2019 ACL Short paper

Coreference Resolution with Entity Equalization

Ben Kantor
Tel Aviv University
benkantor@mail.tau.ac.il

Amir Globerson
Tel Aviv University
amir.globerson@gmail.com

mention — **John** told **Sally** that **she** should come watch **him** play the **violin**.

antecedent — John told **Sally** that **she** should come watch him play the violin.

coreferent — John told **Sally** that **she** should come watch him play the violin.

ϵ

non-anaphoric — John told Sally that she should come watch him play the **violin**.

span

General Electric said the Postal Service contacted the company

Coreference Resolution in Two Steps

1. Detect the mentions (easy)

   "[I] voted for [Nader] because [he] was most aligned with [[my] values]," [she] said

   - mentions can be nested!

2. Cluster the mentions (hard)

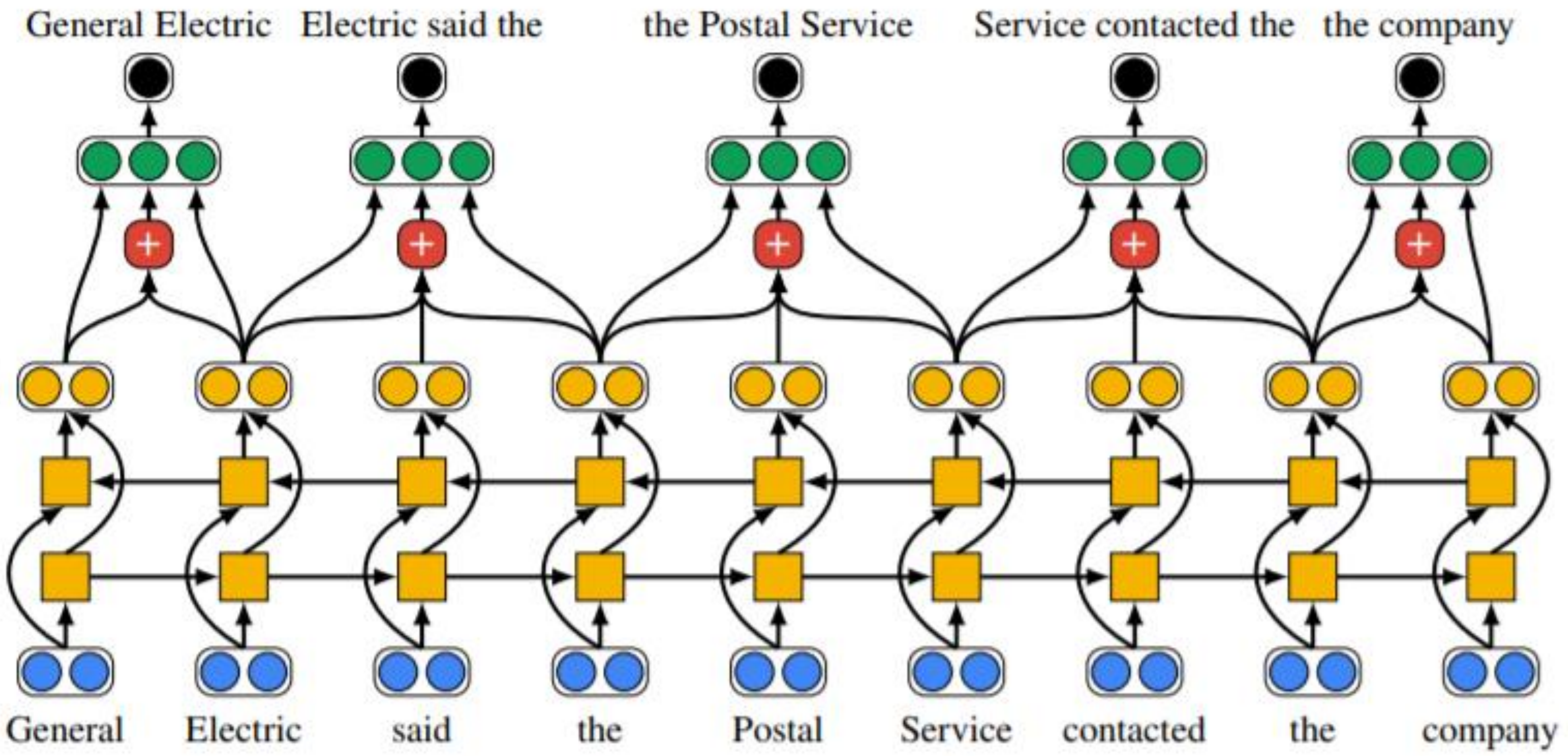
   "[I] voted for [Nader] because [he] was most aligned with [[my] values]," [she] said

Four Kinds of Coreference Models

- Rule-based (pronominal anaphora resolution)
- Mention Pair
- Mention Ranking
- Clustering

End-to-end Model

- Current state-of-the-art model for coreference resolution (Kenton Lee et al. from UW, EMNLP 2017)
- Mention ranking model
- Improvements over simple feed-forward NN
  - Use an LSTM
  - Use attention
  - Do mention detection and coreference end-to-end



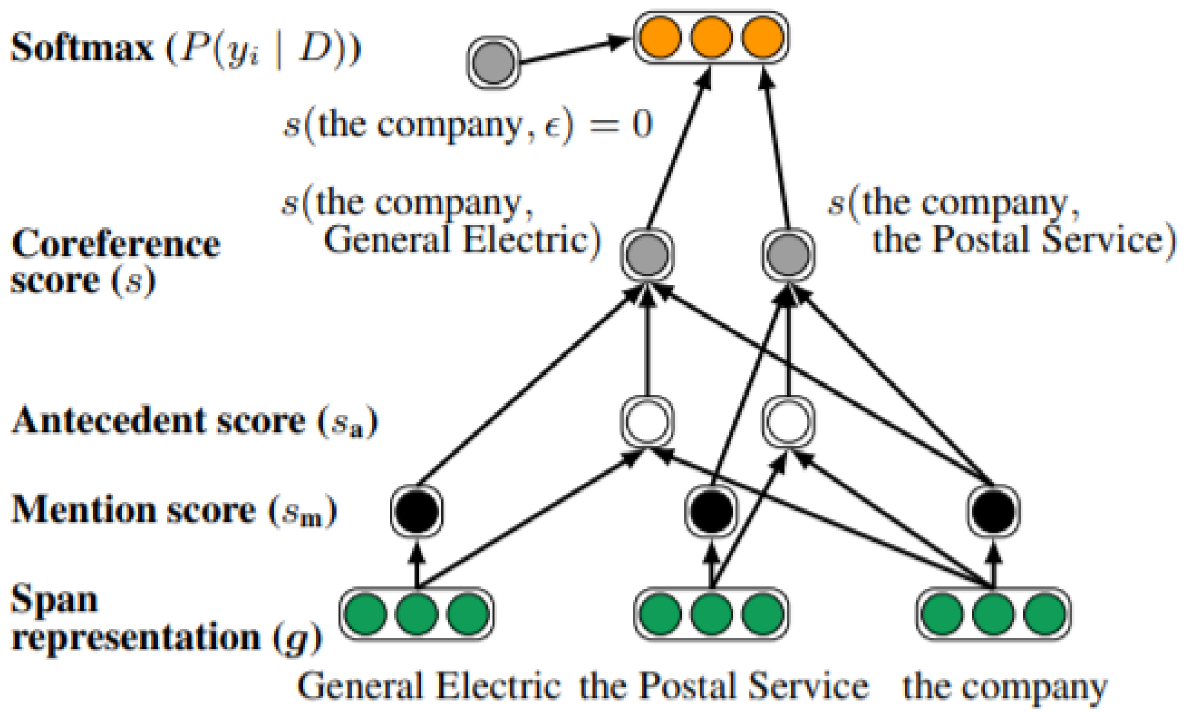
$$\boldsymbol{g}_i = [\boldsymbol{x}^*_{\text{START}(i)}, \boldsymbol{x}^*_{\text{END}(i)}, \hat{\boldsymbol{x}}_i, \phi(i)]$$

$$\alpha_t = \boldsymbol{w}_\alpha \cdot \text{FFNN}_\alpha(\boldsymbol{x}^*_t)$$

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum\limits_{k=\text{START}(i)}^{\text{END}(i)} \exp(\alpha_k)}$$

$$\hat{\boldsymbol{x}}_i = \sum_{t=\text{START}(i)}^{\text{END}(i)} a_{i,t} \cdot \boldsymbol{x}_t$$

[EMNLP2017] End-to-end Neural Coreference Resolution



$$s(i,j) = \begin{cases} 0 & j = \epsilon \\ s_{\text{m}}(i) + s_{\text{m}}(j) + s_{\text{a}}(i,j) & j \neq \epsilon \end{cases}$$

$$s_{\text{m}}(i) = \boldsymbol{w}_{\text{m}} \cdot \text{FFNN}_{\text{m}}(\boldsymbol{g}_i)$$

$$s_{\text{a}}(i,j) = \boldsymbol{w}_{\text{a}} \cdot \text{FFNN}_{\text{a}}([\boldsymbol{g}_i, \boldsymbol{g}_j, \boldsymbol{g}_i \circ \boldsymbol{g}_j, \phi(i,j)])$$

[EMNLP2017] End-to-end Neural Coreference Resolution

Motivation

- Entity Equalization

Speaker 1: Um and **[I]** think that is what's - Go ahead Linda.
Speaker 2: Well and uh thanks goes to **[you]** and to the media to help us... So our hat is off to **[all of you]** as well.

[2019ACL] Coreference Resolution with Entity Equalization

Motivation

- Entity Equalization

- Entity Equalization VS. Antecedent Averaging

[John] went to the park and [he] got tired. [John] decided to go back home.

	$John_1$	he	$John_2$
$John_1$	1	0	0
he	1	0	0
$John_2$	1	0	0

[2018NAACL] Higher-order Coreference Resolution with Coarse-to-fine Inference

[2019ACL] Coreference Resolution with Entity Equalization

Baseline Model

Higher-order

$$\boldsymbol{a}_i = \sum_{y_i \in \mathcal{Y}(i)} P(y_i) \cdot \boldsymbol{g}_{y_i}$$

$$\boldsymbol{f}_i = f_f(\boldsymbol{g}_i, \boldsymbol{a}_i)$$

$$\boldsymbol{g}_i' = \boldsymbol{f}_i \circ \boldsymbol{g}_i + (\boldsymbol{1} - \boldsymbol{f}_i) \circ \boldsymbol{a}_i$$

$$P(y_i) = \frac{e^{s(i, y_i)}}{\sum_{y' \in \mathcal{Y}(i)} e^{s(i, y')}}$$

$$P'(y_i) = \frac{e^{s(\boldsymbol{g}_i', \boldsymbol{g}_{y_i}')}}{\sum_{y \in \mathcal{Y}(i)} e^{s(\boldsymbol{g}_i', \boldsymbol{g}_y')}}$$

[2018NAACL] Higher-order Coreference Resolution with Coarse-to-fine Inference

Baseline Model

Coarse-to-fine Inference

$$s_c(i, j) = \boldsymbol{g}_i^\top \mathbf{W}_c \, \boldsymbol{g}_j$$

$$s(i, j) = s_m(i) + s_m(j) + s_c(i, j) + s_a(i, j)$$

First stage Keep the top M spans based on the mention score $s_m(i)$ of each span.

Second stage Keep the top K antecedents of each remaining span i based on the first three factors, $s_m(i) + s_m(j) + s_c(i, j)$.

Third stage The overall coreference $s(i, j)$ is computed based on the remaining span pairs. The

[2018NAACL] Higher-order Coreference Resolution with Coarse-to-fine Inference

Entity Equalization

$$Q(i \in E_j) = \begin{cases} \sum_{k=j}^{i-1} P(y_i = k) \cdot Q(k \in E_j) & \text{if } j < i \\ P(y_i = \epsilon) & \text{if } j = i \\ 0 & \text{if } j > i \end{cases}$$

$$\boldsymbol{a}_i = \sum_{y_i \in \mathcal{Y}(i)} P(y_i) \cdot \boldsymbol{g}_{y_i}$$

$$\boldsymbol{e}_i^{(t)} = \sum_{j=1}^{t} Q(j \in E_i) \cdot \boldsymbol{g}_j$$

$$\boldsymbol{a}_i = \sum_{j=1}^{i} Q(i \in E_j) \cdot \boldsymbol{e}_j^{(i)}$$

[2019ACL Short Paper] Coreference Resolution with Entity Equalization

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Lee et al. (2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
+ BERT	**83.51**	82.8	83.16	**74.51**	74.14	74.32	71.93	70.6	71.26	76.25
– Second-order	82.61	83.48	83.04	73.56	75.44	74.49	71.6	**71.55**	71.57	76.37
+ EE (Ours)	82.63	**84.14**	**83.38**	73.31	**76.17**	**74.71**	**72.37**	71.14	**71.75**	**76.61**

Table 1: Results on the test set of the English CoNLL-2012 shared task. The average F1 of MUC, B^3 and $CEAF_{\phi_4}$ is the main evaluation metric.

[2019ACL Short Paper] Coreference Resolution with Entity Equalization

2019 ACL Short paper

The Referential Reader: A Recurrent Entity Network for Anaphora Resolution

Fei Liu *
The University of Melbourne
Victoria, Australia

Luke Zettlemoyer
Facebook AI Research
University of Washington
Seattle, USA

Jacob Eisenstein
Facebook AI Research
Seattle, USA

- Coreference with named entities

text　Barack Obama　Obama

world

- Anaphora

text　Barack Obama　he

world

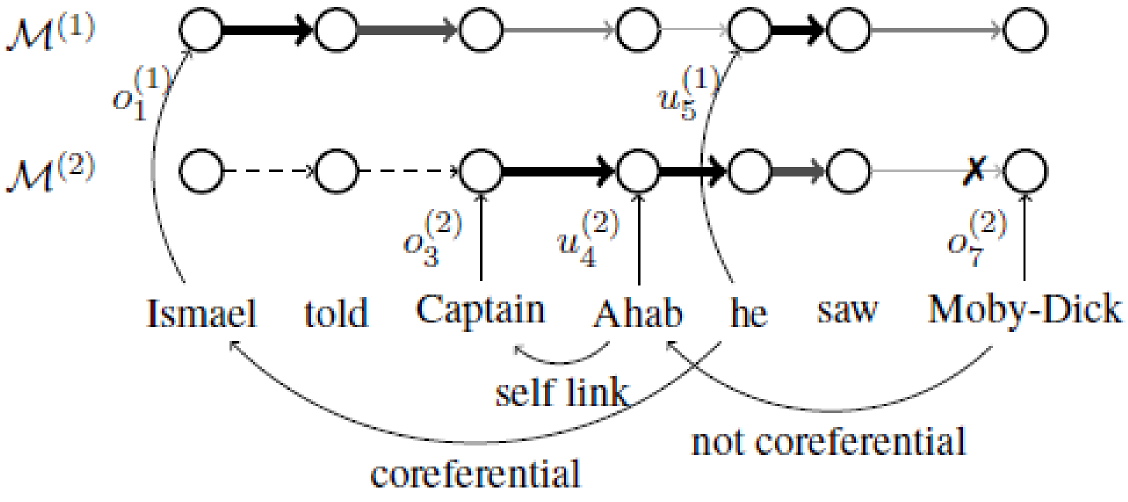


Figure 1: A referential reader with two memory cells. Overwrite and update are indicated by $o_t^{(i)}$ and $u_t^{(i)}$; in practice, these operations are continuous gates. Thickness and color intensity of edges between memory cells at neighboring steps indicate memory salience; ✗ indicates an overwrite.

As each token is encountered, the reader must decide whether to:

(a) link the token to an existing memory, thereby creating a coreference link,

(b) overwrite an existing memory and store a new entity,

(c) disregard the token and move ahead.

As memories are reused, their salience increases, making them less likely to be overwritten.

[2019ACL Short Paper] The Referential Reader: A Recurrent Entity Network for Anaphora Resolution

Model

For a given document consisting of a sequence of tokens $\{w_t\}_{t=1}^{T}$, we represent text at two levels:

- Tokens: represented as $\{\boldsymbol{x}_t\}_{t=1}^{T}$, where the vector $\boldsymbol{x}_t \in \mathbb{R}^{D_x}$ is computed from any token-level encoder.
- Entities: represented by a fixed-length memory $\mathcal{M}_t = \{(\boldsymbol{k}_t^{(i)}, \boldsymbol{v}_t^{(i)}, s_t^{(i)})\}_{i=1}^{N}$, where each memory is a tuple of a key $\boldsymbol{k}_t^{(i)} \in \mathbb{R}^{D_k}$, a value $\boldsymbol{v}_t^{(i)} \in \mathbb{R}^{D_v}$, and a salience $s_t^{(i)} \in [0, 1]$.

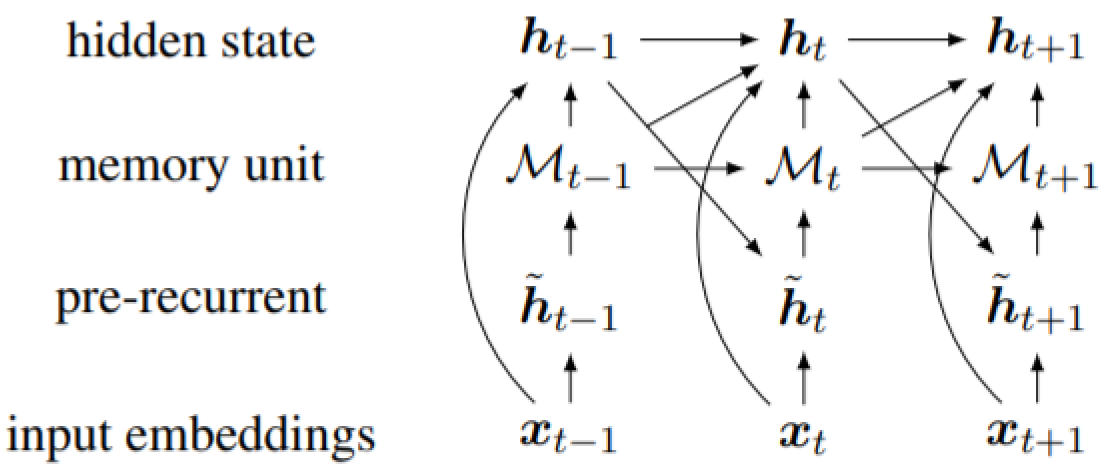


Figure 2: Overview of the model architecture.

[2019ACL Short Paper] The Referential Reader: A Recurrent Entity Network for Anaphora Resolution

Recurrent Unit

$$\boldsymbol{m}_t = \sum_{i=1}^{N} s^{(i)} \boldsymbol{v}_t^{(i)}.$$

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t).$$

$$\boldsymbol{h}_t = \mathrm{GRU}(\boldsymbol{x}_t, (1 - c_t) \times \boldsymbol{h}_{t-1} + c_t \times \boldsymbol{m}_t)$$

$$c_t = \min(\sigma(\boldsymbol{W}_c \tilde{\boldsymbol{h}}_t + b_c), \textstyle\sum_i s_t^{(i)})$$

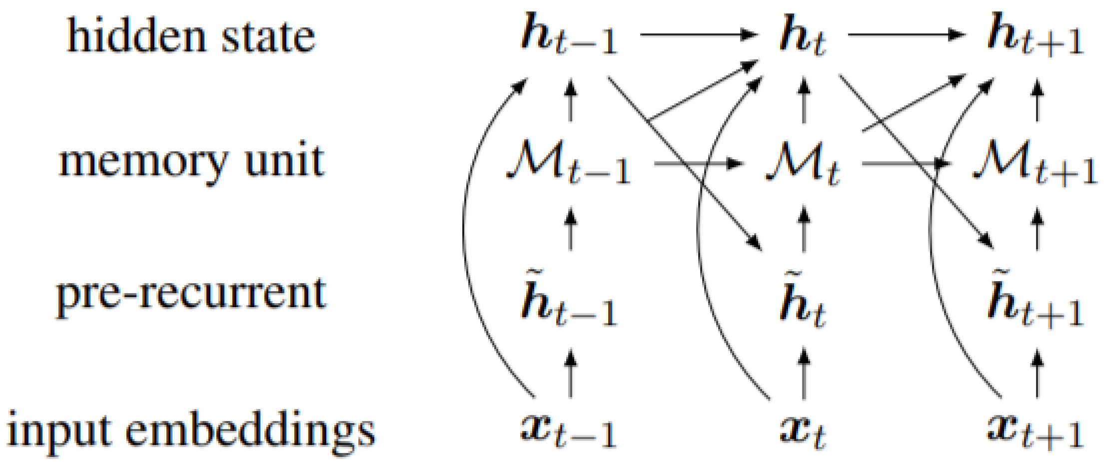


Figure 2: Overview of the model architecture.

[2019ACL Short Paper] The Referential Reader: A Recurrent Entity Network for Anaphora Resolution

Memory Unit

memory gates $\{(u_t^{(i)}, o_t^{(i)})\}_{i=1}^N$;

entity gate $e_t = \sigma(\boldsymbol{\phi}_e \cdot \tilde{\boldsymbol{h}}_t)$

reference gate $r_t = \sigma(\boldsymbol{\phi}_r \cdot \tilde{\boldsymbol{h}}_t) \times e_t$;

Updating existing entities

query vector, $\boldsymbol{q}_t = f_q(\tilde{\boldsymbol{h}}_t)$

attention scores, $\alpha_t^{(i)} = r_t \times \mathrm{SoftMax}(\boldsymbol{k}_{t-1}^{(i)} \cdot \boldsymbol{q}_t + b)$

update gate: $u_t^{(i)} = \min(\alpha_t^{(i)}, 2s_{t-1}^{(i)})$.

Storing new entities.

$$\tilde{o}_t = e_t - \textstyle\sum_{i=1}^N u_t^{(i)}$$

overwrite the memory with the lowest salience.

$$o_t^{(i)} = \tilde{o}_t \times \mathrm{GSM}^{(i)}(-\boldsymbol{s}_{t-1}, \tau)$$

$$\boldsymbol{s}_t = \{s_t^{(i)}\}_{i=1}^N$$

Memory salience.

$$r_t^{(i)} = 1 - u_t^{(i)} - o_t^{(i)}$$

$$\lambda_t = (e_t \times \gamma_e + (1 - e_t) \times \gamma_n)$$

$$s_t^{(i)} = \lambda_t \times r_t^{(i)} \times s_{t-1}^{(i)} + u_t^{(i)} + o_t^{(i)};$$

Memory state.

$$\tilde{\boldsymbol{k}}_t = f_k(\tilde{\boldsymbol{h}}_t) \quad \tilde{\boldsymbol{v}}_t = f_v(\tilde{\boldsymbol{h}}_t)$$

$$\boldsymbol{k}_t^{(i)} = u_t^{(i)} \mathrm{GRU}_k(\boldsymbol{k}_{t-1}^{(i)}, \tilde{\boldsymbol{k}}_t) + o_t^{(i)} \tilde{\boldsymbol{k}}_t + r_t^{(i)} \boldsymbol{k}_{t-1}^{(i)}$$

$$\boldsymbol{v}_t^{(i)} = u_t^{(i)} \mathrm{GRU}_v(\boldsymbol{v}_{t-1}^{(i)}, \tilde{\boldsymbol{v}}_t) + o_t^{(i)} \tilde{\boldsymbol{v}}_t + r_t^{(i)} \boldsymbol{v}_{t-1}^{(i)}.$$

[2019ACL Short Paper] The Referential Reader: A Recurrent Entity Network for Anaphora Resolution

Coreference Chains

$$\omega_{t_1,t_2}^{(i)} = \prod_{t=t_1+1}^{t_2} (1 - o_t^{(i)})$$

$$\hat{\psi}_{t_1,t_2} = \sum_{i=1}^{N} (u_{t_1}^{(i)} + o_{t_1}^{(i)}) \times u_{t_2}^{(i)} \times \omega_{t_1,t_2}^{(i)}.$$

Training

cross-entropy $\sum_{i=1}^{T} \sum_{j=i+1}^{T} H(\hat{\psi}_{i,j}, y_{i,j})$

[2019ACL Short Paper] The Referential Reader: A Recurrent Entity Network for Anaphora Resolution

	F_1^M	F_1^F	$\frac{F_1^F}{F_1^M}$	F_1
Clark and Manning (2015)†	53.9	52.8	0.98	53.3
Lee et al. (2017)†	67.7	60.0	0.89	64.0
Lee et al. (2017), re-trained	67.8	66.3	0.98	67.0
Parallelism†	69.4	64.4	0.93	66.9
Parallelism+URL†	72.3	68.8	0.95	70.6
RefReader, LM objective‡	61.6	60.5	0.98	61.1
RefReader, coref objective‡	69.6	68.1	0.98	68.9
RefReader, LM + coref‡	**72.8**	**71.4**	**0.98**	**72.1**
RefReader, coref + BERT⋆	**80.3**	**77.4**	**0.96**	**78.8**

Table 1: GAP test set performance. †: reported in Webster et al. (2018); ‡: strictly incremental processing; ⋆: average over 5 runs with different random seeds.

$$\sum_{i=1}^{T} \sum_{j=i+1}^{T} H(\hat{\psi}_{i,j}, y_{i,j}).$$

$$P(w_{t+1} \mid \boldsymbol{h}_t)$$

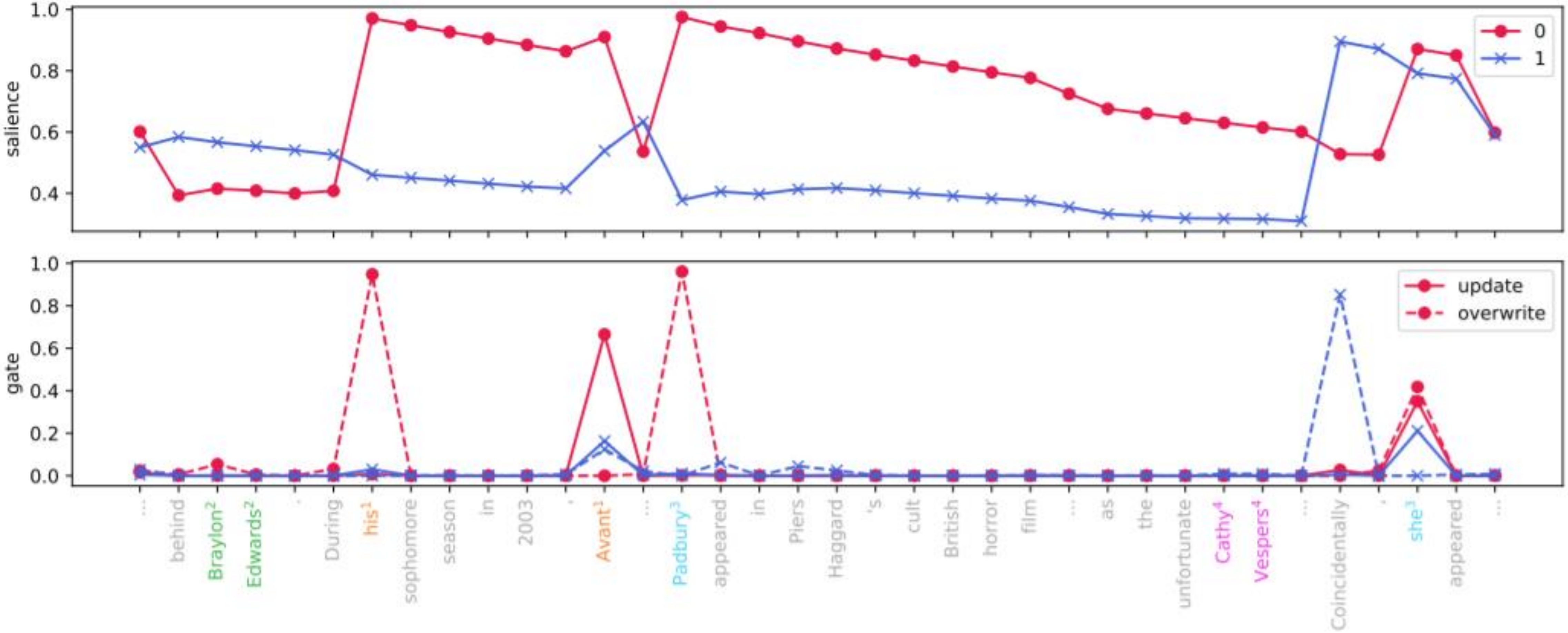


Figure 3: An example of the application the referential reader to a concatenation of two instances from GAP. The ground truth is indicated by the color of each token on the x-axis as well as the superscript.