Joint Training for Open-domain Extraction on the Web:
Exploiting Overlap when Supervision is Limited

Rahul Gupta*
Google Research

Sunita Sarawagi
IIT Bombay

*Work done at IIT Bombay
Query-driven Extraction on the Web

User  →  Collective Extraction

<table>
<thead>
<tr>
<th>Gran Torino</th>
<th>Walt Kowalski</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dirty Harry</td>
<td>Harry Callahan</td>
<td>1971</td>
</tr>
</tbody>
</table>

Firefox 1982

Walt Kowalski (2008)...
Harry Callahan 1971...

Joe Kidd (1972)...
Joe Kidd...

Dirty Harry (1971)...
Inspector Harry Callahan...

Play Misty for Me (1971) Dave...

The Beguiled (1971) John McBurney...

Kelly's Heroes (1970) Kelly...

The Dead Pool (1988)...

Joe Kidd (1972)...

Joe Kidd...

Heartbreak Ridge (1985)...

Dirty Harry (1971)...

Inspector Harry Callahan...

Tightrope (1984)...

Joe Kidd (1972)...

Joe Kidd...

City Heat (1984)...

The Beguiled (1971) John McBurney...

Kelly's Heroes (1970) Kelly...

Sudden Impact (1983)...

Joe Kidd (1972)...

Joe Kidd...

Joe Kidd (1972)...

Joe Kidd...

Combine & de-duplicate, Rank, Display to the user
(World Wide Tables, Gupta & Sarawagi VLDB ’09)
A shared segment can be
• Arbitrarily long
• Across arbitrary number of sources
• Potentially a false-positive!
Content Overlap: Another Example

Eric Allman

(web page) Eric Allman is the main author of the sendmail program for delivering email, although certain alternatives have become popular, such as McKusick's partner.

Charles Babbage

Born: Monday, December 26, 1791, in London (England). Died: October 18, 1871, in London (England). Babbage is considered one of the forefathers of computer science for having designed (with the help of Ada Lovelace) the analytical engine, which, although not constructed, was the first (mechanical) computer. See also Babbage's biography on the NPG website.

John W. Backus


<table>
<thead>
<tr>
<th>CS inventors and their inventions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Codd</strong></td>
</tr>
<tr>
<td><strong>Cray</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>John Atanasoff</strong></th>
<th>Atanasoff–Berry Computer, though it was neither programmable nor Turing-complete.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Charles Babbage</strong></td>
<td>Designed the Analytical Engine and built a prototype for a less powerful mechanical calculator.</td>
</tr>
<tr>
<td><strong>John Backus</strong></td>
<td>Invented FORTRAN (Formula Translation), the first practical high-level programming language, and formulated the Backus-Naur form that described the formal language syntax.</td>
</tr>
</tbody>
</table>
Extraction Setting and Goal

Setting:

– Low supervision (~3 records)
– Multiple semi/un-structured sources (~20)
– Widely varying/disjoint feature sets across sources
– Significant but arbitrary and noisy content overlap

Goal: Jointly train one extraction model per source so that they agree on the labels of shared segments

Conditional Random Field
### Base Model: Linear CRF

**Sample sentence:** My review of Fermat’s last theorem by S. Singh

<table>
<thead>
<tr>
<th>$t$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>My</td>
<td>review</td>
<td>of</td>
<td>Fermat’s</td>
<td>last</td>
<td>theorem</td>
<td>by</td>
<td>S.</td>
<td>Singh</td>
</tr>
<tr>
<td>$y$</td>
<td>Other</td>
<td>Other</td>
<td>Other</td>
<td>Title</td>
<td>Title</td>
<td>Title</td>
<td>other</td>
<td>Author</td>
<td>Author</td>
</tr>
</tbody>
</table>

**Feature vector at position $t$**

$$
\log P_w(y|x) = \sum_t w \cdot f(y_t, y_{t-1}, t, x) - \log Z_w
$$

- **Trained weights**
- **Feature vector at position $t$**
- **“Log Partition”**

*(Lafferty et.al. ’01)*
Possible Alternatives

- Club sources, learn one CRF: Our features are disjoint
- Collective inference: Limited to overlapping content
- Hard label transfer: Co-training, multi-stage learning: prone to error cascades
- Two-source methods: 2-view perceptron/regression: We have multiple sources
- Known joint methods: Compared later
**Goal**

**Input:**  
$S$ data sources, each source $i$ has  
Labeled records $L_i$, Unlabeled records $U_i$  
Set $\mathcal{A} \equiv$ Shared segments across unlabeled records

**Goal:**  
Train CRF weights $w_i$ for each source $i = 1..S$

$$
\max_{\{w_1, \ldots, w_S\}} \sum_{i=1}^{S} \text{LogLikelihood}(L_i|w_i)
$$

$$
+ \sum_{(x,y) \in L_i} \log P(y|x, w_i)
$$

$$
+ \text{AgreementLikelihood}(\mathcal{A}, U_1, \ldots, U_S|w_1, \ldots, w_S)
$$
Goal

\[
\max_{\{w_1, \ldots, w_S\}} \sum_{i=1}^S LL(L_i|w_i) + C \cdot \log \sum_{y_A} \prod_{i=1}^S p_i^\text{marg}(y_A|w_i)
\]

Key Issue: Tractable approximation of the agreement
Re-writing the Agreement Term

\[ \sum_{y_a} p_1^{\text{marg}}(y_a) p_2^{\text{marg}}(y_a) \]

\[ = \sum_{y_a, y_b, y_c} p_1(y_a y_b) p_2(y_a y_c) \]

\[ = \sum_{y_a, y_b, y_c} \text{Score ( } \begin{array}{c} a \\
p_1(y_a y_b) \\
p_2(y_a y_c) \\c \end{array} \text{ )} \]

\[ \approx \text{PartitionFunction( } \begin{array}{c} a \\
p_1(y_a y_b) \\
p_2(y_a y_c) \\c \end{array} \text{ )} \]
Another Example

Three sentence snippets from different sources:

FOX – Matthew “Matt” Groening ,The Simpsons , 23rd Emmy winner Matt Groening ,The Simpsons (creator)

Four shared segments:
Matthew “Matt” Groening (1,2)
Matt Groening (1,2,3)
Matt Groening ,The Simpsons (2,3)
Simpsons (1,2,3)
Collapsing on Shared Segments

Collapsing on
“Matthew Matt Groening”

Collapse further on
“Matt Groening”

..and so on for the other shared segments
Agreement Term = \log \text{Partition}

Final "Fused" Graph: Collapse all shared segments

\[ \log \sum_{\mathbf{y}_A} \prod_{i=1}^{S} p_i(\mathbf{y}_A | \mathbf{w}_i) = \log Z_{\text{fused}} - \sum_{i=1}^{S} \log Z_i \]

Log Partition of the Fused Graph
Hard if the graph has cycles!
Approximating the Log-Partition

$$\log \sum_{y_A} \prod_{i=1}^{S} p_i(y_A|w_i) = \log Z_{\text{fused}} - \sum_{i=1}^{S} \log Z_i$$

Log $Z_{\text{fused}}$ can be approximated by
- Belief propagation (BP) on the fused graph
- Inexpensive variant of BP (Liang et al. ’09)

But…
- BP slow to converge, sometimes inconsistent
- Noisy agreement set => Wrong fused graph!
Alternate Approximation Method

- Collapse on all segments => Intractable cyclic graph
- Collapse on few segments => Maybe get a tractable tree?
Approximation via Partitioning

Partition A into disjoint sets of shared segments $A_1, \ldots, A_k$

\[
\log Z_{\text{fused}}(A) \approx \sum_{i=1}^{k} \log Z_{\text{fused}}(A_i)
\]

$A_1 = \text{Matt Groening, Matthew Matt Groening}$

$A_2 = \text{Simpsons, Matt Groening, The Simpsons}$
Each fused graph = a shared segment + its chains = Tree

...But total number of nodes is the highest possible
Partitioning Desiderata

\[
\min_{k, A_1, \ldots, A_k} \sum_i |\text{FusedGraph}(A_i)|
\]

\[A_1, \ldots, A_k \text{ a partition of } \mathcal{A}\]
\[\forall i, \text{ FusedGraph}(A_i) \text{ is a tree}\]

• **Low runtime**: Runtime linear in sizes of fused graphs
• **Preserve correlation**: Nearby shared segments should go to the same partition
  e. g. “Matthew Matt Groening” and “Matt Groening”
Partitioning Desiderata

\[
\min_{k, A_1, \ldots, A_k} \sum_i |\text{FusedGraph}(A_i)|
\]

\(A_1, \ldots, A_k\) a partition of \(A\)

\(\forall i, \text{FusedGraph}(A_i)\) is a tree

- **NP-hard** in size of agreement set
- **Greedy strategy:**
  - Grow \(A_i\) to maximally reduce objective
  - Tweaks and efficiency measures in paper
And we are done!

\[ \max_{\{w_1, \ldots, w_S\}} \sum_{i=1}^{S} LL(L_i|w_i) + C \cdot \log \sum_{y_A} \prod_{i=1}^{S} p_i^{\text{marg}}(y_A|w_i) \]

Equate to the Log Partition of the Fused Graph

\[ \log \sum_{y_A} \prod_{i=1}^{S} p_i(y_A|w_i) = \log Z_{\text{fused}} - \sum_{i=1}^{S} \log Z_i \]

Decompose via Greedy Partitioning into Fused Trees

\[ \log Z_{\text{fused}}(A) \approx \sum_{i=1}^{k} \log Z_{\text{fused}}(A_i) \]
Experiments: Structured Queries

<table>
<thead>
<tr>
<th>User</th>
<th>Gran Torino</th>
<th>Walt Kowalski</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dirty Harry</td>
<td>Harry Callahan</td>
<td>1971</td>
</tr>
</tbody>
</table>

Collective Extraction

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitchell Gant</td>
<td>Joe Kidd</td>
</tr>
<tr>
<td>1982</td>
<td>Joe Kidd</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dirty Harry (1971)</th>
<th>Joe Kidd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harry Callahan</td>
<td>Joe Kidd</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The Dead Pool (1988)</th>
<th>Joe Kidd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe Kidd</td>
<td>Joe Kidd</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Heartbreak Ridge (1978)</th>
<th>Joe Kidd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe Kidd</td>
<td>Joe Kidd</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pale Rider (1985)</th>
<th>Dave</th>
</tr>
</thead>
<tbody>
<tr>
<td>John McBurney</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tightrope (1984)</th>
<th>Kelly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kelly's Heroes (1970)</td>
<td>Kelly</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>City Heat (1984)</th>
<th>Kelly</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Beguiled (1971)</td>
<td>John McBurney</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sudden Impact (1983)</th>
<th>Kelly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kelly's Heroes (1970)</td>
<td>Kelly</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Firefox</th>
<th>Mitchell Gant</th>
<th>1982</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>City Heat</th>
<th>-</th>
<th>1984</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>-</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>-</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Joe Kidd</th>
<th>Joe Kidd</th>
<th>1972</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Merge & de-duplicate, Rank, Display to the user
Experimental Setting

• Extraction on 58 datasets, each representing a relation
  – Oil spills, James Cagney movies, University mottos, Parrots in Trinidad & Tobago, Star Trek novels etc.
  – Each dataset = 2-20 HTML list sources from a 500M crawl
  – Wide range of #columns, #sources, #records, #shared segments, base accuracy, noise
  – Handful (~ 3) labeled records per list source
  – F1 measured using manually annotated ground truth

• Datasets binned by Base model F1 and Average Number of Shared Segments for ease of presentation
Finding the Agreement Set

- **Traditional:** *Shared segment = Unigram repetitions*
  - Arbitrary, context-oblivious, highly noisy
  - Does not transfer weights of first-order features
- **Our strategy:**
  - *Shared segment = Repeating segment in near-duplicate records*
  - Maximally long segment inside a record cluster
  - Approximate multi-partite matching of sources yields record clusters
Comparison vs Simpler Methods

- Label transfer: cascade-prone, 10% drop in some cases
- Collective inference: boosts $83.3\%$ to $86.1\%$
- Joint training: boosts to $87.5\%$
  - With 7 training records: boosts $87.4\%$ to $89.2\%$
Runtime/Accuracy of All Methods

Greedy-partitioning has the best runtime/accuracy tradeoff

Belief Propagation (BP) quite slow, Fast variant (BP’) not as accurate
# Relative Error Reduction

<table>
<thead>
<tr>
<th></th>
<th>50F</th>
<th>50M</th>
<th>60F</th>
<th>60M</th>
<th>70F</th>
<th>70M</th>
<th>80F</th>
<th>80M</th>
<th>90F</th>
<th>90M</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>44.8</td>
<td>45.4</td>
<td>33.1</td>
<td>32.7</td>
<td>26.5</td>
<td>23.9</td>
<td>14.4</td>
<td>13.4</td>
<td>5.7</td>
<td>3.9</td>
<td>16.7</td>
</tr>
<tr>
<td><strong>Absolute F1 Error of Base</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CInfer</strong></td>
<td>1.7</td>
<td>3.2</td>
<td>10.4</td>
<td>3.3</td>
<td>-2.9</td>
<td>16.4</td>
<td>31.3</td>
<td>28.2</td>
<td>10.1</td>
<td>13.1</td>
<td>17.0</td>
</tr>
<tr>
<td><strong>Tree</strong></td>
<td>6.0</td>
<td>2.3</td>
<td>11.2</td>
<td>9.5</td>
<td>4.4</td>
<td>28.0</td>
<td>38.0</td>
<td>40.6</td>
<td>43.4</td>
<td>13.8</td>
<td>25.5</td>
</tr>
<tr>
<td><strong>Seg</strong></td>
<td>6.6</td>
<td>0.6</td>
<td>14.3</td>
<td>9.8</td>
<td>4.5</td>
<td>31.5</td>
<td>38.8</td>
<td>42.7</td>
<td>36.2</td>
<td>9.3</td>
<td>26.8</td>
</tr>
<tr>
<td><strong>BP</strong></td>
<td>6.0</td>
<td>2.4</td>
<td>10.6</td>
<td>9.3</td>
<td>3.6</td>
<td>28.7</td>
<td>38.6</td>
<td>42.0</td>
<td>43.3</td>
<td>14.9</td>
<td>26.0</td>
</tr>
<tr>
<td><strong>BP’</strong></td>
<td>1.6</td>
<td>2.1</td>
<td>11.8</td>
<td>3.5</td>
<td>-3.1</td>
<td>18.6</td>
<td>34.3</td>
<td>35.0</td>
<td>13.2</td>
<td>-0.5</td>
<td>19.1</td>
</tr>
<tr>
<td><strong>PR</strong></td>
<td>2.3</td>
<td>7.9</td>
<td>4.7</td>
<td>10.3</td>
<td>4.1</td>
<td>28.7</td>
<td>30.5</td>
<td>33.3</td>
<td>30.2</td>
<td>9.3</td>
<td>22.4</td>
</tr>
</tbody>
</table>

**Red:** Increase in error  
**Green:** Best method
Experiments: Noisy Agreement Set

- Our scheme: ~5% token-level noise, small F1 drop
- Arbitrary unigrams: ~15% node noise, significant F1 drop
Related Work

• Agreement-based learning (Liang et.al. ’09)
  – EM-based scheme applied on two sources with clean overlap

• Posterior Regularization (Ganchev et.al. ’08)
  – Different agreement term; used in multi-view

• Two-view perceptron/regression, co-training/boosting/SVMs (Brefeld et.al. ’05, Blum & Mitchell ’98, Collins & Singer ’99, Sindhwani et.al. ’05, Kakade & Foster ’07)
  – Two source and/or hard label transfer

• Multi-task learning (Ando & Zhang ’05)
  – Single source, shared features sought

• Semi-supervised learning (Chapelle et.al. ’06)
  – No training, no support for partially structured overlaps

• Co-regularization, Pooling (Suzuki et.al. ’07)
Summary

• Joint training: Text overlap compensates for supervision
  – Reward agreement of distributions on overlapping text
  – Tractable approximations of the reward
  – Scheme to find low-noise overlapping segments
  – Extensive empirical comparison on many datasets

Best accuracy/speed tradeoff using content overlap

= Decomposing agreement over greedy tree partitions

• Future work
  – Online and parallel collective training
Thanks