Cross Document Entity Disambiguation

August 22, 2007

Johns Hopkins Summer Workshop
Person Entities for CDC

- **Corpora:** David Day, Janet Hitzeman
- **Relations:** Su Jian, Stanley Yong, Claudio Guiliano, Gideon Mann
- **CDC Features:** Jason Duncan, Paul McNamee, Rob Hall, Mike Wick
- **Clustering/Machine Learning:** Mike Wick, Rob Hall
Problem Description

- Disambiguate entities across a corpus
  - **Document level, Entity level, Mention level**
- **Document level disambiguation / Web People Corpora**
  - SPOCK corpus (description – challenge page – discussion forum)
- **High ambiguity level**
  - “The Jim Smith Society” | “James Smith”
Google Search for ‘James Smith’

- **James Smith Cree Nation** James Smith Cree Nation P.O. Box 1059 Melfort, Saskatchewan S0E 1A0. Ph: (306) 864–3636 Fx: (306) 864–3336. www.sicc.sk.ca/bands/bjames.html - 1k - Cached - Similar pages

- **James Smith (political figure) - Wikipedia, the free encyclopedia**
  James Smith (about 1719 – July 11, 1806), was a signer to the United States Declaration of Independence as a representative of Pennsylvania. ... en.wikipedia.org/wiki/James_Smith_(political_figure) - 22k - Cached -

- **Band Details** Official Name, James Smith. Number, 370. Address, PO BOX 1059, MELFORT, SK. Postal Code, S0E 1A0. Phone, (306) 864-3636. Fax, (306) 864-3336 sdiprod2.inac.gc.ca/fnprofiles/FNProfiles_DETAILS.asp?BAND_NUMBER=370 - 12k

- **Comox Valley Real Estate: James Smith, your Realtor for Comox ...**
  James Smith is your realtor for the Comox Valley area, including Comox, Courtenay, Cumberland, Union Bay, Royston, and Black Creek. www.jamessmith.ca/ - 10k - Cached - Similar pages

- **Watercolor Snapshots - by James Smith** Watercolor Snapshots by James Smith - your portrait custom painted in watercolor, or the portrait of your relative or friend, painted from your 4 x 6 ... 28k - Cached - Similar pages
Problem Description

• Entity level disambiguation (ACE 2005 + CDC annotation)
  – PER, GPE, LOC, and ORG entities that have a NAME string on the coreference chain AND are +SPECIFIC

• Low ambiguity level (B-cubed baseline of 0.80 F versus 0.09 F for Spock corpus for “shatter all” condition, one node per cluster)
Features from previous work

• Document level bag of words/NER entities features (basically all previous systems)
• Local contextual information (bags of words/NER entities in local context).
• Syntactic Features (base NPs in document/local contexts Chen & Martin)
• Basic Relational Information (Mann & Yarowsky). DOB, POB, etc.
Three areas where workshop can make a contribution

1. **More and better features**
2. **More use of relation information** from varying sources (ground truth and system generated relations, supervised and unsupervised)
3. **Different clustering procedures** than standard greedy single-link agglomerative clustering
Experiments

- Document Level Information (Bow, Boe)
- Mention Level Information (Bow, Boe)
- Topic Models (SPOCK)
- Relations
  - ACE (supervised): ORG-AFF 82-F, PER-SOC 91-F
  - SPOCK (unsupervised)
- Different Clustering Techniques (SPOCK)
Entity Disambiguation

Michael Wick
Which J. Smith?

- Jazz Musician
- Politician
- Student
- Investor
- CEO
- Historian

Web Docs
First Order Model

Classifier/evaluator

SCORE

\[ P(Y|X) = \frac{1}{Z_x} \prod f(y_i, x_i) \prod (1-f(y_{ij}, x_{ij})) \]

Find configuration to maximize this objective function

Greedy Agglomerative Approximation
(Traditional Features)

- Bags of Words
- Title overlaps
- Named Entities
- Chunking
- TFIDF term selection
Two articles on John Smith, the jazz musician

…and John Smith on saxophone..

…his melodic improvisation..

NO WORDS IN THE EXCERPTS OVERLAP!!!
Which J. Smith?

- Jazz Musician
  
  ...his melodic improvisation..

- Student

- Investor

- CEO

- Historian

1 university program learning students education
2 ashley love ash miss hey
3 tx la christi corpus san
4 company insurance financial income investor
5 contact email city state fullname
6 masters music jazz folk orchestra
7 registered status category read japan
8 museum county historical society kansas
9 photography times braunfels jp rail
10 film alex films kill night
11 senate senator section court company

...and John Smith on saxophone..
## Results With Topics

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B-Cubed</strong></td>
<td>.32</td>
<td>.24</td>
<td>.28</td>
</tr>
<tr>
<td>+topics</td>
<td>.23</td>
<td>.44</td>
<td>.30</td>
</tr>
<tr>
<td><strong>Pairwise</strong></td>
<td>.12</td>
<td>.19</td>
<td>.15</td>
</tr>
<tr>
<td>+topics</td>
<td>.13</td>
<td>.44</td>
<td>.20</td>
</tr>
<tr>
<td><strong>MUC</strong></td>
<td>.70</td>
<td>.65</td>
<td>.67</td>
</tr>
<tr>
<td>+topics</td>
<td>.84</td>
<td>.86</td>
<td>.85</td>
</tr>
</tbody>
</table>

Chunks+TitleOverlap + TFIDF + NER + Relations + **TOPICS!!!**
Metropolis-Hastings Sampler

\[ P(Y|X) = \frac{1}{Z_x} \prod f(y_i, x_i) \prod (1-f(y_{ij}, x_{ij})) \]

Requires summing over all possible configurations

\[ P(y'|y) = \max \left[ \frac{p(y')}{p(y)} \times \frac{p(y|y')}{p(y'|y)}, 1 \right] \]

Probability of Accepting Jump

Likelihood ratio: partition function cancels!

Inverse ratio of making a jump vs. reversing that jump
Metropolis-Hastings

1. Initialize with Greedy Agglomerative

2. Pick a block from cohesion distribution

3. Pick new assignment uniformly at random

4. Accept with probability:

\[ P(y'|y) = \text{Max}[ \frac{p(y')}{p(y)} * \frac{p(y'|y)}{p(y'|y)}, 1]\]

(in this example we reject)
Deriving Cohesion Distribution

Block Distribution

BE  .97
AC  .91
BED  .84
BEDF  .84
## Results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B-Cubed</strong></td>
<td>.318</td>
<td>.312</td>
<td>.315</td>
</tr>
<tr>
<td><em>w/MH</em></td>
<td>.318</td>
<td>.435</td>
<td>.367</td>
</tr>
<tr>
<td><strong>Pairwise</strong></td>
<td>.271</td>
<td>.243</td>
<td>.256</td>
</tr>
<tr>
<td><em>w/MH</em></td>
<td>.278</td>
<td>.370</td>
<td>.317</td>
</tr>
<tr>
<td><strong>MUC</strong></td>
<td>.838</td>
<td>.851</td>
<td>.844</td>
</tr>
<tr>
<td><em>w/MH</em></td>
<td>.863</td>
<td>.877</td>
<td>.870</td>
</tr>
</tbody>
</table>

*Metropolis Hastings*  
*Greedy Agglomerative*
## CDC results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B-Cubed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+DOC</td>
<td>.96</td>
<td>.88</td>
<td>.92</td>
</tr>
<tr>
<td>+MEN</td>
<td>.97</td>
<td>.90</td>
<td>.93</td>
</tr>
<tr>
<td>+CHAIN</td>
<td>.96</td>
<td>.90</td>
<td>.93</td>
</tr>
<tr>
<td>+LEX</td>
<td>.88</td>
<td>.93</td>
<td>.91</td>
</tr>
<tr>
<td>+REL</td>
<td>.97</td>
<td>.95</td>
<td>.96</td>
</tr>
<tr>
<td><strong>Pairwise</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+DOC</td>
<td>.92</td>
<td>.57</td>
<td>.71</td>
</tr>
<tr>
<td>+MEN</td>
<td>.94</td>
<td>.70</td>
<td>.80</td>
</tr>
<tr>
<td>+CHAIN</td>
<td>.94</td>
<td>.70</td>
<td>.80</td>
</tr>
<tr>
<td>+LEX</td>
<td>.80</td>
<td>.87</td>
<td>.84</td>
</tr>
<tr>
<td>+REL</td>
<td>.95</td>
<td>.88</td>
<td>.91</td>
</tr>
<tr>
<td><strong>MUC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+DOC</td>
<td>.89</td>
<td>.78</td>
<td>.83</td>
</tr>
<tr>
<td>+MEN</td>
<td>.91</td>
<td>.79</td>
<td>.85</td>
</tr>
<tr>
<td>+CHAIN</td>
<td>.90</td>
<td>.79</td>
<td>.84</td>
</tr>
<tr>
<td>+LEX</td>
<td>.74</td>
<td>.83</td>
<td>.79</td>
</tr>
<tr>
<td>+REL</td>
<td>.92</td>
<td>.87</td>
<td>.89</td>
</tr>
</tbody>
</table>

BCubed (shattered): pr=1.0  re=.67  F1=.80
Conclusions

- Topics enable additional and powerful features
- Metropolis-Hastings improves upon greedy method
Generative, Unsupervised Models for Web People Disambiguation

Rob Hall (UMass Amherst)
A Simple Generative Process

Each document “belongs” to some hidden entity.

Cluster by determining the mode of the posterior distribution of $e$ given the words.

A sequence of observations is generated, conditioned on the entity.
Word Model

Entity in document

“Bag of words” (ignore sequential nature)

“Dirichlet Process”

Per-entity multinominal over words

(Symmetric) Dirichlet Prior
Approximate Inference: Collapsed Gibbs

Start with some initialization of all $e$.

Then resample each $e^k$ in turn from the distribution:

$$p(e^k | \bar{e}^{-k}, \bar{w}) \propto p(\tilde{w}^k | e^k) \cdot p(e^k | \bar{e}^{-k})$$

With $e^{-k}$ sampled we can use CRP:

$$p(e^k = i | e^{-k}) \propto \frac{n_i}{n - 1 + \alpha} \quad p(e^k = \text{new} | e^{-k}) \propto \frac{\alpha}{n - 1 + \alpha}$$

Then can integrate out $\Theta$ using Polya-urn scheme:

$$p(\tilde{w}^k = w_0..w_m | e^k) \propto \prod_i \frac{\beta + C_{w_i}^{e^k} + C_{0..i-1}^{w_i}}{\sum_{w \in V} \beta + C_{e^k}^w + C_{0..i-1}^w}$$
Same as before

Bernoulli (binary) variable that determines whether \( w \) is drawn from the local or global distribution (gibbs sampled).

Asymmetric Dirichlet prior (learned).

"Global" distribution Over words
Approximate Inference: Collapsed Gibbs

The new probability model is:

\[ p(e^k | \tilde{e}^{-k}, \tilde{w}) \propto p(\tilde{w}^k | \tilde{z}^k, e^k, \tilde{w}) \cdot p(\tilde{z}^k | e^k, \tilde{w}) \cdot p(e^k | \tilde{e}^{-k}) \]

This requires sampling \( z \):

\[
p(\tilde{z}^k_i = 0 | e^k, \tilde{w}, \tilde{z}) \propto \frac{\beta + C_{e^k}^{w_i}}{\sum_w \beta + C_w^{e^k}} \quad p(\tilde{z}^k_i = 1 | e^k, \tilde{w}, \tilde{z}) \propto \frac{\gamma^{w_i} + C_{G}^{w_i}}{\sum_w \gamma^{w} + C_G^{w}}
\]

Then when calculating \( p(e|w) \) only use the \( w \) which correspond to \( z = 0 \). (The probability of words from the global topic is absorbed into the normalizing constant).
Incorporating other Evidence

Duplicate the “bag of words” probability model for each other class of observation.

“Bag of” observations for each evidence class.
Gibbs Sampling in this Model

The new probability model is:

\[ p(e_k^k | \tilde{e}^{-k}, \tilde{w}, \tilde{z}) \propto p(e_k^k | \tilde{e}^{-k}) \prod_c p(\tilde{w}_c^k | \tilde{z}_c^k, e^k, \tilde{w}) \cdot p(\tilde{z}_c^k | e^k, \tilde{z}_c^{-k}, \tilde{w}) \]

Start with an initialization of \( e \)

Iterate over each document \( k \):

For each type of observation \( c \):

Resample \( z_c^k \)

Resample \( e^k \)
<table>
<thead>
<tr>
<th>Model</th>
<th>B-Cubed</th>
<th>Pairwise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Words</td>
<td>63.3</td>
<td>16.4</td>
</tr>
<tr>
<td>Words-Stop</td>
<td>69.7</td>
<td>17.3</td>
</tr>
<tr>
<td>Words-URL</td>
<td>51.7</td>
<td>18.2</td>
</tr>
<tr>
<td>Words-URL-Stop</td>
<td>60.9</td>
<td>18.6</td>
</tr>
<tr>
<td>Words-URL-WN</td>
<td>48.6</td>
<td>19.8</td>
</tr>
<tr>
<td>Words-URL-WN-Stop</td>
<td>63.2</td>
<td>19.2</td>
</tr>
</tbody>
</table>