) \\ &s.t. \quad Y \in C_Y, \qquad Z, R \in C_{Z,R}, \qquad A \in C_A \end{aligned}$$

- Relax Y, R, Z to be $[0, 1]$
- Slack constraints of Y, Z
- Block coordinate descent
- Quadratic programming to optimize Y

Optimization

$$\begin{aligned} \min\ & \gamma_t C_{track}(Y) + \gamma_m C_{mention}(Z,R) + C_{align}(A,Y,Z) \\ s.t.\ & Y \in C_Y, \qquad Z,R \in C_{Z,R}, \qquad A \in C_A \end{aligned}$$

- Relax Y, R, Z to be $[0,1]$
- Slack constraints of Y, Z
- Block coordinate descent
- Quadratic programming to optimize Y
- Quadratic programming to optimize Z, R

Optimization

$$\begin{aligned} \min \ & \gamma_t C_{track}(Y) + \gamma_m C_{mention}(Z, R) + C_{align}(A, Y, Z) \\ s.t. \ & Y \in C_Y, \qquad Z, R \in C_{Z,R}, \qquad A \in C_A \end{aligned}$$

- Relax Y, R, Z to be $[0, 1]$
- Slack constraints of Y, Z
- Block coordinate descent
- Quadratic programming to optimize Y
- Quadratic programming to optimize Z, R
- Dynamic time wrapping to optimize A

Optimization

$$\begin{aligned} \min \gamma_t C_{track}(Y) + \gamma_m C_{mention}(Z, R) + C_{align}(A, Y, Z) \\ s.t. \quad Y \in C_Y, \qquad Z, R \in C_{Z,R}, \qquad A \in C_A \end{aligned}$$

- Relax Y, R, Z to be $[0, 1]$
- Slack constraints of Y, Z
- Block coordinate descent
- Quadratic programming to optimize Y
- Quadratic programming to optimize Z, R
- Dynamic time wrapping to optimize A
- Round Y, Z to integer matrix

Dataset

We reveal Lynette holding Porter by his feet, while he clings to Preston's desk.

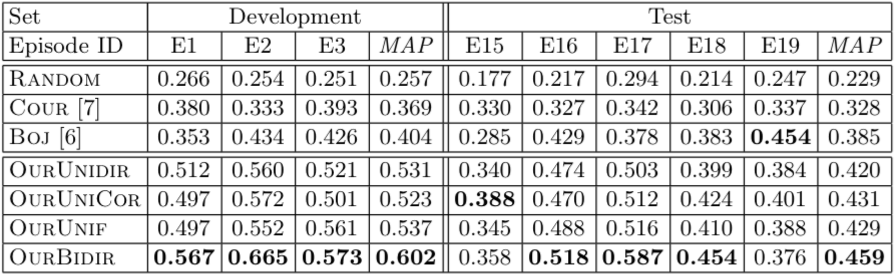
Missy points to the larger kid. The big kid walks off. Other kids jeer.

Cary eyes the siblings, as Alicia looks across the bullpen

		pronoun/nominal
Dev. Set (14 episodes)	**3329** tracks (3 eps.)	**811** mentions
Test Set (5 episodes)	**4757** tracks	**300** mentions

Quantitative Results

Name assignment to tracks in video.

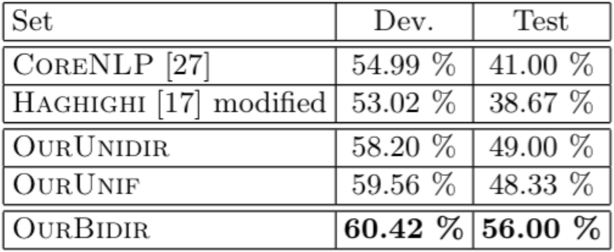


Set	Development				Test					
Episode ID	E1	E2	E3	*MAP*	E15	E16	E17	E18	E19	*MAP*
Random	0.266	0.254	0.251	0.257	0.177	0.217	0.294	0.214	0.247	0.229
Cour [7]	0.380	0.333	0.393	0.369	0.330	0.327	0.342	0.306	0.337	0.328
Boj [6]	0.353	0.434	0.426	0.404	0.285	0.429	0.378	0.383	**0.454**	0.385
OurUnidir	0.512	0.560	0.521	0.531	0.340	0.474	0.503	0.399	0.384	0.420
OurUniCor	0.497	0.572	0.501	0.523	**0.388**	0.470	0.512	0.424	0.401	0.431
OurUnif	0.497	0.552	0.561	0.537	0.345	0.488	0.516	0.410	0.388	0.429
OurBidir	**0.567**	**0.665**	**0.573**	**0.602**	0.358	**0.518**	**0.587**	**0.454**	0.376	**0.459**

- Random: Randomly picks a name based on crude alignment
- Cour: Weakly-supervised method for name assignment
- BOJ: min C_{track} without scene constraint
- OurUnidir: min C_{track} with scene constraint
- OurUnicor: min C_{track} with coreference constraints
- OurUnif: All tracks given equal values in alignment matrix
- OurBidir: Full model

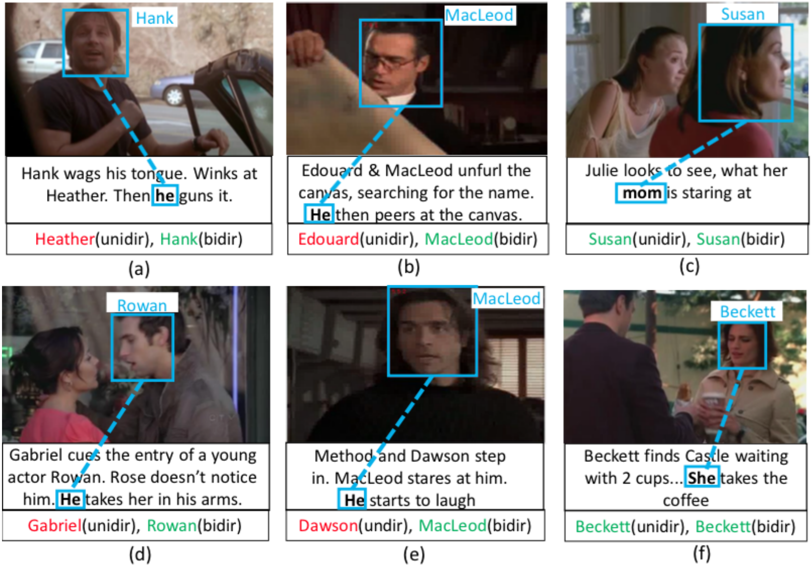
Quantitative Results

Name assignment to mentions in text.



Set	Dev.	Test
CoreNLP [27]	54.99 %	41.00 %
Haghighi [17] modified	53.02 %	38.67 %
OurUnidir	58.20 %	49.00 %
OurUnif	59.56 %	48.33 %
OurBidir	**60.42 %**	**56.00 %**

Qualitative Results



Hank wags his tongue. Winks at Heather. Then **he** guns it.

Heather(unidir), Hank(bidir)

(a)

Edouard & MacLeod unfurl the canvas, searching for the name. **He** then peers at the canvas.

Edouard(unidir), MacLeod(bidir)

(b)

Julie looks to see, what her **mom** is staring at

Susan(unidir), Susan(bidir)

(c)

Gabriel cues the entry of a young actor Rowan. Rose doesn't notice him. **He** takes her in his arms.

Gabriel(unidir), Rowan(bidir)

(d)

Method and Dawson step in. MacLeod stares at him. **He** starts to laugh

Dawson(undir), MacLeod(bidir)

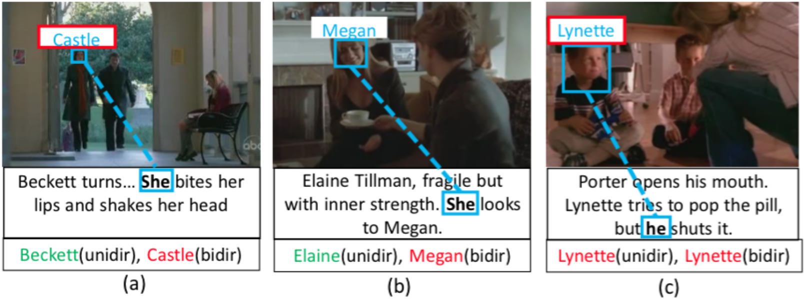
(e)

Beckett finds Castle waiting with 2 cups... **She** takes the coffee

Beckett(unidir), Beckett(bidir)

(f)

Errors



Castle
Beckett turns... She bites her lips and shakes her head
Beckett(unidir), Castle(bidir)
(a)
Megan
Elaine Tillman, fragile but with inner strength. She looks to Megan.
Elaine(unidir), Megan(bidir)
(b)
Lynette
Porter opens his mouth. Lynette tries to pop the pill, but he shuts it.
Lynette(unidir), Lynette(bidir)
(c)

- Missing/low resolution faces
- Error in coreference resolution

Summary

Contribution:

- Joint person naming and coreference resolution
- New dataset
- State-of-the-art performance on visual/textual side

Summary

Contribution:

- Joint person naming and coreference resolution
- New dataset
- State-of-the-art performance on visual/textual side

Future work:

- Actions/attributes for alignment

V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking People in Videos with “Their” Names Using Coreference Resolution. In *Computer Vision – ECCV 2014*, pages 95–110. Springer International Publishing, Cham, Sept. 2014.