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

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.)

[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.)

BOTTLE
Supervision in Learning

• Supervised

• Co-occurrence?

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

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

[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 -)
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.)

00:24:38 --> 00:24:39
The chinese people have water.

00:24:44 --> 00:24:45
Where did you get this?
Learning from Ambiguous Labels

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

• Each face has two or more possible labels
Faces in the News

- Image captions on the web

Hillary Clinton is 100% behind Barack Obama.

[Barnard+al 03, Blei+Jordan 03, Berg + al 04]
Ambiguous Labeling Setting

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

\[ \text{e.g., Jin+Ghahramani 02, Hullermeier+ Beringer 06} \]
Can we learn without true labels?

- No analysis (ttbomk)
- **Confounders:**
  \[
P(z=\text{Jack} \mid y=\text{Sawyer}, x) = 1
  \]
  Can’t tell them apart
- **Assumption:** \((\epsilon, \delta)\)-ambiguity
  \[
P(z \mid y, x) \text{ is less than } 1 \text{ (most of the time)}
  \]
  \[
  \epsilon = \sup_{(x,y) \in G, z} P(z \mid y, x) < 1
  \]
  \[
  G \text{ - set of good pairs } (x,y) \text{ where above holds}
  \]
  \[
P((x, y) \in G) = 1 - \delta
  \]
Ambiguity network for LOST
Generalization from Ambiguous Samples

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

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

Theorem: (assuming $(\epsilon, \delta)$-ambiguity)

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

Theorem: with probability $1 - \eta$

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

- **Multiclass Model:**
  \[ f(x) = \arg \max_k f^k(x); \quad f^k(x) = w^k \cdot x \]

- **One-Against-All Loss:**
  \[ \mathcal{L}(f(x), y) = \ell(f^y(x)) + \sum_{k \neq y} \ell(-f^k(x)) \]

- **Multilabel Loss (‘Naïve’ 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):**
  \[ \mathcal{L}(f(x), \{y, z\}) = \ell(\frac{f^y(x) + f^z(x)}{2}) + \sum_{k \neq y, z} \ell(-f^k(x)) \]

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

decompilation

rigid registration

part detection

frontal face detection
Tracking

false positives
false negatives
overlapping detections
no grouping

dynamic program

register

track# 1
track# 2
track# 3
Additional Features

4 x 50 PCA

[Ramanan et al., 2007]  [Everingham et al., 2006]
Naming Error

Lost (16 chars)

- Chance
- Gen. Model (IBM1)
- KNN
- Naïve Hinge
- Ambig Hinge
- Ambig Hinge++

8 episodes

- Sawyer
- Jack
- Jack
- Kate
- Jack
- Jack
- Jack
- Jack

[Image of bar chart showing error rates for different models]
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

**Screenplay**

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

**Video**

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

(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

identify:

- PEOPLE
- LOCATIONS
- OBJECTS
- ACTIONS
Action Dictionary

σηουτ  (JACK) (shouts) ()

σωιμ  (Sawyer) (wakes up) ()

φολλοω  (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) ()

[Cour +al, 08]  [Related: Laptev +al, 08]

Precision: 90%
Body Parsing via Locally Parametric CRFs
Naming without a screenplay

1st person reference

Hey, Jack!

I’m Jack.

2nd person reference

Where is Jack?

Jack in scene speaking

3rd person reference

Jack not in scene not speaking

Jack in scene not speaking

false positive

Jack-in-the-box

Supervision from dialogue:

- sparse
- indirect
- noisy

Grouping using continuity editing cues

~50 references per episode only

only constraints on possible labels

“It’s Jack”: 1st or 3rd person?
Grouping using Gestalt of Continuity Editing

180°-rule

An Attentional Theory of Continuity Editing [Smith, 2005]
Dialog-Only Naming

![Graph showing precision recall for different methods: ours, ours w/o gender, ours w/o gender or grouping, and prior (Jack).]
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
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.)

[SUN]

[SAYID]

[SUN]

[SAYID]

[SUN]

[SAYID]

[SUN]

[BOTTLE]

[SUN]

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

• 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?
Supervised Word Alignment

200 train, En/Fr

<table>
<thead>
<tr>
<th>Model</th>
<th>AER</th>
<th>Prec / Rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM model 4 (intersected)</td>
<td>6.5</td>
<td>98 / 88%</td>
</tr>
<tr>
<td>Our Alignment Model</td>
<td>4.3</td>
<td>96 / 95%</td>
</tr>
</tbody>
</table>

Best published accuracy on English-French (Hansards)

[Lacoste-Julien, Taskar, Klein, Jordan 06]
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: \( p(a, e, f; \theta) \)
HMM model [Ney, Vogel '96]

Generative model: \( p(a, e, f; \theta) \)

\( p(e) \) the railroad term is "demand loading"
HMM model [Ney, Vogel ’96]

Generative model: \( p(a, e, f; \theta) \)

\( p(e) \)  

the railroad term is “demand loading”
HMM model [Ney, Vogel ’96]

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

$p(e)$  

the railroad term is "demand loading"

le
HMM model [Ney, Vogel '96]

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

$p(e)$

the railroad term is “demand loading”
HMM model [Ney, Vogel '96]

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

$p(e)$

the railroad term is "demand loading"

le terme
HMM model [Ney, Vogel ’96]

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

\[ p(\mathbf{e}) \]

\[ p(a_j | a_{j-}; \theta_d) \quad \text{the railroad term is "demand loading"} \]

Distortion \( \theta_d \)

\[
\begin{align*}
p(\uparrow \uparrow) & = 0.6 \\
p(\swarrow \searrow) & = 0.2 \\
p(\uparrow \downarrow) & = 0.1 \\
& \ldots
\end{align*}
\]
HMM model [Ney, Vogel ’96]

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

\[
\begin{align*}
p(\mathbf{e}) \\
p(\mathbf{a}_j | \mathbf{a}_{j-}; \theta_d) \\
p(\mathbf{f}_j | \mathbf{e}_{a_j}; \theta_t) \\
\end{align*}
\]

Distortion \( \theta_d \)

\[
\begin{align*}
p(\uparrow \downarrow) &= 0.6 \\
p(\downarrow \uparrow) &= 0.2 \\
p(\downarrow \downarrow) &= 0.1 \\
\ldots
\end{align*}
\]

Translation \( \theta_t \)

\[
\begin{align*}
p(\text{the} \rightarrow \text{le}) &= 0.53 \\
p(\text{the} \rightarrow \text{la}) &= 0.24 \\
p(\text{railroad} \rightarrow \text{ferroviaire}) &= 0.19 \\
p(\text{NULL} \rightarrow \text{le}) &= 0.12 \\
\ldots
\end{align*}
\]

Note: model not symmetric
EM training

Maximize $p(e, f; \theta)$

Parameters: $\theta$

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

$a$
EM training

Maximize $p(e, f; \theta)$

Parameters: $\theta$

Expectation over alignments: $q$

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

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

$e, f \leftarrow a$
Posterior Over Alignments

<table>
<thead>
<tr>
<th>Matrix dimensions</th>
<th>source words</th>
<th>target words</th>
</tr>
</thead>
<tbody>
<tr>
<td>rows</td>
<td></td>
<td></td>
</tr>
<tr>
<td>columns</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word to word posterior probabilities $p$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.01 &lt; p$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.01 &lt; p &lt; 0.05$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.05 &lt; p &lt; 0.1$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.1 &lt; p &lt; 0.2$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.2 &lt; p &lt; 0.3$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.3 &lt; p &lt; 0.4$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.4 &lt; p &lt; 0.5$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.5 &lt; p &lt; 0.6$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.6 &lt; p &lt; 0.7$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.7 &lt; p &lt; 0.8$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.8 &lt; p &lt; 0.9$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.9 &lt; p &lt; 0.95$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$0.95 &lt; p$</td>
<td></td>
</tr>
</tbody>
</table>

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:** $x$  
  **Hidden:** $y$  
  **Model:** $p_\theta(x, y)$

- **Objective:**
  \[ L(\theta) = \mathbb{E}_x \log \sum_y p_\theta(x, y) = \mathbb{E}_x \log p_\theta(x) \]

- **E-step:**
  \[ q_x(y) = \arg \min_q KL(q(y) \| p_\theta(y \mid x)) \]
  \[ = p_\theta(y \mid x) \]

- **M-step:**
  \[ \max_\theta \mathbb{E}_x \mathbb{E}_q \log p_\theta(x, y) \]

- **Lower Bound:**
  \[ F(\theta, q) = \mathbb{E}_x \mathbb{E}_q \log \frac{p_\theta(x, y)}{q_x(y)} \leq L(\theta) \]

- A local max of $F(\theta, q)$ is a local max of $L(\theta)$
• **E-step:** (different)
  \[ q_x(y) = \arg \min_{q \in \mathcal{Q}_x} KL(q(y) || p_\theta(y | x)) \]
  constraints

• **M-step:** (same)
  \[ \max_\theta \mathbb{E}_x \mathbb{E}_q \log p_\theta(x, y) \]
Bijectivity Constraints

\[ p_\theta(y \mid x) \quad q_x(y) = \arg \min_{q \in \mathcal{Q}} KL(q(y) \mid \mid p_\theta(y \mid x)) \]
Symmetry (Agreement) Constraints

\[ p_{ef}(y \mid x) \quad \text{and} \quad p_{fe}(y \mid x) \]

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

\[ q_{ef}(y \mid x) \quad \text{and} \quad q_{fe}(y \mid x) \]
Posterior Regularization Objective

[Graca, Ganchev, Taskar, NIPS 07]

• **E-step:**

\[ q_x(y) = \arg \min_{q \in Q_x} KL(q(y) \| p_\theta(y \mid x)) \]

**constraints**

• **M-step:**

\[ \max_\theta \mathbb{E}_x \mathbb{E}_q \log p_\theta(x, y) \]

• **Theorem:** converges to a local max of

\[ L(\theta) - \mathbb{E}_x KL(Q_x \| p_\theta(y \mid x)) \]

\[ KL(Q_x \| p) = \min_{q \in Q_x} KL(q \| p) = \text{penalty for deviation from constraints} \]
E-step: I-projections

- If constraints are linear inequalities (or eqs)
  \[ \mathcal{Q}_x = \{ q_x(y) : \mathbb{E}_q f(x, y) \leq b \} \]
- Then \[ q_x(y) = \arg \min_{q \in \mathcal{Q}_x} KL(q(y) \| p_\theta(y \mid x)) \]
  \[ \propto p_\theta(y \mid x) \exp(-\lambda \cdot f(x, y)) \]
- If \( f(x, y) \) is a sum over factors of \( p_\theta(x, y) \)
  then \( q_x(y) \) 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:

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

- Gradient:

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

where

$$q_x(y) \propto p_\theta(y | x) \exp(-\lambda \cdot f(x, y))$$

- 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
Europarl (Pt/En, 1m sent.)

![Precision-Recall Curve](image)

- Bijective
- Regular
- Symmetric
- Viter
- Post AER
- Post F1
- Model 4
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]

<table>
<thead>
<tr>
<th></th>
<th>Regular</th>
<th>Bijective</th>
<th>Symmetric</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fr-En</td>
<td>33.42</td>
<td>32.74</td>
<td>33.52</td>
<td>33.12</td>
</tr>
<tr>
<td>En-Fr</td>
<td>26.47</td>
<td>26.76</td>
<td>26.27</td>
<td>26.90</td>
</tr>
<tr>
<td>Es-En</td>
<td>30.18</td>
<td>30.41</td>
<td>30.32</td>
<td>30.24</td>
</tr>
<tr>
<td>En-Es</td>
<td>29.89</td>
<td>30.36</td>
<td>30.27</td>
<td>30.09</td>
</tr>
<tr>
<td>Pt-En</td>
<td>28.66</td>
<td>29.27</td>
<td>28.86</td>
<td>28.78</td>
</tr>
<tr>
<td>En-Pt</td>
<td>26.59</td>
<td>27.09</td>
<td>26.89</td>
<td>26.90</td>
</tr>
</tbody>
</table>
Alignments for Bitext Projection

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

El sector avícola tiene características muy específicas.
The poultry sector has very specific characteristics.

Baseline  Bijective  Symmetric

ave # projected edges
Posterior Regularization

• Framework for exploiting prior knowledge
  – Without complicating the model
  – Simple EM+projections algorithm
  – Intuitive objective: \( L(\theta) - \mathbb{E}_x KL(Q || p_\theta(y | x)) \)

• 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