Linking people in videos with “their” names using coreference resolution (Supplementary Material)

Vignesh Ramanathan*, Armand Joulin†, Percy Liang†, Li Fei-Fei†

*Department of Electrical Engineering, Stanford University
†Computer Science Department, Stanford University
{vigneshr, armand, pliang, feifeili}@cs.stanford.edu

1 Episodes from the dataset

The episodes which are part of Development and Test set are shown in Tab. 1.

Development Set	**Test Set**
1. Numb3rs 3x11	15. Highlander 5x14
2. Castle 1x03	16. Highlander 5x20
3. Highlander 5x02	17. Castle 1x09
4. Highlander 5x06	18. The Mentalist 1x19
5. The Mentalist 1x08	19. Californication 1x01
6. The Mentalist 1x22	
7. The Mentalist 3x11	
8. The Good Wife 1x10	
9. The Good Wife 1x20	
10. Twin Peaks 2x03	
11. Desperate Housewives 1x04	
12. 30 Rock 1x12	
13. Sliders 4x02	
14. Numb3rs 3x05	

Table 1. The episodes used in our experiments are shown.

2 Optimization in terms of the alignment matrix

We first show that $\|A^TY - Z\|_F^2$ is equivalent to $-2\text{tr}\left(A^TYZ\right)$ plus an additive constant, and then we provide the Dynamic Programming algorithm for optimizing $\text{tr}\left(A^TYZ\right)$ with respect to M .

We note that $Z\mathbf{1}_P = \mathbf{1}_M$ and $Z \in \{0,1\}^{M\times P}$. Hence, exactly M elements of the matrix are set to 1. This results in $\|Z\|_F^2 = \text{tr}(Z^TZ) = M$. Similarly, for the matrix A^TY, we can show that $\|A^TY\|_F^2 = \text{tr}\left(A^TYY^TA\right) = M$.

$$\begin{aligned}\|A^TY - Z\|_F^2 &= \mathrm{tr}\left(Z^TZ\right) + \mathrm{tr}\left(AA^TYY^T\right) - 2\mathrm{tr}\left(Y^TAZ\right) \\ &= 2M - 2\mathrm{tr}\left(Y^TAZ\right).\end{aligned} \tag{1}$$

Note that, in the absence of the integer relaxation, the matrix (A^TY) has the same properties as Z, where each row sums to 1, and the elements are in $\{0,1\}$. Hence, $Tr(Y^TAA^TY) = \mathrm{tr}(Z^TZ) = M$. Hence, minimizing $\|A^TY - Z\|_F^2$ is equivalent to maximizing $\mathrm{tr}(Y^TAZ)$.

We denote by $\mathcal{T}_m$ the set of tracks which are aligned with a mention m, based on the crude alignment of the descriptions with the video. The Dynamic program to maximize $\mathrm{tr}(Y^TAZ)$ with respect to A is shown in Algo. 1.

```
Data: $Y \in \mathbb{P}_{TP},\ Z \in \mathbb{P}_{MP},\ \{\mathcal{T}_m,\ \forall\, m \in \mathcal{M}\}$
Result: $A \in \{0,1\}^{T\times M}$
Initialization:
$\tilde{A} \leftarrow ZY^T$;
for $m = 0 \rightarrow M$ do
    for $t = 0 \rightarrow T$ do
        $C_{tm} \leftarrow 0, I_{tm} \leftarrow 0$;
    end
end
Cost update:
for $m = 1 \rightarrow M$ do
    for $t = 1 \rightarrow T$ do
        $A_{tm} \leftarrow 0$;
        if $t \notin \mathcal{T}_m$ then
            $\tilde{A}_{tm} \leftarrow -\infty$
        end
        if $C_{t-1m} \leq C_{tm-1} + \tilde{A}_{tm}$ then
            $C_{tm} \leftarrow C_{tm-1} + \tilde{A}_{tm}$;
            $I_{tm} \leftarrow t$;
        end
        $C_{tm} \leftarrow C_{t-1m}$;
        $I_{tm} \leftarrow I_{t-1m}$;
    end
end
Backtracking:
$t \leftarrow T, m \leftarrow M$;
while $m \geq 1$ do
    $t \leftarrow I_{tm}$;
    $A_{tm} \leftarrow 1$;
    $m \leftarrow m - 1$;
end
```

Algorithm 1: Dynamic program algorithm for optimizing with respect to A

3 Coreference features

The coreference features are based on the standard features used in coreference resolution systems such as [2, 1]. They include two sets of features corresponding to (i) features between a pair of different mentions, and (ii) features extracted from a single mention. These features are concatenated to form the final pairwise coreference feature $\Phi_{ij}^{\text{mention}}$, between the mentions i, j. The first set of features are active when the two mentions are different $(i \neq j)$, and the second set of features are active when the two mentions are the same $(i = j)$. All the features are discretized and represented by binary vectors.

The coreference features between a pair of different mentions, are briefly explained below.

1. *Sentence distance*: The number of sentences between the two mentions.
2. *Parse tree distance*: The distance between the mentions on the semantic parse tree.
3. *Word distance*: The number of words between the two mentions.
4. *Animacy agreement*: Indicates if the two mentions agree on animacy values.
5. *Gender agreement*: Indicates if the two mentions agree on gender values.
6. *Cardinality agreement*: Indicates if the two mentions agree on cardinality.
7. *Head word agreement*: Indicates if the mentions have the same headword.
8. *Inside*: Indicates if one mention is contained inside the other.
9. *Appositions*: Indicates if a mention is the apposition of the other.
10. *Role Appositions*: Indicates if a mention is the role apposition of the other.
11. *Predicate Nominative*: Indicates if a mention is the predicate nominative of the other.

The coreference features extracted from a single mention are explained below:

1. *Mention type*: The mention type such as pronoun, nominal, or a proper noun.
2. *Subject*: Indicates if the mention is a subject in the sentence.
3. *Direct Object*: Indicates if the mention is a direct object.
4. *Gender*: Gender of the mention.
5. *Animacy*: Animacy of the mention.
6. *Cardinality*: Cardinality indicating if the mention is singleton or pronoun.
7. *Presence in cast list*: Indicates if the word corresponding to the mention is part of the cast list $\mathcal{P}$.

4 Additional constraints for the mention naming model

First, we show the computation of the matrix $B(\Phi^{\text{mention}}, \lambda^{\text{mention}})$ used in the clustering cost for coreference resolution of the main paper. Next, we explain the complete formulation of the mention naming model to include additional constraints such as gender agreement.

The $M^2 \times M^2$ matrix B is obtained by first computing the $M^2 \times M^2$ coreference feature kernel K_c. Each element in K_c corresponds to two pairs of mentions.

Since the i^{th} element in the $M^2 \times M^2$ vector vec(R) corresponds to $(i_{row}, i_{col})^{th}$ element in the matrix R, where $i_{row} = \lceil \frac{i}{M} \rceil$, and $i_{col} = i - M \lfloor \frac{2i-1}{2M} \rfloor$, we use the same notation while computing the kernel matrix. The $(i,j)^{th}$ element in this kernel K_c^{ij} is shown below.

$$K_c^{ij} = \Phi^{\text{mention}}_{i_{row} i_{col}} \cdot \Phi^{\text{mention}}_{j_{row} j_{col}} \tag{2}$$

The matrix B is then computed as follows,

$$B(\Phi^{\text{mention}}, \lambda^{\text{mention}}) = \lambda^{\text{mention}} \mathbf{\Pi} \left(\mathbf{\Pi} K_c \mathbf{\Pi} + M^2 \lambda^{\text{mention}} \mathbf{I} \right)^{-1} \mathbf{\Pi}, \tag{3}$$

where $\mathbf{\Pi} = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{\mathbf{1}^T\mathbf{1}}$ and λ^{mention} is the regularization parameter.

Gender constraint. Let g_m denote the gender of the mention $m \in \mathcal{M}$. Two mentions i, j are connected to each other only if their genders g_i and g_j are equal. This is a very valuable cue, as noted in most of the coreference resolution systems such as [1, 2].

Pronoun constraint. Let $\mathcal{M}_{pro.}$ be the set of pronouns from the script. A mention belonging to this set is not allowed to connect to itself as an antecedent. This constraint forces the name corresponding to a pronoun to be obtained from another mention.

Cast constraint. Let $\mathcal{M}_{cas.}$ be the set of mentions such that, the word corresponding to the mention is the same as a person name from our cast list $\mathcal{P}$. For instance, the mention "John" in the sentence "John eats an apple", if the cast list includes the name "John". Let, $p_m \in \mathcal{P}$ be the cast name corresponding to a mention $m \in \mathcal{M}_{cas.}$. For such a mention m, we enforce R_{mm} to be equal to 1, and the corresponding element Z_{mp_m} in the matrix Z to be equal to 1.

5 Modified version of coreference resolution from Haghighi and Klein [1]

This baseline used for comparison in Tab. 2 of our paper can be viewed as a probabilistic version of our unidirectional model. The model assumes that every *mention* is associated with an antecedent mention, occurring before it in the text. The choice of the antecedent mention is given by the $M \times M$ matrix U with entries in $0, 1$, where the $(i,j)^{th}$ element is set to 1, if the mention i is the antecedent mention j. The probability of choosing i as the antecedent to j is associated with the coreference feature $\Phi^{\text{mention}}_{ij}$ and a weight vector w_m. The coreference model is as shown in Eq. 4.

$$p(Z, U|\Phi^{\text{mention}}; w_m) \propto \Phi(U, \Phi^{\text{mention}}, w_m) \prod_i \Psi(z_i, \cdots, z_1, u_i) \tag{4}$$

$$\Phi(u_{ij}, \Phi^{\text{mention}}, w_m) = \exp\left(\mathbf{1}(u_{ij} = 1) w_m \cdot \Phi_{ij}^{\text{mention}}\right) \ \ \forall i \leq j$$

$$\Psi(z_i, \cdots, z_1, u_i) = \begin{Bmatrix} 1 - \epsilon, & \text{if } \forall \ \ u_{ji} = 1, \ z_i = z_j \\ \epsilon, & \text{otherwise} \end{Bmatrix}$$

where ϵ is a small value, z_i is the i^{th} column in Z, u_{ij} is the $(i,j)^{th}$ element of U, and u_i is its i^{th} column.

In addtion to these factors, we also assign priors based on different constraints, similar to our unidirectional model described in the previous section. Following [1], we learn the model through a mean-field approximation. We assume a distribution $q_z(Z)$ for Z and $q_u(U)$ for U. The coreference resolution can then be performed by solving the optimization problem shown below:

$$\max_{q_z, q_s, w_m} \mathbb{E}_{q_z, q_u}\left[L(Z, U|\Phi^{\text{mention}}; w_m)\right] + H_q,$$

where $L(Z, U|\Phi^{\text{mention}}; w_m)$ is the log-likelihood of p form Eq. 4 and H_q is the entropy of q_z, q_u.

References

1. Haghighi, A., Klein, D.: Coreference resolution in a modular, entity-centered model. In: HLT-NAACL (2010)
2. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D.: Stanford's mulit-pass sieve coreference resolution system at the conll-2011 shared task. In: CoNLL-2011 Shared Task (2011)