

From Co-occurrence to Correspondence

Ben Taskar

T. Cour, K. Ganchev, J. Graca, C. Jordan, B. Sapp



University of Pennsylvania

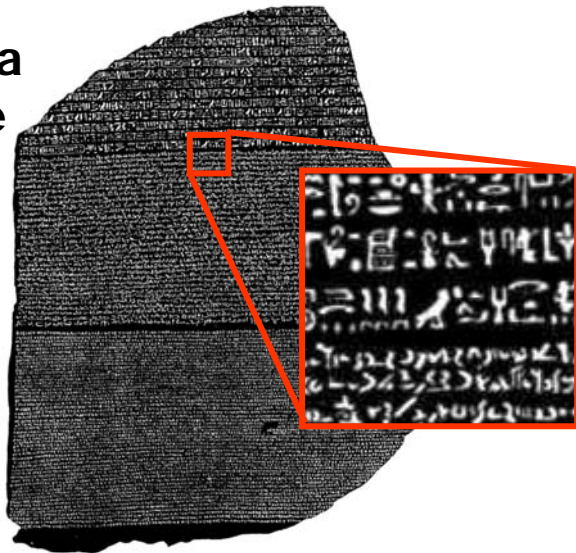


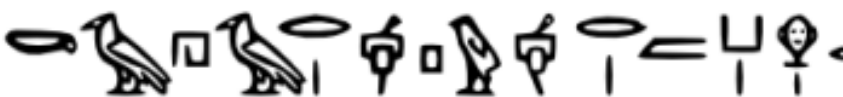
Learning from Co-Occurrence

Foreign Language Lexicon

Visual Lexicon

Rosetta Stone



K A L A R E S A W S A R E M H A H E R




HURLEY: Uh ... the Chinese people have water.
(Sayid and Kate go to check it out.)

[EXT. BEACH - CRASH SITE]

(Sayid holds the empty bottle in his hand and questions Sun.)

SAYID: (quietly)
Where did you get this?
(He looks at her.)

[EXT. JUNGLE]

(Sawyer is walking through the jungle. He reaches a spot. He kneels down and looks back to check that no one's followed him.)



SAYID



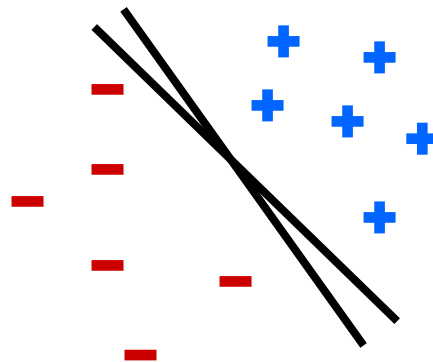
SUN



BOTTLE

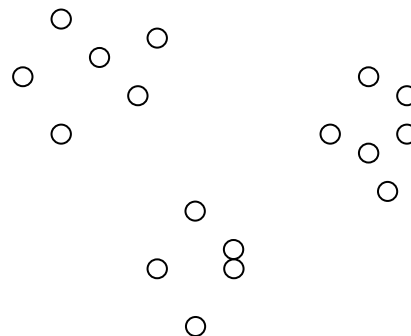
Supervision in Learning

- Supervised



- Co-occurrence?

- Unsupervised

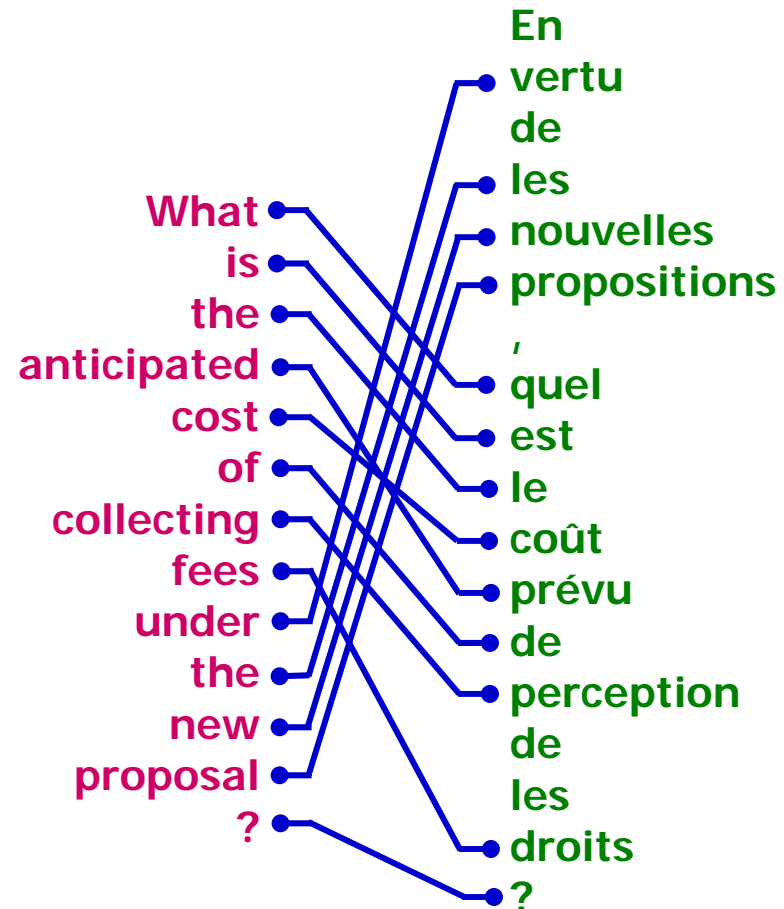


Word-Level Correspondence

- Key step in statistical machine translation systems

What is the anticipated cost of collecting fees under the new proposal?

En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?



Movie/Script Correspondence



scene

shot

...
HURLEY: Uh ... the Chinese people have water.
(Sayid and Kate go to check it out.)

[EXT. BEACH - CRASH SITE]

(Sayid holds the empty bottle in his hand and questions Sun.)

SAYID: (quietly) Where did you get this?
(He looks at her.)

[EXT. JUNGLE]

(Sawyer is walking through the jungle. He reaches a spot. He kneels down and looks back to check that no one's followed him.)

...



Levels of alignment
Temporal:
Scene/Shot/Thread
Script/closed captions

Within modalities:
Pronoun resolution
Face tracking/recognition

Across modalities:
Person/Object/Action
correspondence

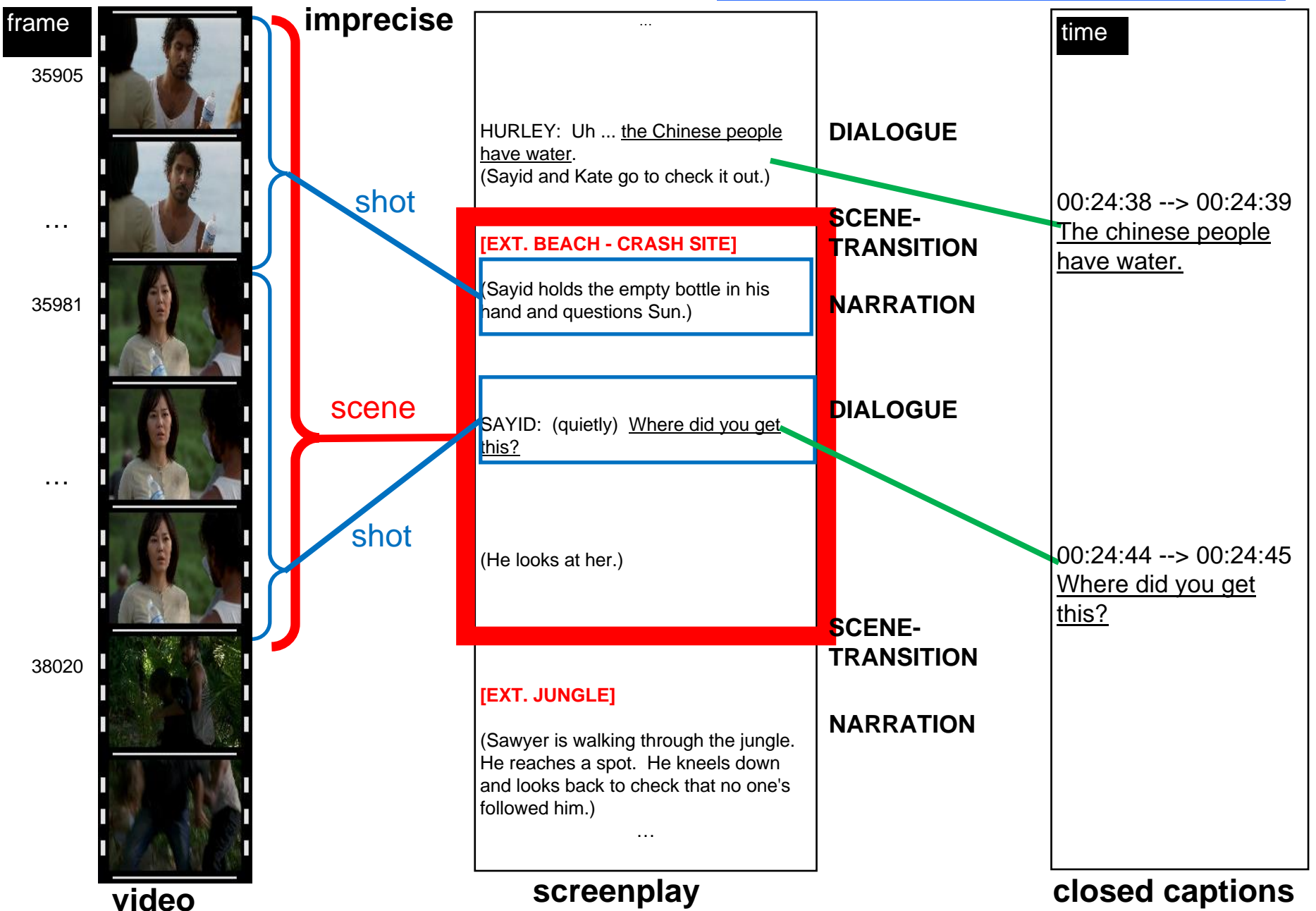
Query: "walks"

(Hurley) (walks up) (to ...)



image/text alignment

screenplay/closed captions alignment

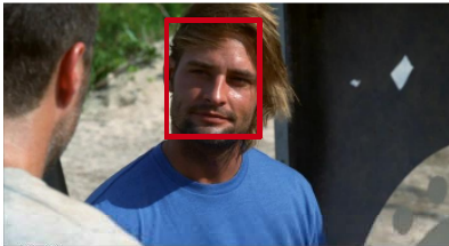


Learning from Ambiguous Labels

[T. Cour, B. Sapp, C. Jordan, B. Taskar, CVPR09]

- Each face has two or more possible labels

[INT. BEACH - SAWYER'S TENT -- DAY]



JACK: "Where is it?"

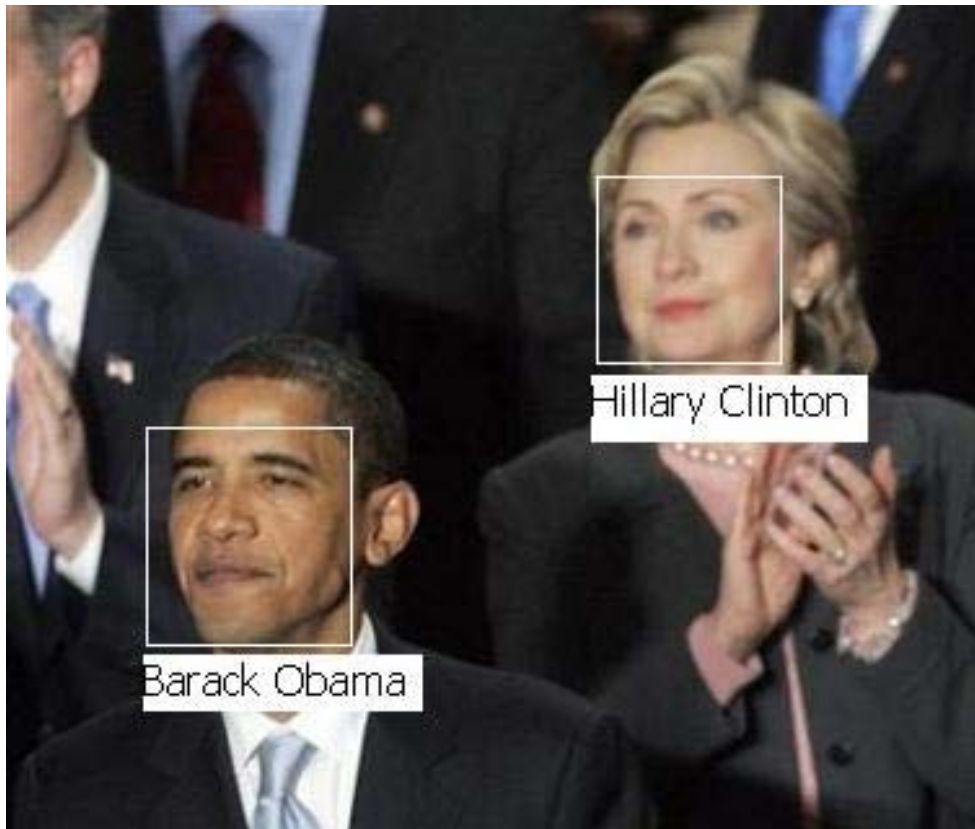
SAWYER: "Where's what?"



Jack? Jack?
Sawyer? Sawyer?

Faces in the News

- Image captions on the web



Hillary Clinton is 100% behind Barack Obama.

Ambiguous Labeling Setting

- x - input
- y - true label $\in \{1, \dots, K\}$
- z - extra label(s) $\in \{1, \dots, K\}$
- IID samples from unknown $P(x, y, z)$
- Ambiguous observations: $(x, \{y, z\})$



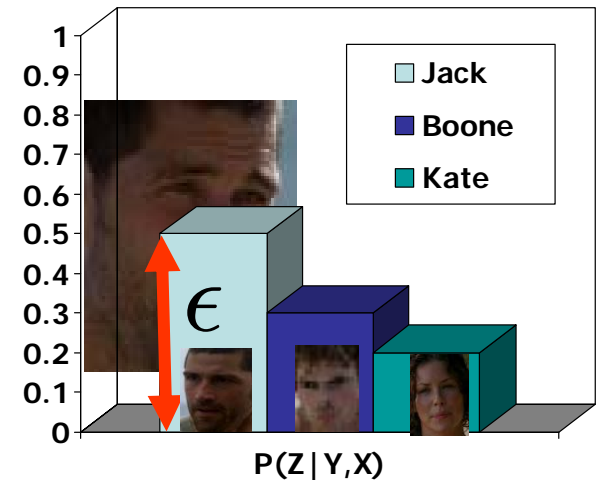
Can we learn without true labels?

- No analysis (ttbomk)

- **Confounders:**

$$P(z=Jack \mid y=Sawyer, x) = 1$$

Can't tell them apart



- **Assumption:** (ϵ, δ) -ambiguity

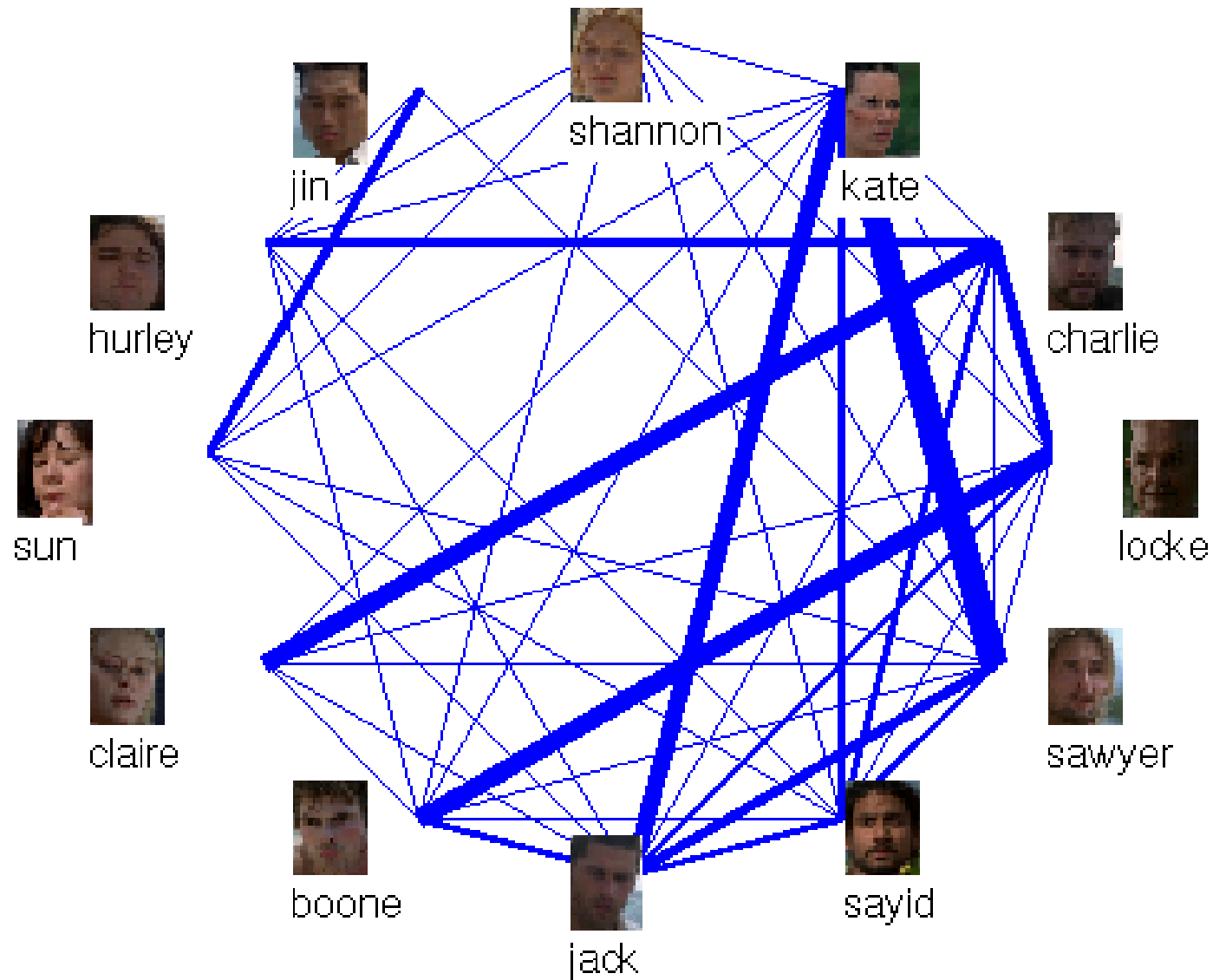
$P(z|y,x)$ is less than 1 (most of the time)

$$\epsilon = \sup_{(x,y) \in G, z} P(z \mid y, x) < 1$$

G – set of good pairs (x,y) where above holds

$$P((x, y) \in G) = 1 - \delta$$

Ambiguity network for LOST



Generalization from Ambiguous Samples

Error: $\mathbf{E}[y \neq f(x)]$ Ambiguous Error: $\mathbf{E}[y, z \neq f(x)]$

$$\mathbf{E}[y \neq f(x)] \geq \mathbf{E}[y, z \neq f(x)]$$

Theorem: (assuming (ϵ, δ) -ambiguity)

$$\mathbf{E}[y \neq f(x)] \leq \frac{1}{1 - \epsilon} \mathbf{E}[y, z \neq f(x)] + \delta$$

Theorem: with probability $1 - \eta$

$$\mathbf{E}[y \neq f(x)] \leq \frac{1}{1 - \epsilon} \left(\hat{\mathbf{E}}_n[\text{margin}(f)] + O\left(\sqrt{\frac{\ln(1/\eta)}{n}}\right) \right) + \delta$$

Convex Discriminative Formulation

- Multiclass Model: $f(x) = \arg \max_k f^k(x)$; $f^k(x) = w^k \cdot x$
- One-Against-All Loss:

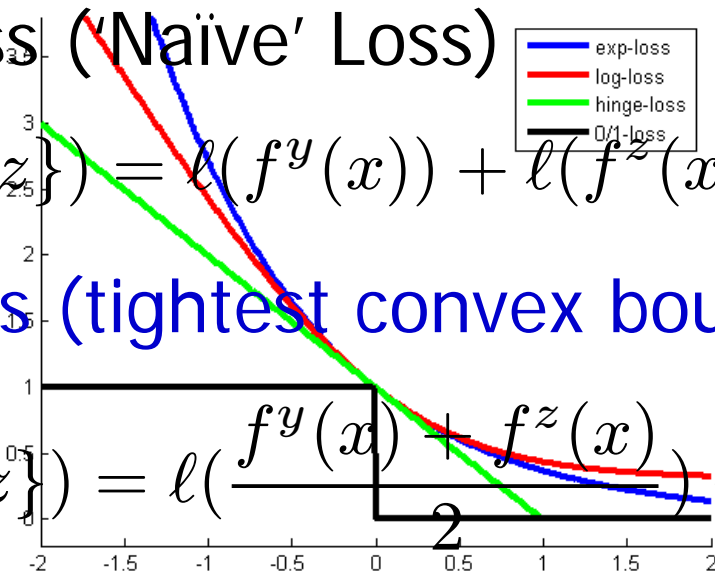
$$\mathcal{L}(f(x), y) = \ell(f^y(x)) + \sum_{k \neq y} \ell(-f^k(x))$$

- Multilabel Loss ('Naive' Loss)

$$\mathcal{L}(f(x), \{y, z\}) = \ell(f^y(x)) + \ell(f^z(x)) + \sum_{k \neq y, z} \ell(-f^k(x))$$

- Proposed Loss (tightest convex bound on ambig err):

$$\mathcal{L}(f(x), \{y, z\}) = \ell\left(\frac{f^y(x) + f^z(x)}{2}\right) + \sum_{k \neq y, z} \ell(-f^k(x))$$

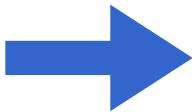


where $\ell(\cdot)$ is standard binary loss (e.g. hinge, exp, log)

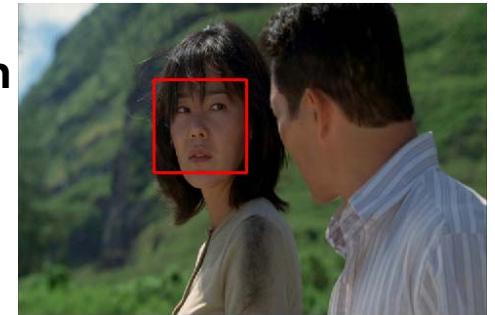
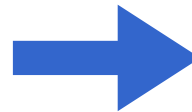
DVD to Faces



decompilation



frontal
face
detection



part
detection

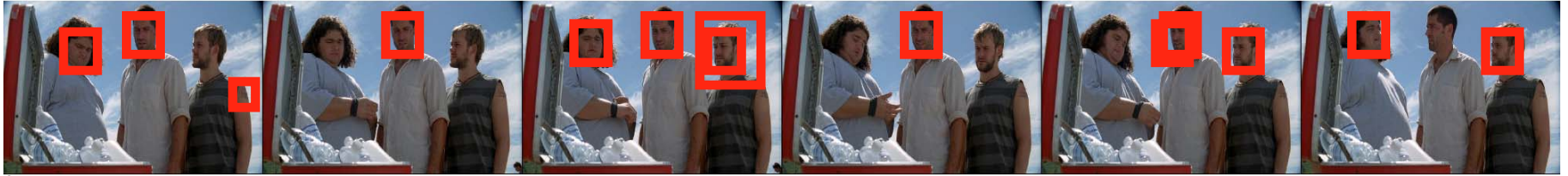


rigid
registration



60 x 90

Tracking



false positives
false negatives
overlapping detections
no grouping



dynamic program



register

track# 1



track# 2



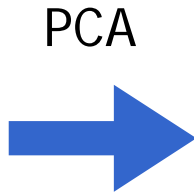
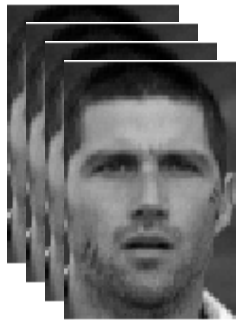
track# 3



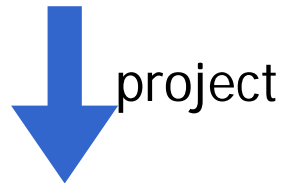
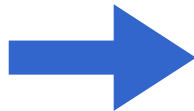
Face Features

eigenfaces

mean

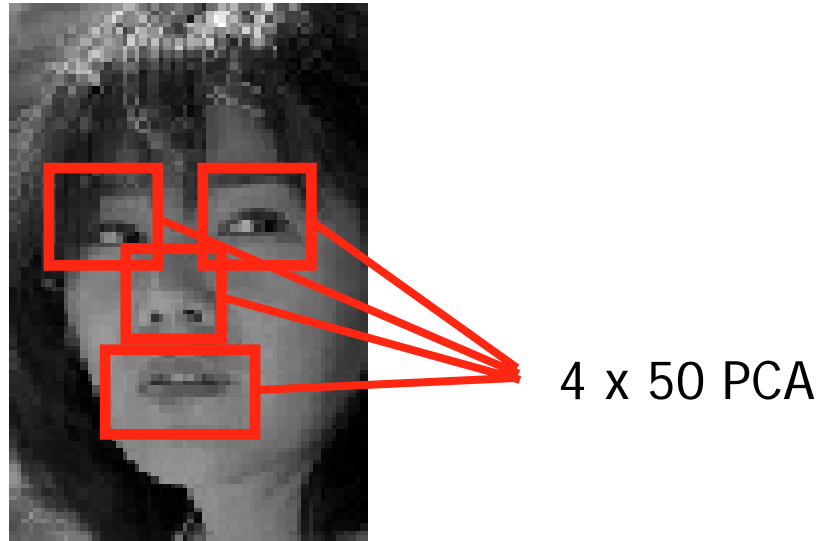


new face

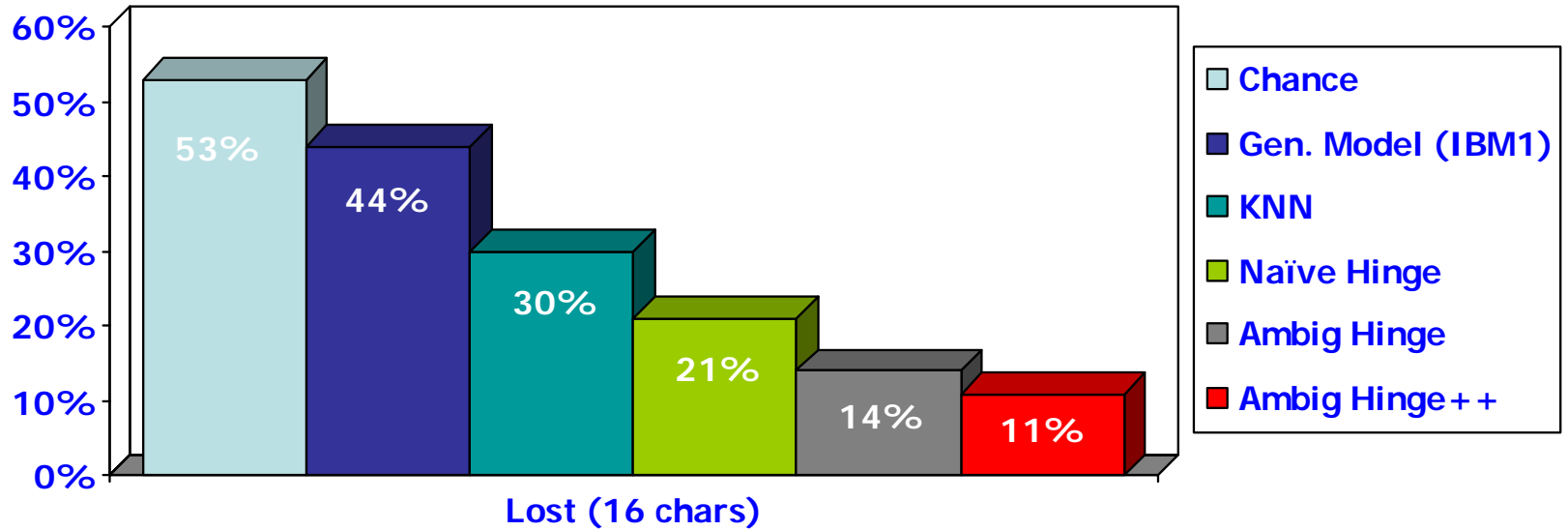


50 PCA

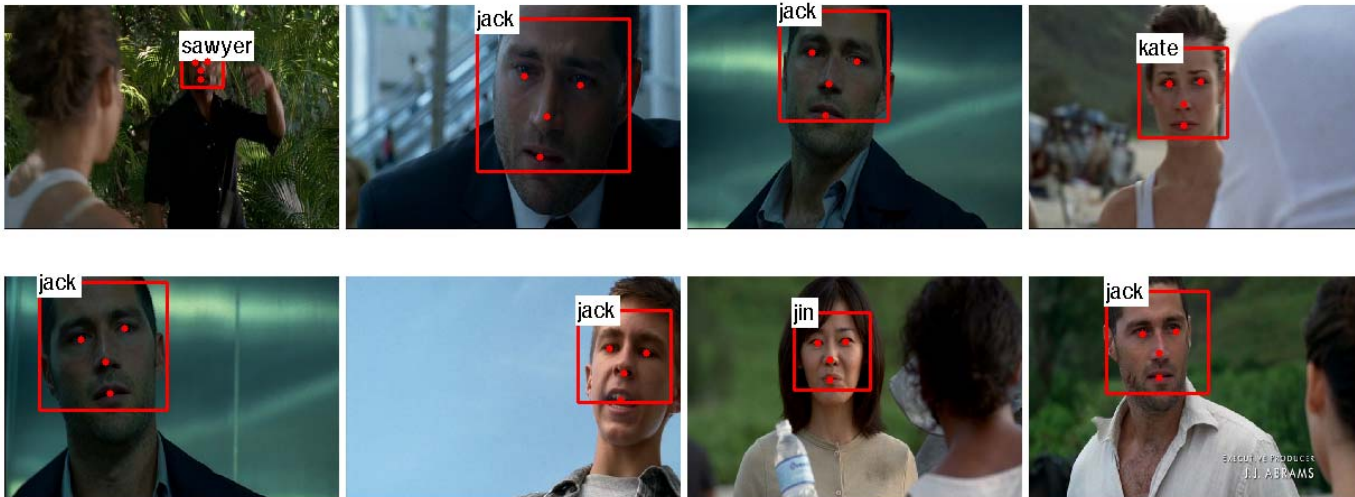
Additional Features



Naming Error

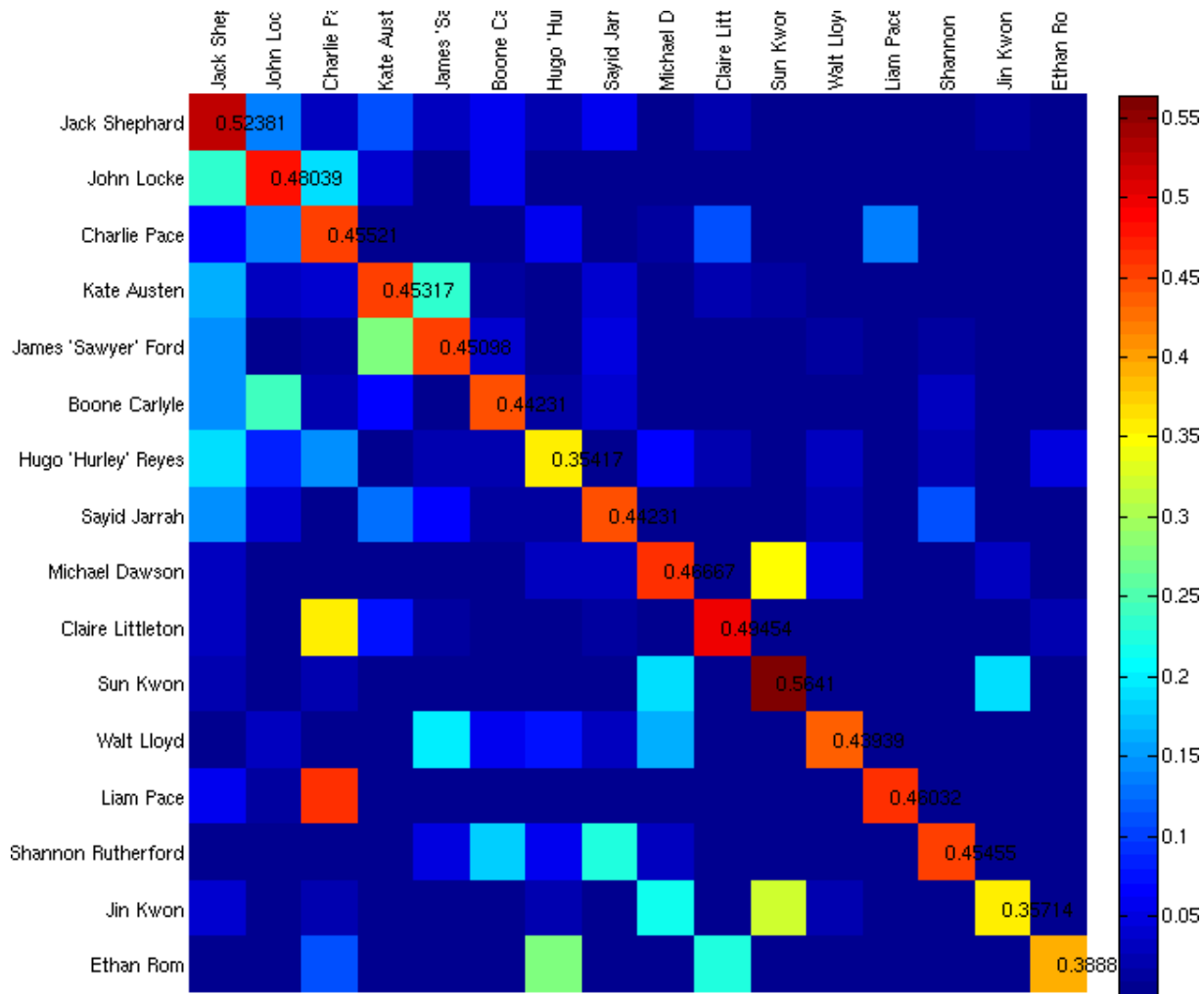


8 episodes

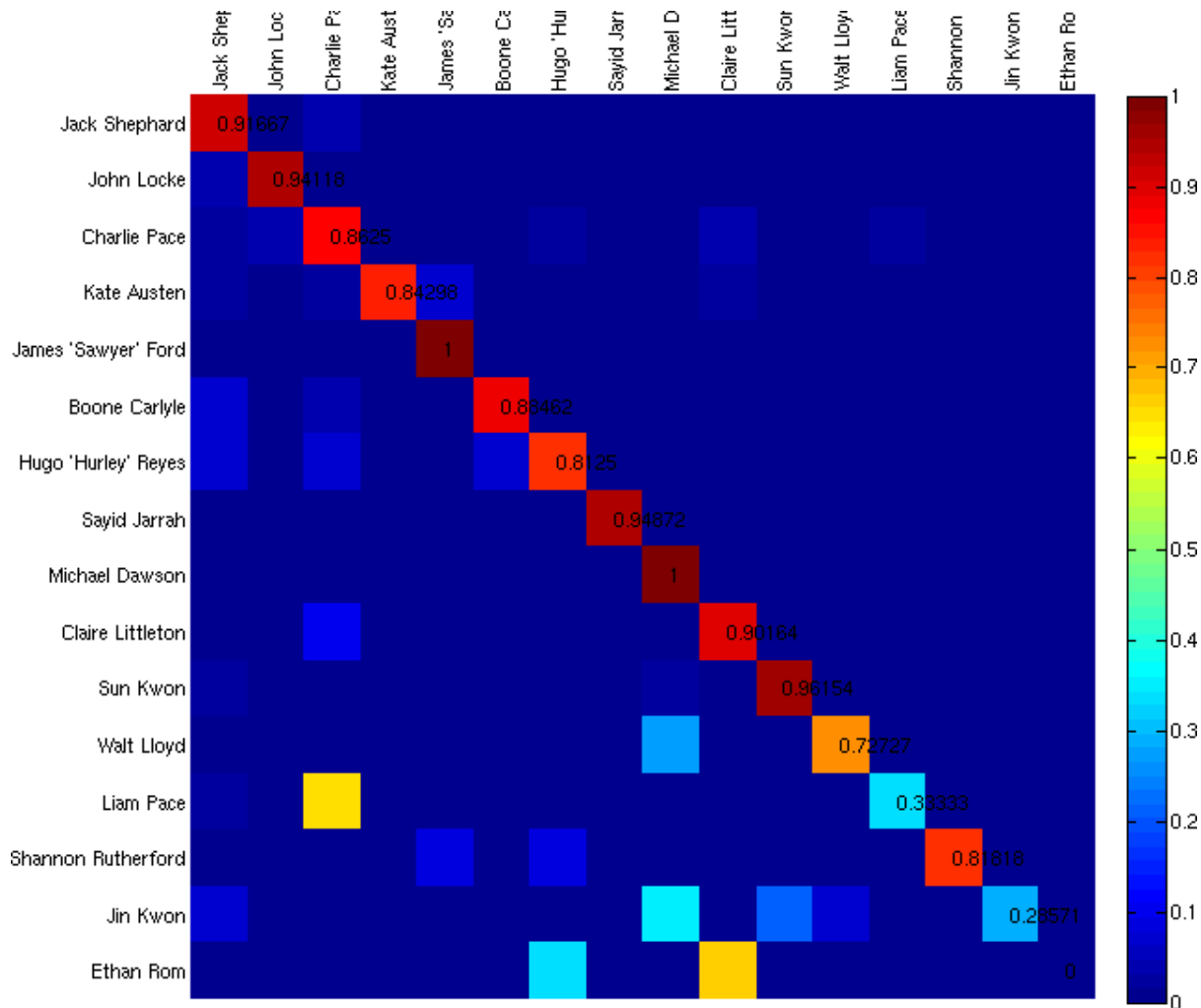


Scenes from *Lost*





Confusion matrix of chance (row normalized)
 (i,j): proportion of times i was seen with j



Confusion matrix of learned model (row normalized)
 (i,j): proportion of times i was seen with j

CSI

Catharine Willows

Precision: 85.3



CSI

Sarah Sidle

Precision: 78.3



Labeled Actions from Videos

[EXT. BEACH - CRASH SITE - DAY]

(Sayid holds the empty bottle in his hand and questions her.)

SAYID. (quietly) where did you get this?
(He looks at her.)

Screenplay

(Sayid holds the empty bottle in **his** hand and questions **her**.)

pronoun resolution

(Sayid holds the empty bottle in **Sayid's** hand and questions **Sun**.)

verb frames (subject verb object)

(Sayid holds bottle) (Sayid questions Sun)

alignment

???

Video



identify:

PEOPLE
LOCATIONS
OBJECTS
ACTIONS



Action Dictionary

σηουτ

(JACK) (shouts) ()



ωακε

(Sawyer) (wakes up) ()



Precision: 90%

φολλω

(Kate) (follows) (Jack)



σιτ

(Locke) (sits down) ()



σμιλε

(Kate) (smiles) ()



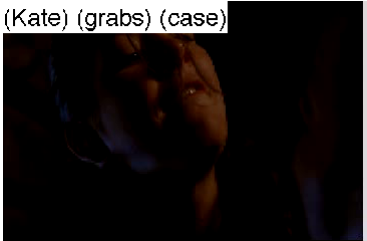
σωιμ

(Sawyer) (turns) (swimming)



γραβ

(Kate) (grabs) (case)



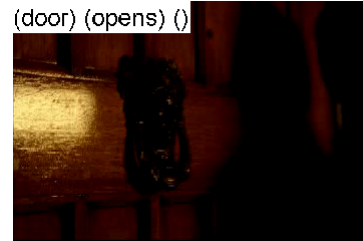
κισσ

(Shannon) (kisses) (ear)



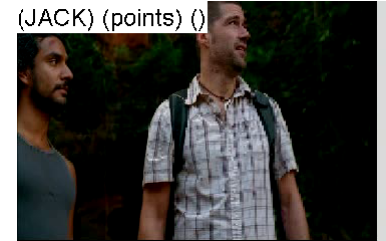
οπεν δοορ

(door) (opens) ()

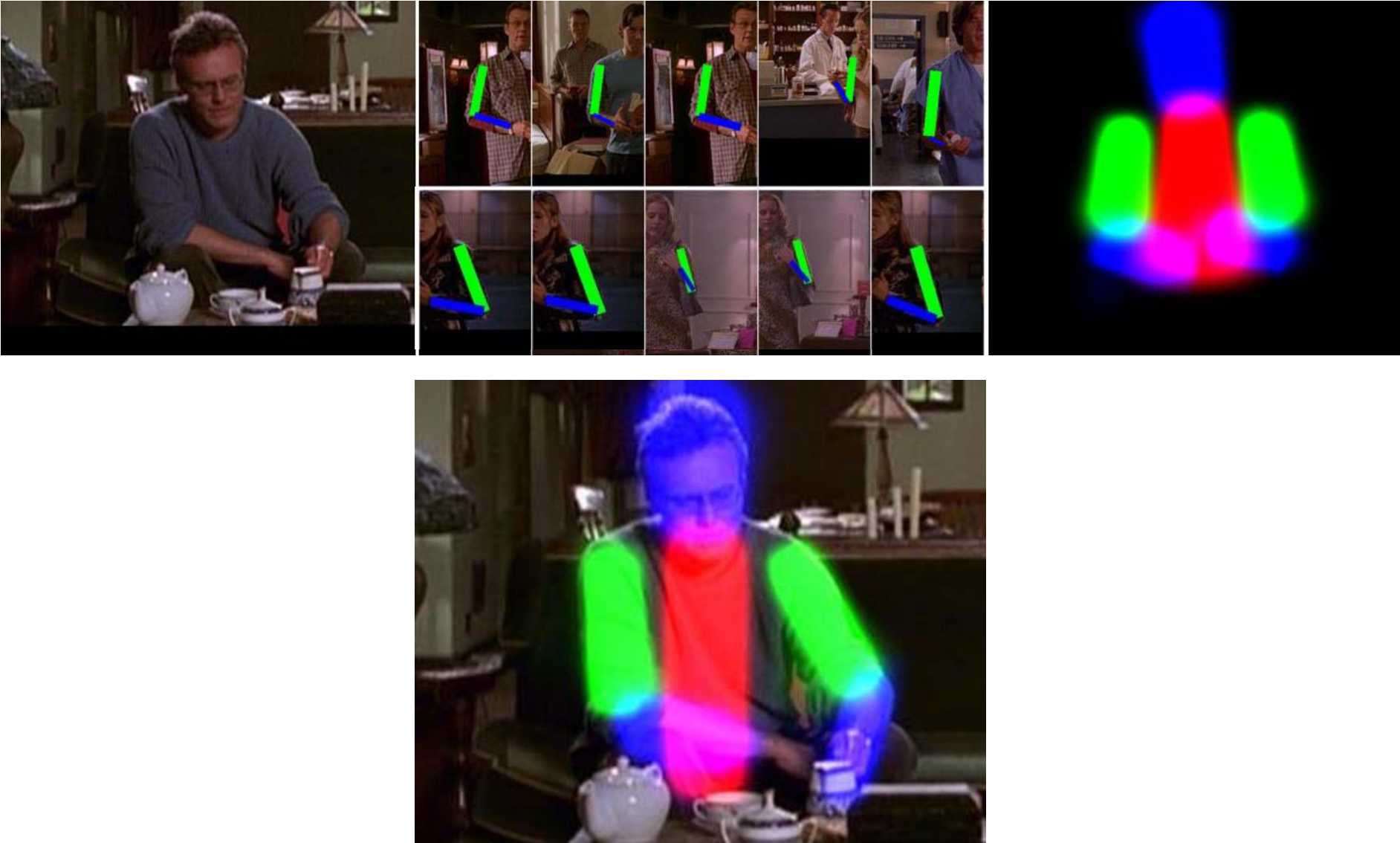


ποιντ

(JACK) (points) ()



Body Parsing via Locally Parametric CRFs



Naming without a screenplay

1st person reference



I'm Jack.

Jack in scene
speaking

2nd person reference



Hey, Jack!

Jack in scene
not speaking

3rd person reference



Where is Jack?

Jack *not* in scene
not speaking

false positive



Jack-in-the-box

Supervision from dialogue: { sparse
indirect
noisy

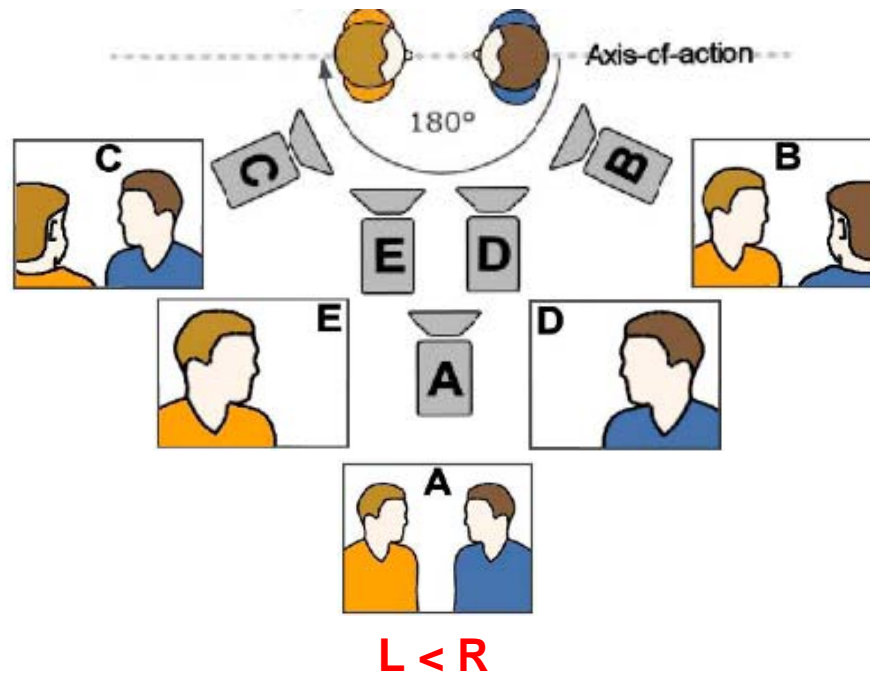
~50 references per episode only

only *constraints* on possible labels

“It’s Jack”: 1st or 3rd person ?

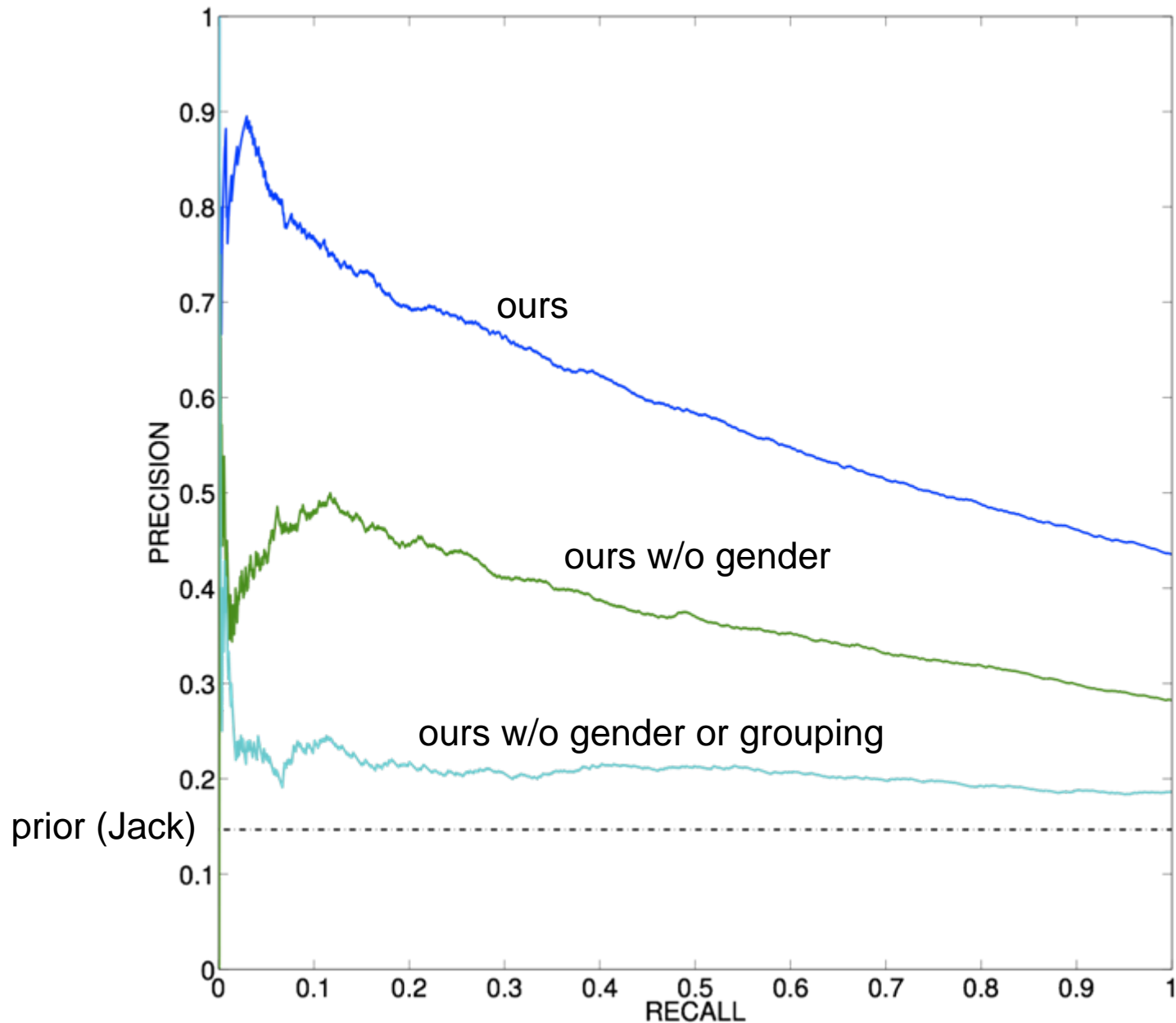
Grouping using continuity editing cues

Grouping using Gestalt of Continuity Editing 180°-rule



An Attentional Theory of Continuity Editing [Smith, 2005]

Dialog-Only Naming



Understanding Movies

[With: T. Cour, B. Sapp, C. Jordan, E. Miltsakaki]

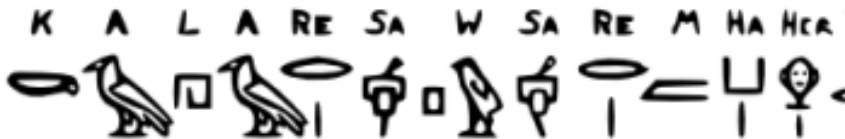
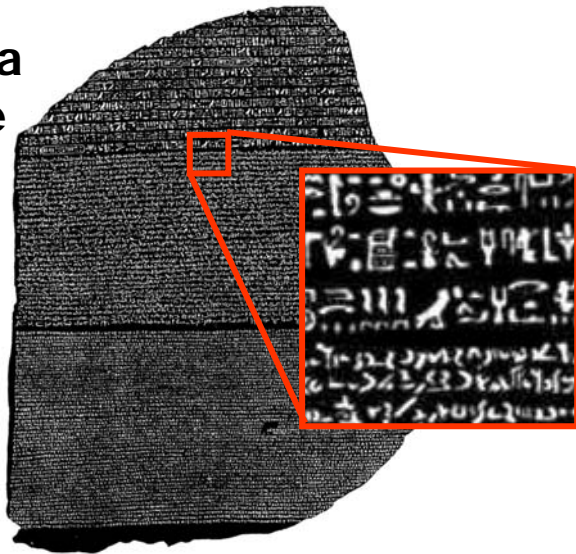
- Who, what, where, when?
 - With minimal supervision
 - Novel weak learning models and analysis
- Naming people without a screenplay
 - Dialog only: self-introductions and addresses
 - Using voices and faces, editing cues to group
- Learning articulated action models
 - Human figure parsing in videos

Learning from Co-Occurrence

Foreign Language Lexicon

Visual Lexicon

Rosetta Stone



HURLEY: Uh ... the Chinese people have water.
(Sayid and Kate go to check it out.)

[EXT. BEACH - CRASH SITE]

(Sayid holds the empty bottle in his hand and questions Sun.)

SAYID: (quietly)
Where did you get this?
(He looks at her.)

[EXT. JUNGLE]

(Sawyer is walking through the jungle. He reaches a spot. He kneels down and looks back to check that no one's followed him.)



SAYID



SUN



BOTTLE

Word Alignment

- Key step in statistical machine translation systems

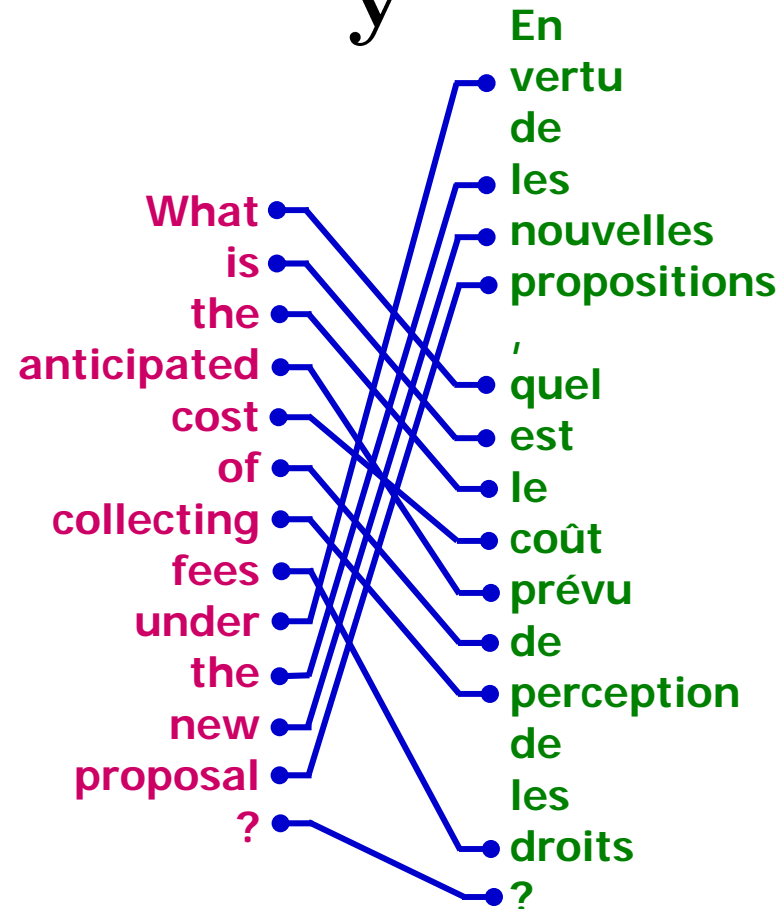
x

What is the anticipated
cost of collecting fees
under the new proposal?

En vertu des nouvelles
propositions, quel est le
coût prévu de perception
des droits?



y



Supervised Word Alignment

200 train, En/Fr

AER

Prec / Rec

	<u>AER</u>	<u>Prec</u> / <u>Rec</u>
IBM model 4 (intersected)	6.5	98 / 88%
Our Alignment Model	4.3	96 / 95%

Best published accuracy on English-French (Hansards)

Unsupervised Alignment

- Spanish, German, Finnish, Czech
- No supervised data
- Need to learn from co-occurrence only
- IBM Translation Models: 1-4
[Brown, Della Pietra, Della Pietra and Mercer, 94]

HMM model [Ney, Vogel '96]

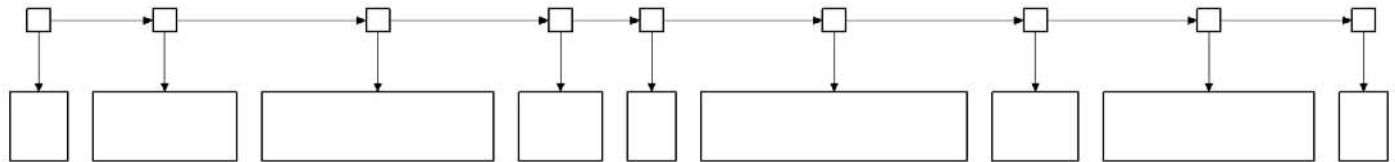
Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$

HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$

$p(\mathbf{e})$

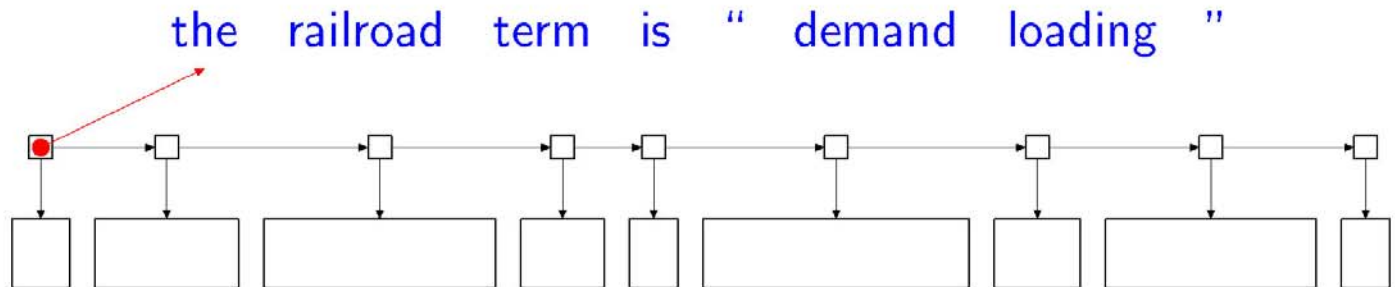
the railroad term is " demand loading "



HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$

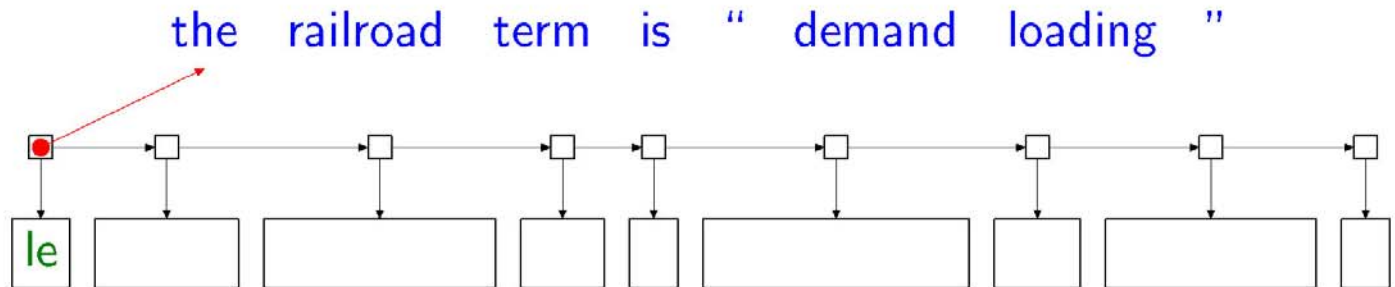
$p(\mathbf{e})$



HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$

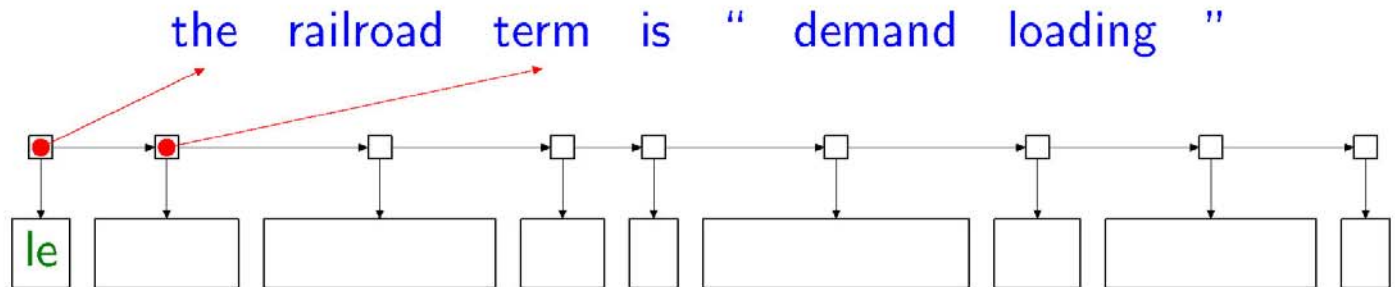
$p(\mathbf{e})$



HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$

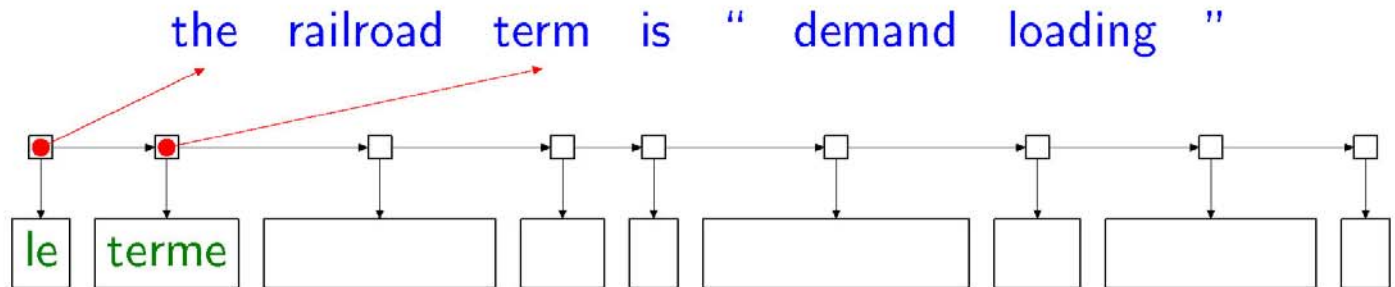
$p(\mathbf{e})$



HMM model [Ney, Vogel '96]

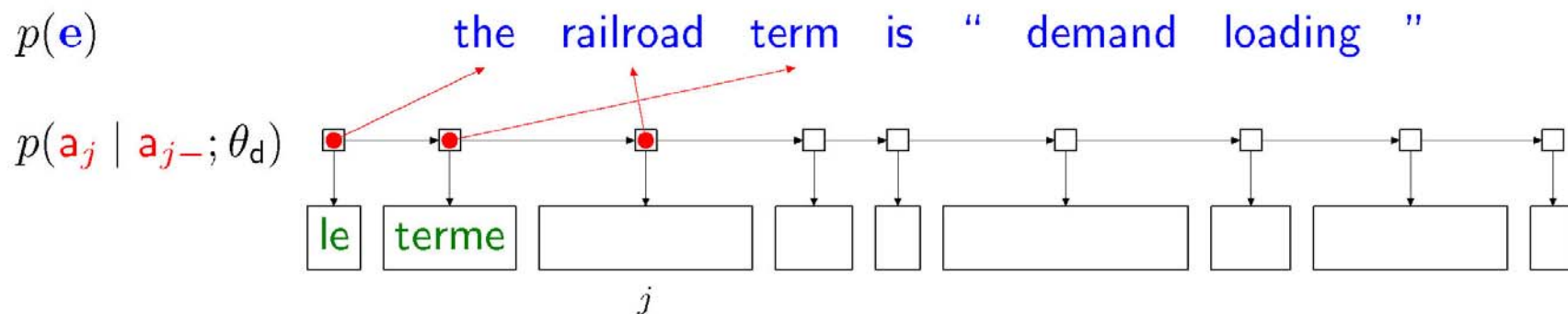
Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$

$p(\mathbf{e})$



HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$



Distortion θ_d

$$p(\begin{matrix} \uparrow & \uparrow \\ \bullet & \bullet \end{matrix}) = 0.6$$

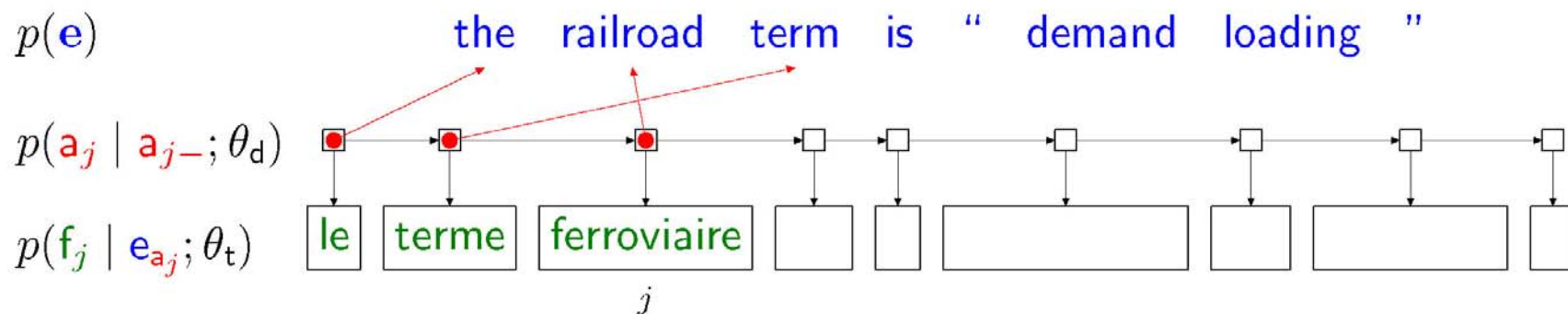
$$p(\begin{matrix} \uparrow & \nearrow \\ \bullet & \bullet \end{matrix}) = 0.2$$

$$p(\begin{matrix} \nearrow & \searrow \\ \bullet & \bullet \end{matrix}) = \mathbf{0.1}$$

...

HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$



Distortion θ_d

$$\begin{aligned}
 p(\uparrow \uparrow) &= 0.6 \\
 p(\uparrow \nearrow) &= 0.2 \\
 p(\nearrow \uparrow) &= \mathbf{0.1} \\
 &\dots
 \end{aligned}$$

Translation θ_t

$$\begin{aligned}
 p(\text{the} \rightarrow \text{le}) &= 0.53 \\
 p(\text{the} \rightarrow \text{la}) &= 0.24 \\
 p(\text{railroad} \rightarrow \text{ferroviaire}) &= \mathbf{0.19} \\
 p(\text{NULL} \rightarrow \text{le}) &= 0.12 \\
 &\dots
 \end{aligned}$$

Note: model not symmetric

EM training

Maximize $p(\mathbf{e}, \mathbf{f}; \theta)$

Parameters: θ

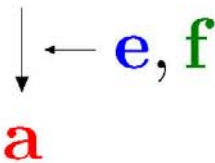
E-step:

$q(\mathbf{a} | \mathbf{e}, \mathbf{f}) := p(\mathbf{a} | \mathbf{e}, \mathbf{f}; \theta)$
(forward-backward)

q



θ



EM training

Maximize $p(\mathbf{e}, \mathbf{f}; \theta)$

Parameters: θ

Expectation over alignments: q

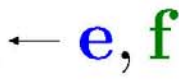
E-step:
 $q(\mathbf{a} | \mathbf{e}, \mathbf{f}) := p(\mathbf{a} | \mathbf{e}, \mathbf{f}; \theta)$
(forward-backward)

M-step:
 $\theta := \operatorname{argmax}_{\theta} \mathbb{E}_q \log p(\mathbf{a}, \mathbf{e}, \mathbf{f} | \theta)$
(normalizing counts)

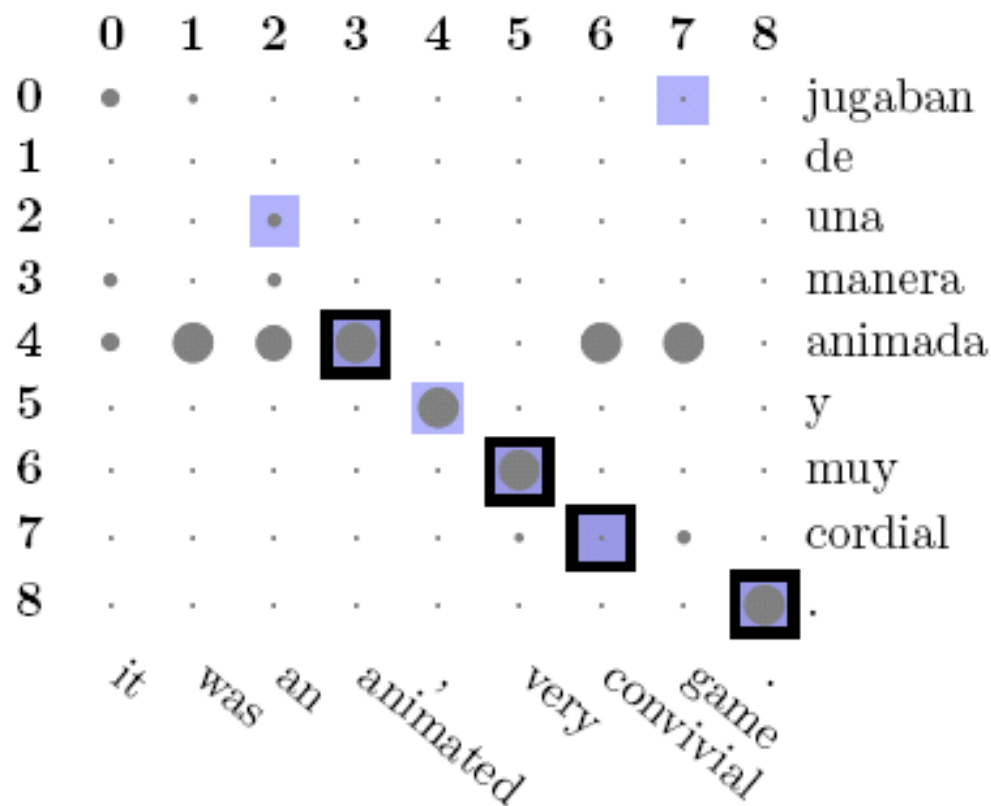
q

θ

\mathbf{a}

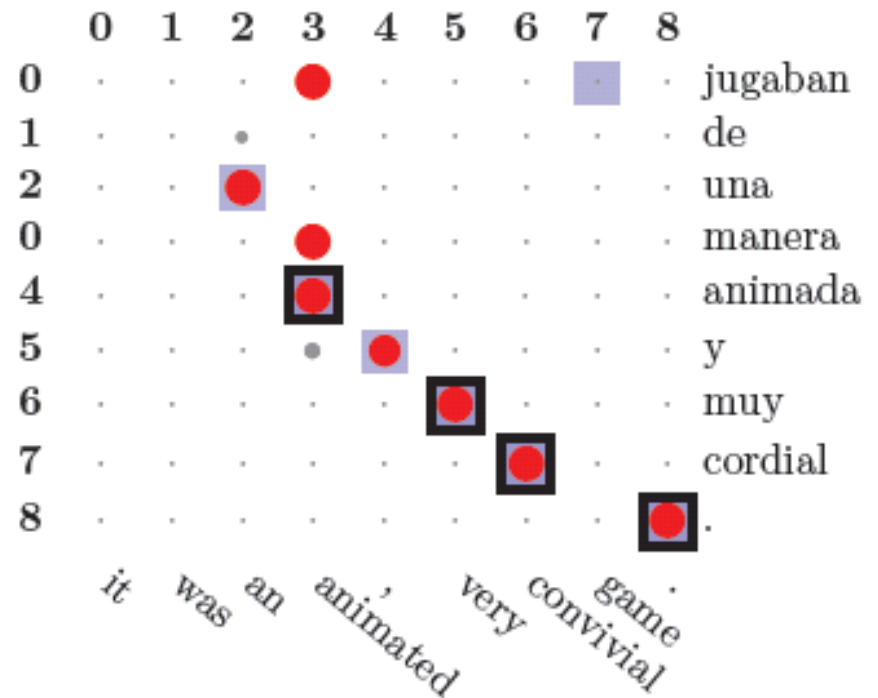
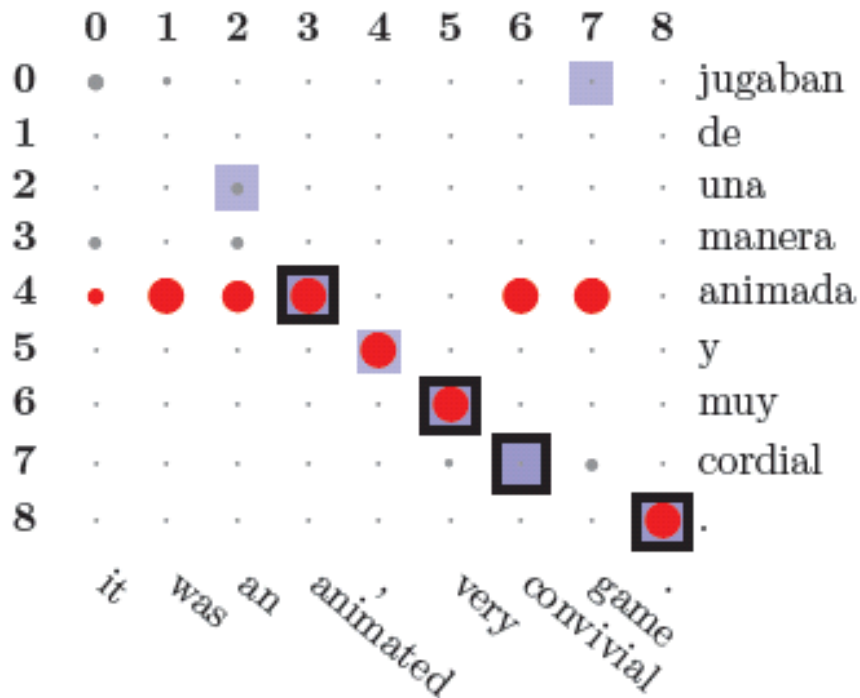


Posterior Over Alignments



Matrix dimensions	
rows	source words
columns	target words
Word to word posterior probabilities p	
•	$0.01 < p$
•	$0.01 \leq p < 0.05$
•	$0.05 \leq p < 0.1$
•	$0.1 \leq p < 0.2$
•	$0.2 \leq p < 0.3$
•	$0.3 \leq p < 0.4$
•	$0.4 \leq p < 0.5$
•	$0.5 \leq p < 0.6$
•	$0.6 \leq p < 0.7$
•	$0.7 \leq p < 0.8$
•	$0.8 \leq p < 0.9$
•	$0.9 \leq p < 0.95$
•	$0.95 \leq p$
Alignment points	
■	Sure gold alignment point
■	Possible gold alignment point

Problems



- Not 1-1 (For En/Sp/Fr/Pt 86-98% are 1-1)
- Not symmetric
- Rare words collect garbage [Moore 05]

Fixes?

- More complex models (IBM 4,5,6, etc.)
 - Improper distributions
 - Computing posteriors over bijective alignments is **#P-complete (permanent problem)**
 - Decoding with symmetric pairwise costs is **NP-hard (quadratic assignment problem)**
- Post-processing heuristics [Och&Ney 03]
 - Intersection of directional models plus fill-in
 - Procedural, difficult to control

Controlling Latent Variables

- Common problem in generative models:
 - **What do latent variables represent?**
 - Control via additional features
 - Very indirect and unpredictable outcome
 - Control via additional model structure
 - Often makes model intractable or inefficient
- For latent alignment variables, want:
 - Bijectivity
 - Symmetry
- Idea: **impose control on directly on posteriors**

Standard EM

- Observed: \mathbf{x} Hidden: \mathbf{y} Model: $p_{\theta}(\mathbf{x}, \mathbf{y})$

- Objective: $L(\theta) = \hat{\mathbb{E}}_{\mathbf{x}} \log \sum_{\mathbf{y}} p_{\theta}(\mathbf{x}, \mathbf{y}) = \hat{\mathbb{E}}_{\mathbf{x}} \log p_{\theta}(\mathbf{x})$

- E-step:

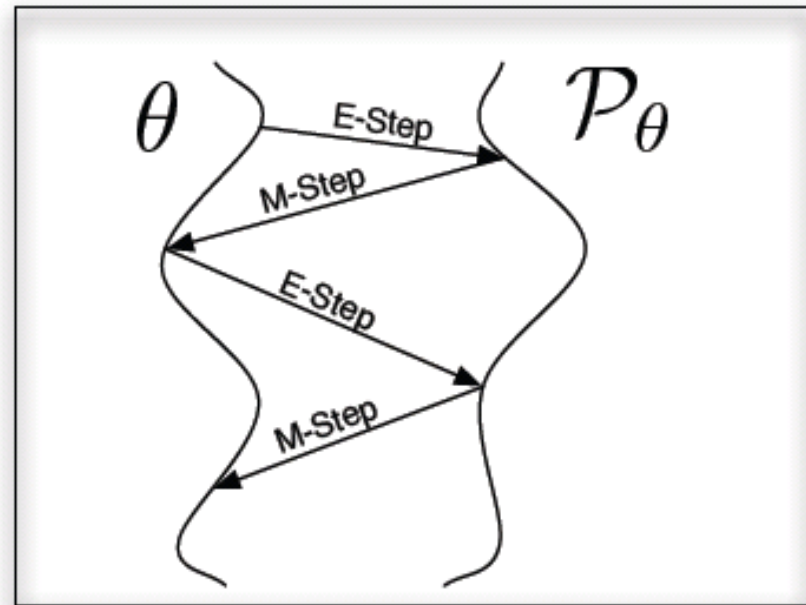
$$\begin{aligned} q_{\mathbf{x}}(\mathbf{y}) &= \arg \min_q KL(q(\mathbf{y}) || p_{\theta}(\mathbf{y} | \mathbf{x})) \\ &= p_{\theta}(\mathbf{y} | \mathbf{x}) \end{aligned}$$

- M-step:

$$\max_{\theta} \hat{\mathbb{E}}_{\mathbf{x}} \mathbb{E}_q \log p_{\theta}(\mathbf{x}, \mathbf{y})$$

- Lower Bound: $F(\theta, q) = \hat{\mathbb{E}}_{\mathbf{x}} \mathbb{E}_q \log \frac{p_{\theta}(\mathbf{x}, \mathbf{y})}{q_{\mathbf{x}}(\mathbf{y})} \leq L(\theta)$

- A local max of $F(\theta, q)$ is a local max of $L(\theta)$



???

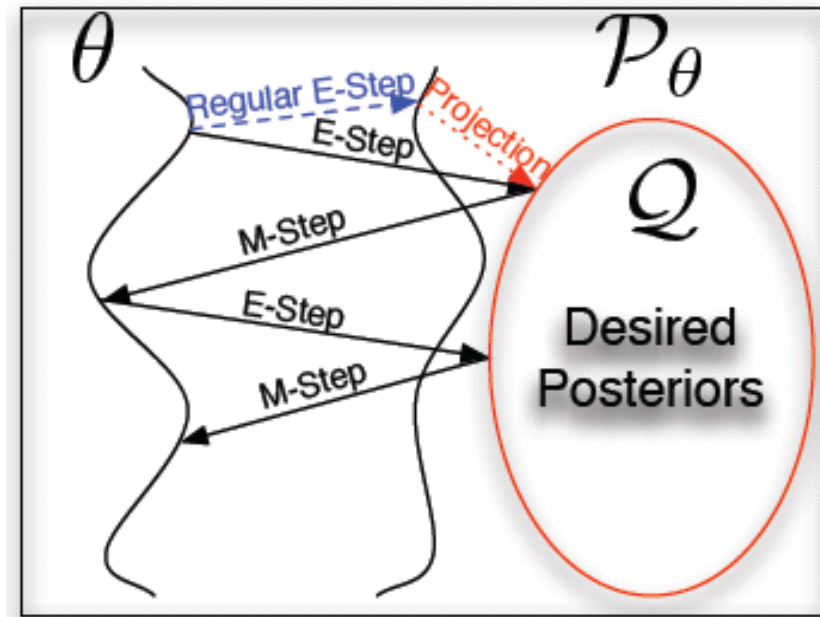
- E-step: (different)

$$q_{\mathbf{x}}(\mathbf{y}) = \arg \min_{q \in \mathcal{Q}_{\mathbf{x}}} KL(q(\mathbf{y}) || p_{\theta}(\mathbf{y} | \mathbf{x}))$$

constraints

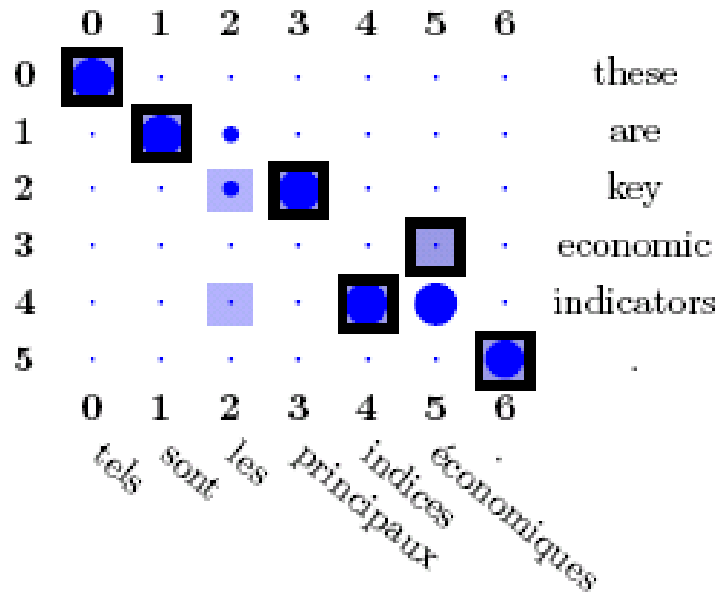
- M-step: (same)

$$\max_{\theta} \hat{\mathbb{E}}_{\mathbf{x}} \mathbb{E}_q \log p_{\theta}(\mathbf{x}, \mathbf{y})$$



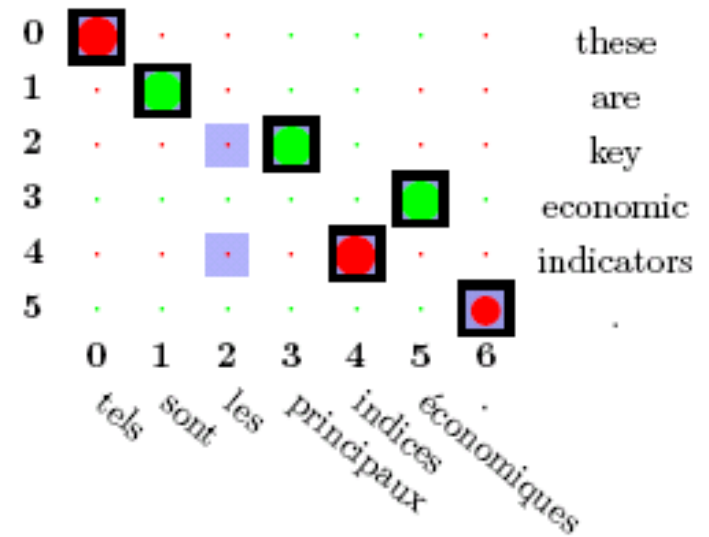
Bijection Constraints

\mathcal{Q}



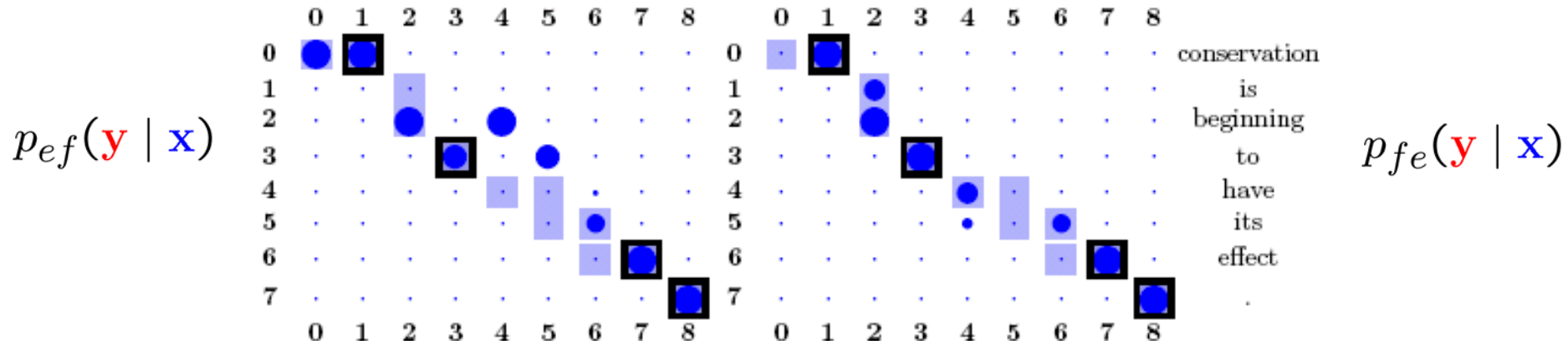
$$p_{\theta}(\mathbf{y} \mid \mathbf{x})$$

$$\begin{cases} \sum \leq 1 \\ \sum \leq 1 \\ \sum \leq 1 \\ \sum \leq 1 \\ \sum \leq 1 \\ \sum \leq 1 \\ \sum \leq 1 \end{cases}$$

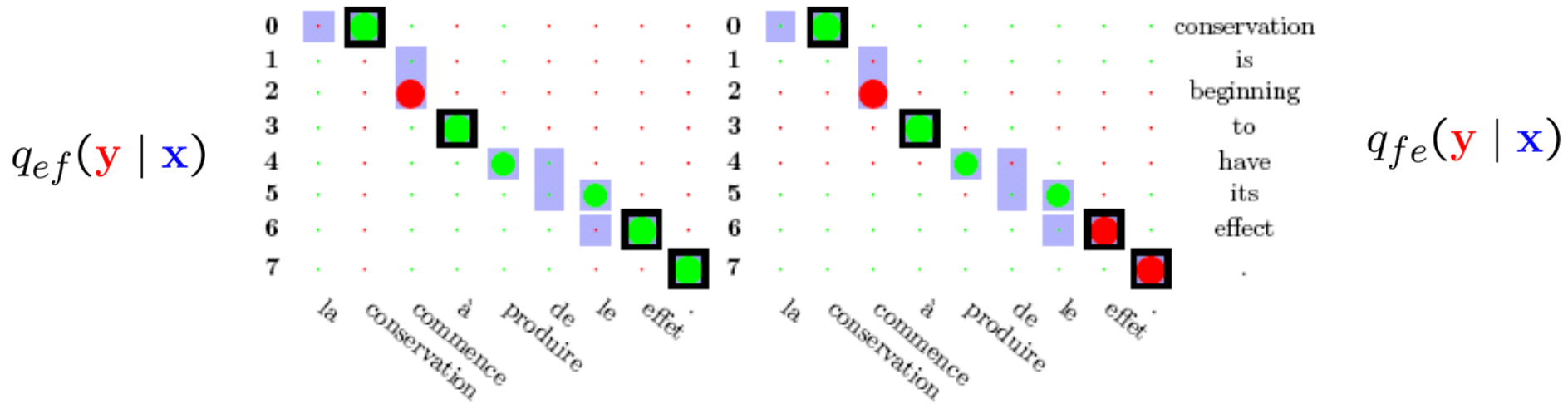


$$q_{\mathbf{x}}(\mathbf{y}) = \arg \min_{q \in \mathcal{Q}} KL(q(\mathbf{y}) \parallel p_{\theta}(\mathbf{y} \mid \mathbf{x}))$$

Symmetry (Agreement) Constraints



$$\mathcal{Q} = \{q : q_{ef}(y_{ij}) = q_{fe}(y_{ji}), \quad \forall i, j\}$$



Posterior Regularization Objective

[Graca, Ganchev, Taskar, NIPS 07]

- E-step:

$$q_{\mathbf{x}}(\mathbf{y}) = \arg \min_{q \in \mathcal{Q}_{\mathbf{x}}} KL(q(\mathbf{y}) || p_{\theta}(\mathbf{y} | \mathbf{x}))$$

constraints

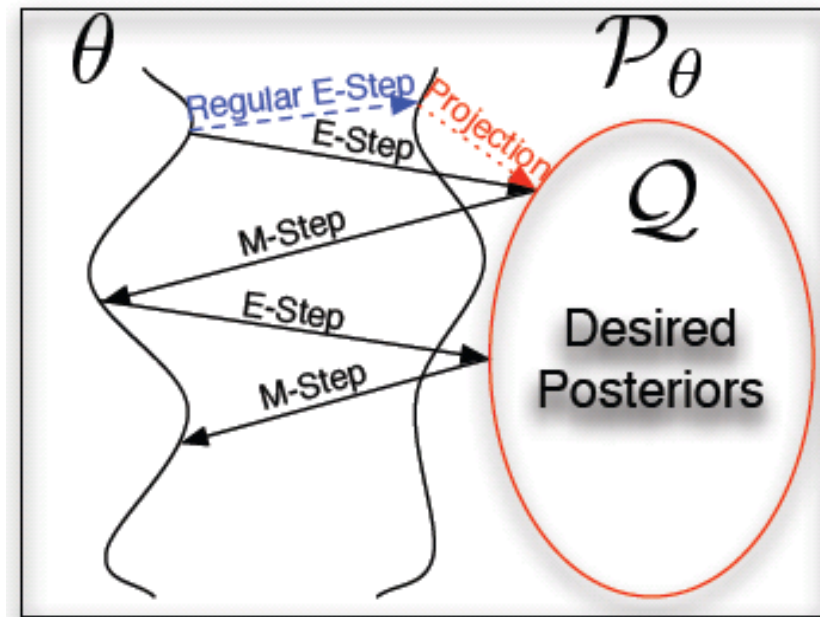
- M-step:

$$\max_{\theta} \hat{\mathbb{E}}_{\mathbf{x}} \mathbb{E}_q \log p_{\theta}(\mathbf{x}, \mathbf{y})$$

- Theorem: converges to a local max of

$$L(\theta) - \hat{\mathbb{E}}_{\mathbf{x}} KL(\mathcal{Q}_{\mathbf{x}} || p_{\theta}(\mathbf{y} | \mathbf{x}))$$

$$KL(\mathcal{Q}_{\mathbf{x}} || p) = \min_{q \in \mathcal{Q}_{\mathbf{x}}} KL(q || p) = \text{penalty for deviation from constraints}$$



E-step: I-projections

- If constraints are **linear** inequalities (or eqs)

$$\mathcal{Q}_{\mathbf{x}} = \{q_{\mathbf{x}}(\mathbf{y}) : \mathbb{E}_q \mathbf{f}(\mathbf{x}, \mathbf{y}) \leq \mathbf{b}\}$$

- Then $q_{\mathbf{x}}(\mathbf{y}) = \arg \min_{q \in \mathcal{Q}_{\mathbf{x}}} KL(q(\mathbf{y}) || p_{\theta}(\mathbf{y} | \mathbf{x}))$
 $\propto p_{\theta}(\mathbf{y} | \mathbf{x}) \exp(-\lambda \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}))$

- If $\mathbf{f}(\mathbf{x}, \mathbf{y})$ is a sum over factors of $p_{\theta}(\mathbf{x}, \mathbf{y})$
then $q_{\mathbf{x}}(\mathbf{y})$ has same graphical structure, complexity

- Projection solved by gradient descent on the dual

E-step: I-projection Dual

- Dual for each example \mathbf{x} is:

$$\min_{\lambda \geq 0} \lambda \cdot \mathbf{b} + \log \sum_{\mathbf{y}} p_{\theta}(\mathbf{y} \mid \mathbf{x}) \exp(-\lambda \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}))$$

- Gradient:

$$\mathbf{b} - \mathbb{E}_q \mathbf{f}(\mathbf{x}, \mathbf{y})$$

where

$$q_{\mathbf{x}}(\mathbf{y}) \propto p_{\theta}(\mathbf{y} \mid \mathbf{x}) \exp(-\lambda \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}))$$

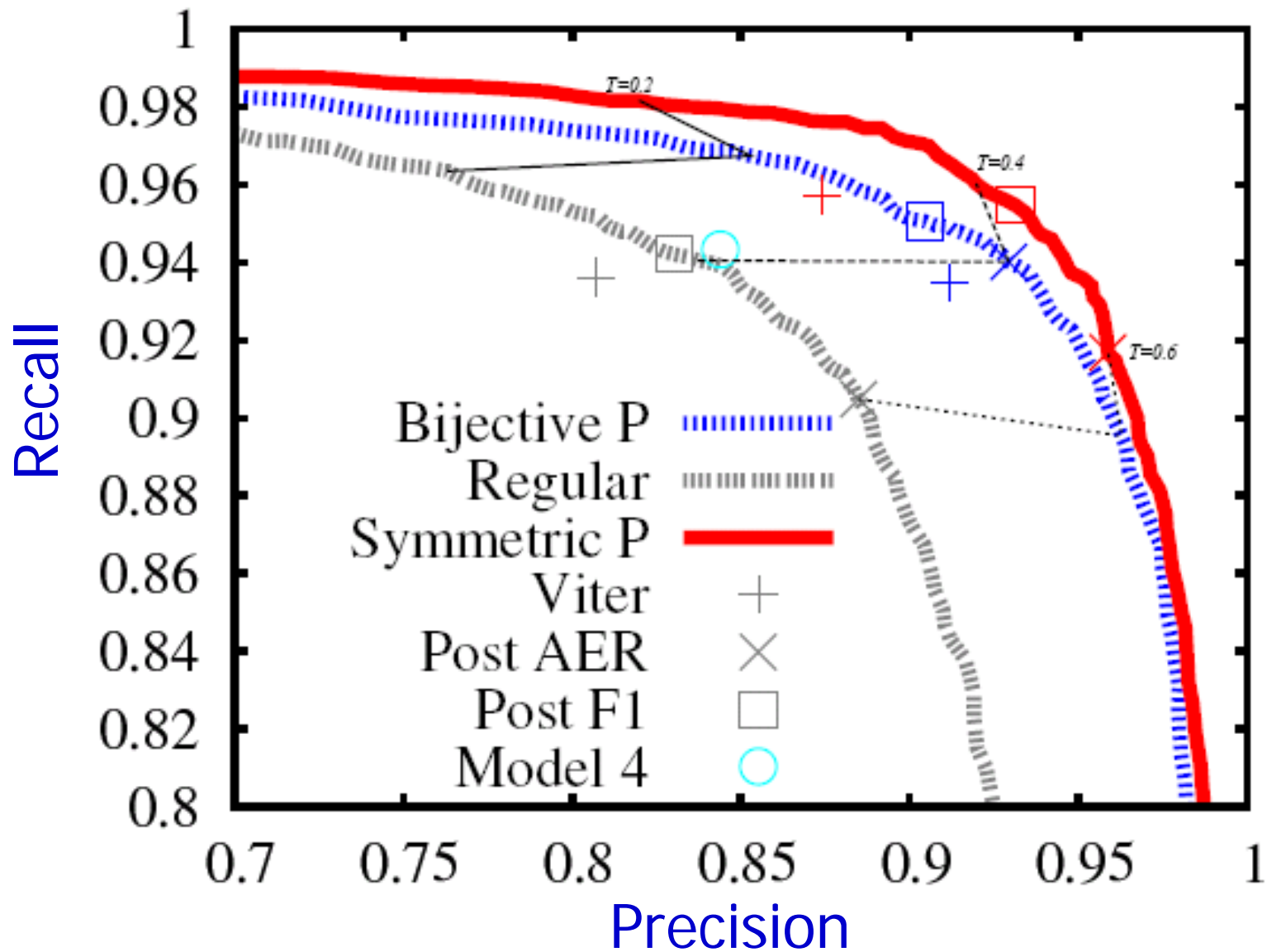
- Since projections are per example, online EM is easy

Corpora

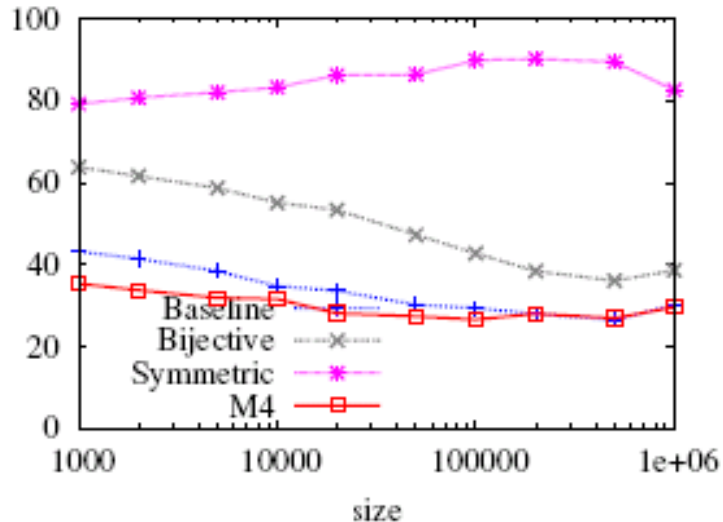
- Hansards, Europarl
- Standard dev set for tuning, test set
- En/Fr, En/Pt, En/Sp

- Metric:
 - Precision/Recall tradeoff is application-driven
 - Generate curves using posterior threshold
- Application-specific metrics:
 - Bleu for MT
 - Accuracy of bitext dependency projection

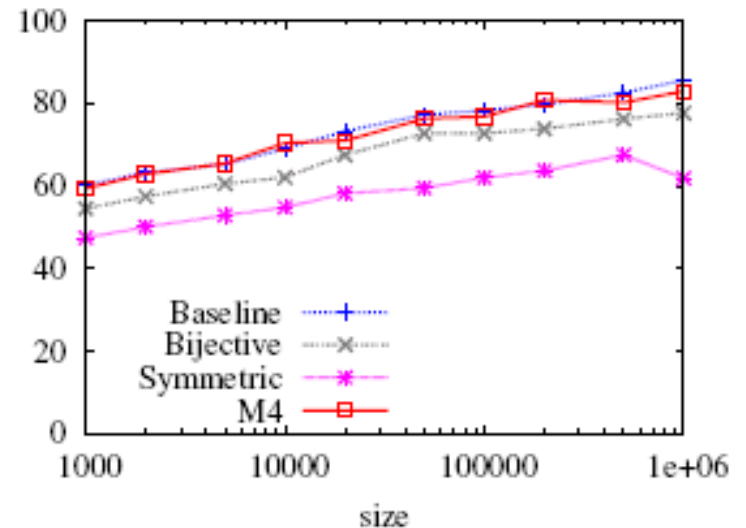
Hansards (En/Fr, 1m sent.)



Effect on Rare Words (at most 5)



Precision



Recall

Overall recall set to match Model 4

Do better alignments help MT?

[Ganchev, Graca, Taskar, ACL 08]

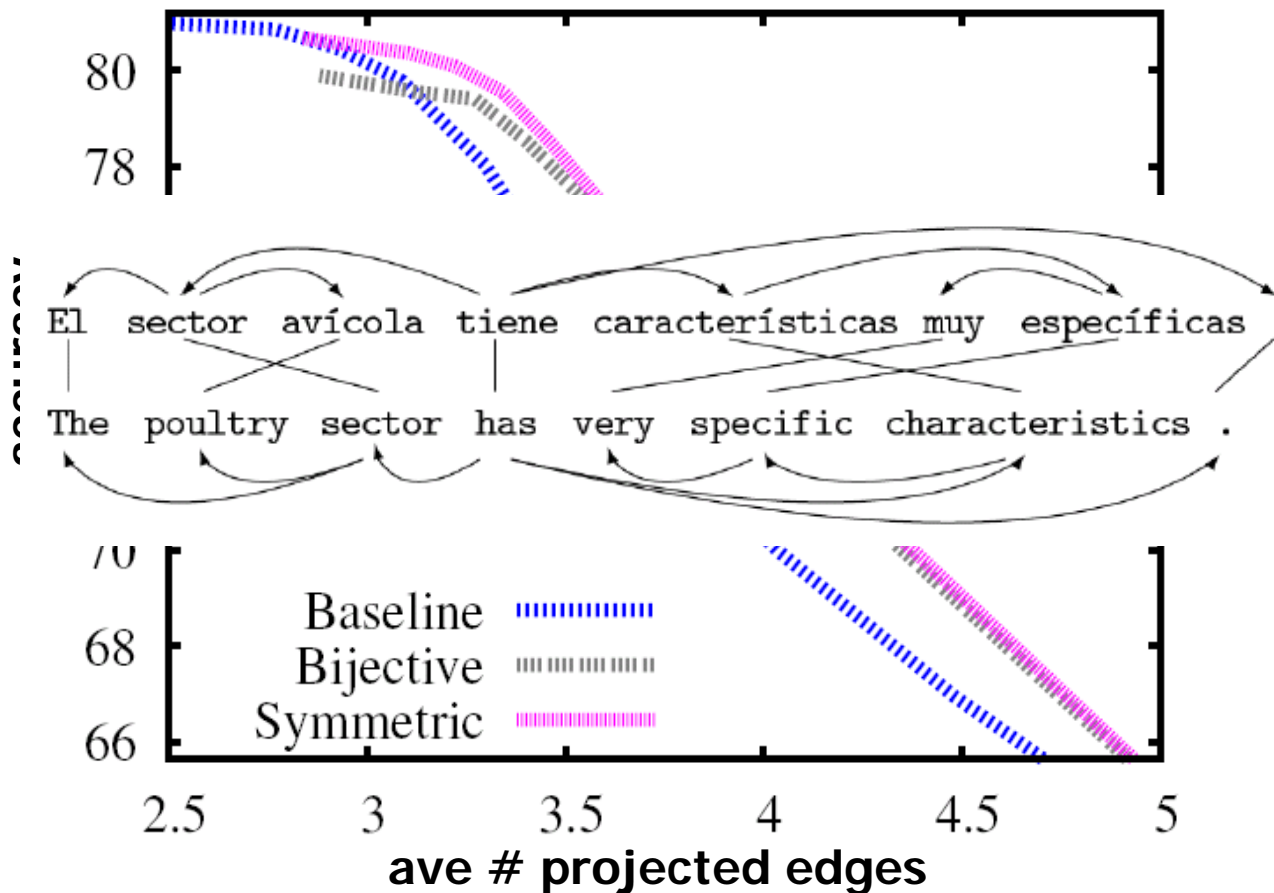
- Using MOSES [Koehn+ 07]
 - Phrase-based translation decoder
 - MERT to optimize params
 - 100K sents, using standard heuristics
 - BLEU Metric [Papineni+al 02]



	Regular	Bijective	Symmetric	Model 4
Fr-En	33.42	32.74	33.52	33.12
En-Fr	26.47	26.76	26.27	26.90
Es-En	30.18	30.41	30.32	30.24
En-Es	29.89	30.36	30.27	30.09
Pt-En	28.66	29.27	28.86	28.78
En-Pt	26.59	27.09	26.89	26.90

Alignments for Bitext Projection

Bulgarian Bitext Corpus (Tiedeman 07) ,
using parsers trained on CONLL 07 (Nivre et al)

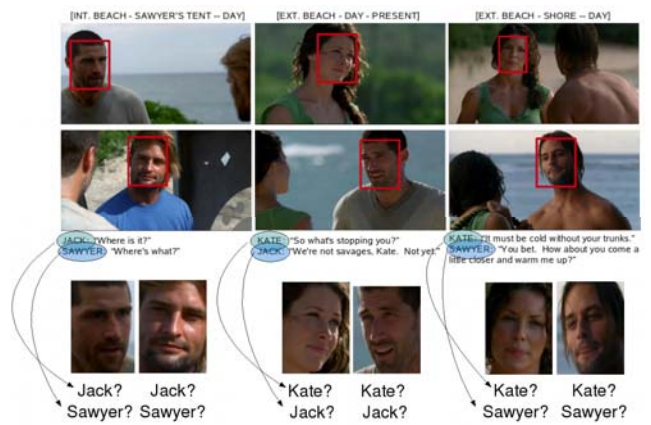
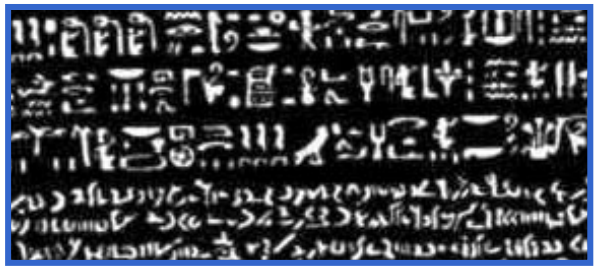


Posterior Regularization

- Framework for exploiting prior knowledge
 - Without complicating the model
 - Simple EM+projections algorithm
 - Intuitive objective: $L(\theta) - \mathbf{E}_{\mathbf{x}} KL(Q || p_{\theta}(\mathbf{y} | \mathbf{x}))$
- Related work
 - [Structural annealing: Smith & Eisner 04]
 - [Generalized Expectation Criteria: Mann & McCallum 08]
- Can directly enforce intractable constraints
 - Bijectivity, Agreement
 - Any linear constraint (eq/ineq) on posteriors
 - Grammar projection, other machine translation models
- Complementary to informative parameter priors

Correspondence across Languages and Modalities

- Words of different languages
- Faces, voices and names
- Movies and scripts
- Sound and transcription



- Towards principled, flexible framework for learning from weak supervision

Students who did all that:



Timothee Cour



Kuzman Ganchev



João Graça



Chris Jordan

With help from:
Akash Nagle,
Eleni Miltsakaki



Ben Sapp