Learning on the Web

Fernando Pereira

Thanks to
Koby Crammer, Kuzman Ganchev, João Graça, Yi Liu, Marius Pasca, Franz Och, Joseph Reisinger, Deepak Ravichandran, Stefan Riezler, Partha Talukdar, Ben Taskar, Alexander Vasserman, and other colleagues at Google and University of Pennsylvania.
Penn work funded by NSF
Web Inference and Learning

• Building and using the Web requires algorithms that can draw reliable inferences from the wealth of evidence implicit in Web data

• How to interpret this term in this context?

• Does this sentence answer that question?

• Will this user click on that ad?

• *Learning*: create concise representations to support good inferences
Beyond Supervised Learning

- Explicit annotation is costly and misleading
  - *Expert annotation*: difficult to agree on annotation criteria, repeated failures to achieve self-sustainability:
  - *User annotation*: what’s its immediate value to the user, attractive to spammers
- Learning from what users do: we constantly seek and organize information
  - Examples: *machine translation, query expansion*
- Lots of (mostly) unlabeled data
Learning from Parallel Data
Statistical Machine Translation

- *Inference*: is this phrase a good translation of that phrase in this context?
- Indirect evidence: translation pairs, monolingual text
- Memory-based (non-parametric) learning:
  - Bilingual: phrase translation table
  - Monolingual: language model
Machine learning from *human* communication

- Parallel texts:

**DE**

SEHR GEEHRTER GAST!
KUNST, KULTUR UND KOMFORT IM HERZEN BERLIN.

DIE ÖRTLICHE NETZSPANNUNG BETRÄGT 220/240 VOLT BEI 50 HERTZ.

**EN**

DEAR GUESTS,
ART, CULTURE AND LUXURY IN THE HEART OF BERLIN.

THE LOCAL VOLTAGE IS 220/240 VOLTS 50 HZ.
Building a Translator

- Align (cf. genomic alignments…)

KUNST, KULTUR UND KOMFORT IM HERZEN BERLINS.

ART, CULTURE AND LUXURY IN THE HEART OF BERLIN.

- Extract phrase pairs from alignments
- Model the statistics of the target language
- To translate a source text:
  - search over phrase-to-phrase translations
  - filter with model of target language

Friday, May 7, 2010
Meaning from Web Search

- *Term mismatches* between queries and documents in information retrieval

- Query terms and relevant document terms differ

- *Query expansion*: add search terms known to occur in relevant documents

  - Increase recall

  - Decrease query term ambiguity
Query Expansion by Translation

- Training
  - Align queries and clicked snippets
  - Create translation tables
  - Train $n$-gram language model from query logs
- Use
  - “Translate” many queries
  - Extract and store table of “translated” terms in context
From query to snippets

Google search results for herbs for mexican cooking

- Healing with Herbs and Rituals: A ...
  - $18.95 - Barnes & Noble.com
- Healing with Herbs and Rituals: A ...
  - $18.95 - Bookstrip.com
- FRONTIER HERBS TACO & MEXICAN SEASONING ...
  - $17.34 - www.southnatural.com

See herbs for mexican cooking results available through Google Checkout

Spices & Herbs

Spices and Herbs at MexGrocer.com, a nationwide online grocery store for authentic Mexican food, household products, cooking recipes, cookbooks and culture.

www.mexgrocer.com/catagories-spices---herbs.html - 29k - Cached - Similar pages

MEXICO HOT OR NOT - Las Hierbas de Cocina: A Culinary Guide to ...

Those with no inclination toward tending even the hardiest herbs can now find several of the more common herbs used in Mexican cooking being sold in the ...

www.mexconnect.com/mex_/recipes/puebla/kgyerbas.html - 22k - Cached - Similar pages
## Ambiguity Resolution

- **5-best phrase-level translations**

| (herbs, herbs) (for, for) (chronic, chronic) (heartburn, heartburn) |
| (herbs, herb) (for, for) (chronic, chronic) (heartburn, heartburn) |
| (herbs, remedies) (for, for) (chronic, chronic) (heartburn, heartburn) |
| (herbs, medicine) (for, for) (chronic, chronic) (heartburn, heartburn) |
| (herbs, supplements) (for, for) (chronic, chronic) (heartburn, heartburn) |

| (herbs, herbs) (for, for) (mexican, mexican) (cooking, cooking) |
| (herbs, herbs) (for, for) (cooking, cooking) (mexican, mexican) |
| (herbs, herbs) (for, for) (mexican, mexican) (cooking, food) |
| (mexican, mexican) (herbs, herbs) (for, for) (cooking, cooking) |
| (herbs, spices) (for, for) (mexican, mexican) (cooking, cooking) |
Beyond Parallel Corpora
• Elementary semantic inference

• First step: what are the possible classes for each instance?
Someone Told Us

- Text patterns (Hearst 92, Van Durme & Pasça 08)

Google

"such as whistler"

Web

Whistler and Vail are best American ski resorts for advanced ...
Dec 21, 2009 ... Famous ski resorts such as Whistler in Canada and Vail in the US are great for skiers of all abilities as well as snowboarders, ...

Whistler Internet Cafes
As is the case with many ski resorts such as Whistler, many of the local hotels have computers available to their guests so that they don't need to go to an ...

Florida-based group eyes Whistler auction
When it comes to major resorts such as Whistler, observers say there are few potential bidders especially in the current economic environment. ...

WHISTLER STUDIES Arriving in Europe in 1855 at the age of twenty ...
by J Sandberg - 1968 - Cited by 1 - Related articles
Great artists, such as Whistler, have a rare gift, "one supreme quality of spirit," which is none other than "the love of beauty for the very beauty's sake ...

John Singer Sargent, "Carnation, Lily, Lily, Rose," and the ...
by AL Helmreich - 2003 - Cited by 2 - Related articles
artists such as Whistler or Clausen who were eager to claim relative autonomy (despite their recourse to alternative professional artists' societies). ...

www.jstor.org/stable/3048512


www.calgaryherald.com/story_print.html?id=2595550&sponsor= - Cached

www.jstor.org/stable/3830184
Informative Co-occurrences

• WebTables (Cafarella et al. 08)

• 154M HTML tables from Web pages

• Cluster instances in table columns
Combining Information Sources

- **Bootstrapping:**
  - Seed set of (instance, class) pairs
  - Compute instance similarity from additional sources
  - Use similarity to infer new (instance, class) pairs

- **Approach:** label propagation in a graph
  - instance nodes
  - cluster/class nodes
  - bipartite structure
Label Propagation
### Label Propagation

<table>
<thead>
<tr>
<th>WT</th>
<th><strong>Musician</strong></th>
<th></th>
<th><strong>Singer</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Singer</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Billy Joel (0.75)</td>
<td></td>
<td>Bob Dylan (0.95)</td>
</tr>
<tr>
<td></td>
<td>Johnny Cash (0.73)</td>
<td></td>
<td>Johnny Cash (0.87)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Billy Joel (0.82)</td>
</tr>
</tbody>
</table>

Friday, May 7, 2010
## Label Propagation

<table>
<thead>
<tr>
<th>Musician</th>
<th>Singer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Billy Joel (0.75)</strong></td>
<td><strong>Bob Dylan (0.95)</strong></td>
</tr>
<tr>
<td><strong>Johnny Cash (0.73)</strong></td>
<td><strong>Johnny Cash (0.87)</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Billy Joel (0.82)</strong></td>
</tr>
</tbody>
</table>
Label Propagation

<table>
<thead>
<tr>
<th>Musician</th>
<th>Singer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Billy Joel (0.75)</td>
<td>Bob Dylan (0.95)</td>
</tr>
<tr>
<td>Johnny Cash (0.73)</td>
<td>Johnny Cash (0.87)</td>
</tr>
<tr>
<td></td>
<td>Billy Joel (0.82)</td>
</tr>
</tbody>
</table>
Label Propagation

- **Musician**
  - Billy Joel (0.75)
  - Johnny Cash (0.73)

- **Singer**
  - Bob Dylan (0.95)
  - Johnny Cash (0.87)
  - Billy Joel (0.82)

Extraction Confidence

Cluster ID

Friday, May 7, 2010
Label Propagation

Musician
- Billy Joel (0.75)
- Johnny Cash (0.73)

Singer
- Bob Dylan (0.95)
- Johnny Cash (0.87)
- Billy Joel (0.82)

Singer
- 0.95
- 0.87
- 0.82

Musician
- 0.75
- 0.73

Cluster ID
- A8

Extraction Confidence
Label Propagation

<table>
<thead>
<tr>
<th>Singer</th>
<th>Musician</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob Dylan</td>
<td>0.95</td>
</tr>
<tr>
<td>Johnny Cash</td>
<td>0.87</td>
</tr>
<tr>
<td>Billy Joel</td>
<td>0.82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Singer</th>
<th>Musician</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob Dylan</td>
<td>0.95</td>
</tr>
<tr>
<td>Johnny Cash</td>
<td>0.87</td>
</tr>
<tr>
<td>Billy Joel</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Seed classes

Extraction Confidence

Cluster ID
Classes, Clusters, Instances

Singer

- Bob Dylan: 0.95

Musician

- Johnny Cash: 0.87
- Billy Joel: 0.82

Classes, Clusters, Instances

Initial clusters

Singer

Musician

Bob Dylan

Johnny Cash

Billy Joel

0.95

0.87

0.82

0.73

0.75

Musician
Initial clusters

- **Singer**
  - Bob Dylan: 0.95
  - Johnny Cash: 0.82
  - Billy Joel: 0.75

- **Musician**
  - Musician: 0.87
  - Musician: 0.73

**Smoothness:** Nodes linked by heavy edges tend to have similar labels.
Classes, Clusters, Instances

Initial clusters

Coupling Nodes:
Encourage linked instance nodes to have similar class labels (smoothness)

Bob Dylan
Musician
Smoothness:
Nodes linked by heavy edges tend to have similar labels

Johnny Cash
Musician

Billy Joel
Musician
Classes, Clusters, Instances

- **Initial clusters**
- **Singer**
- **Musician**
- **Bob Dylan**
- **Johnny Cash**
- **Billy Joel**

**Smoothness**: Nodes linked by heavy edges tend to have similar labels

**Coupling Nodes**: Encourage linked instance nodes to have similar class labels (smoothness)

**Seed classes can be different from the cluster IDs of first phase extractors (A8, WT, etc.)**
Label Propagation

Singer

Bob Dylan

Johnny Cash

Musician

Billy Joel

Friday, May 7, 2010
Label Propagation

Initialization

Seed classes

Singer

Bob Dylan

Johnny Cash

Billy Joel

Musician

Musician 1.0

Musician 1.0

Musician 1.0

Musician 1.0
Label Propagation

Iteration 1

Seed classes

- Bob Dylan
- Johnny Cash
- Billy Joel

- Singer
- Musician

Musician 0.8
Musician 1.0
Musician 0.8
Musician 1.0
Musician 1.0
Musician 1.0
Musician 1.0
Musician 1.0

Friday, May 7, 2010
Label Propagation

Iteration 2

Derived classes
Seed classes

Singer

Musicians:
- Musician 0.8
- Musician 1.0
- Musician 0.6
- Musician 1.0

Bob Dylan

Johnny Cash

Billy Joel

Friday, May 7, 2010
Label Propagation

Iteration 2

Singer

Bob Dylan

Johnny Cash

Musician

Billy Joel
Class Assignment for Given Instances
Class Assignment for Given Instances

☐ A8
☐ Propagation
☐ WebTables
Class Assignment for Given Instances

924k (class, instance) pairs extracted from 100M web documents

74M (class, instance) pairs extracted from WebTables
Class Assignment for Given Instances

Graph with 1.4M nodes, 75M edges

- A8
- Propagation
- WebTables

Graph with 1.4M nodes, 75M edges
Class Assignment for Given Instances

\[ \text{MRR} = \frac{1}{|\text{test-set}|} \sum_{v \in \text{test-set}} \frac{1}{\text{rank}_v(\text{class}(v))} \]
Class Assignment for Given Instances

\[
MRR = \frac{1}{|\text{test-set}|} \sum_{v \in \text{test-set}} \frac{1}{\text{rank}_v(\text{class}(v))}
\]

Evaluation against Wordnet excerpt (38 classes, 8910 instances)

Graph with 1.4M nodes, 75M edges
Finding More Good Instances

- Instances found solely by label propagation:

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision at 100 (non-A8 extractions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book Publishers</td>
<td>87.36</td>
</tr>
<tr>
<td>Federal Agencies</td>
<td>29.89</td>
</tr>
<tr>
<td>NFL Players</td>
<td>94.95</td>
</tr>
<tr>
<td>Scientific Journals</td>
<td>90.82</td>
</tr>
<tr>
<td>Mammal Species</td>
<td>84.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NFL Players</td>
<td>Tony Gonzales, Thabit Davis, Taylor Stubblefield, Ron Dixon, Rodney Hannah</td>
</tr>
</tbody>
</table>
Discovering Similar Classes

- Label propagation by-product:

<table>
<thead>
<tr>
<th>Seed Class</th>
<th>Non-Seed Class Labels Discovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book Publishers</td>
<td>small presses, journal publishers, educational publishers, academic</td>
</tr>
<tr>
<td></td>
<td>publishers, commercial publishers</td>
</tr>
<tr>
<td>NFL Players</td>
<td>sports figures, football greats, football players, backs, quarterbacks</td>
</tr>
<tr>
<td>Scientific Journals</td>
<td>prestigious journals, peer-reviewed journals, refereed journals,</td>
</tr>
<tr>
<td></td>
<td>scholarly journals, academic journals</td>
</tr>
</tbody>
</table>
Semantic Constraints for Better Classes
Semantic Constraints for Better Classes

- **people-person-name**
  - Isaac Newton
- **film-music_contributor-name**
  - Johnny Cash
  - Bob Dylan
Semantic Constraints for Better Classes

Instances with shared attributes are likely to be from the same class.
Experiments with Public Sources

- Make Talukdar et al. results easy to reproduce and extend
- Freebase: multiple-sourced relational tables
- Pantel et al. 09 gold-standard hypernyms
- TextRunner (U. of Washington): hypernyms from open-domain extraction
- YAGO (Suchanek et al., 07): entity-attribute knowledge base curated from Wikipedia and Wordnet
Better Classes with YAGO Attributes
Better Classes with YAGO Attributes

170 WordNet Classes, 10 Seeds per Class, using Adsorption

- TextRunner Graph
- YAGO Graph
- TextRunner + YAGO Graph

Mean Reciprocal Rank (MRR)

0.45
0.413
0.375
0.338
0.3
Better Classes with YAGO Attributes

170 WordNet Classes, 10 Seeds per Class, using Adsorption

From TextRunner:
- 175k nodes, 529k edges

Graphs:
- TextRunner Graph
- YAGO Graph
- TextRunner + YAGO Graph

Mean Reciprocal Rank (MRR)

Friday, May 7, 2010
Better Classes with YAGO Attributes

170 WordNet Classes, 10 Seeds per Class, using Adsorption

Mean Reciprocal Rank (MRR)

TextRunner Graph
YAGO Graph
TextRunner + YAGO Graph

From YAGO: 142k nodes, 777k edges
Better Classes with YAGO Attributes

170 WordNet Classes, 10 Seeds per Class, using Adsorption

Mean Reciprocal Rank (MRR)

0.45

TextRunner Graph
YAGO Graph
TextRunner + YAGO Graph

Combined graph, 237k nodes, 1.3m edges

Friday, May 7, 2010
Better Classes with YAGO Attributes

170 WordNet Classes, 10 Seeds per Class, using Adsorption

Mean Reciprocal Rank (MRR)

TextRunner Graph
YAGO Graph
TextRunner + YAGO Graph
Classes for Attributes

- Qualitative evidence that attribute nodes propagate the right information

<table>
<thead>
<tr>
<th>YAGO Attribute</th>
<th>Top-2 WordNet Classes Assigned by MAD (example instances for each class are shown in brackets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>has_currency</td>
<td>wordnet_country_108544813 (Burma, Afghanistan) wordnet_region_108630039 (Aosta Valley, Southern Flinders Ranges)</td>
</tr>
<tr>
<td>works_at</td>
<td>wordnet_scientist_110560637 (Aage Niels Bohr, Adi Shamir) wordnet_person_100007846 (Catherine Cornelius, Jamie White)</td>
</tr>
<tr>
<td>has_capital</td>
<td>wordnet_state_108654360 (Agusan del Norte, Bali) wordnet_region_108630039 (Aosta Valley, Southern Flinders Ranges)</td>
</tr>
<tr>
<td>born_in</td>
<td>wordnet_boxer_109870208 (George Chuvalo, Fernando Montiel) wordnet_chancellor_109906986 (Godon Brown, Bill Bryson)</td>
</tr>
<tr>
<td>has_isbn</td>
<td>wordnet_book_106410904 (Past Imperfect, Berlin Diary) wordnet_magazine_106595351 (Railway Age, Investors Chronicle)</td>
</tr>
</tbody>
</table>
Propagation Objective

- MAD [Talukdar & Crammer 09] (simplified)

\[
\arg\min_{\hat{Y}} \sum_{l=1}^{m+1} \left[ \|S\hat{Y}_l - SY_l\|^2 + \mu_1 \sum_{u,v} M_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2 + \mu_2 \|\hat{Y}_l - R_l\|^2 \right]
\]

- \(m\) labels, +1 dummy label
- \(M = W^T + W\) is the symmetrized weight matrix
- \(\hat{Y}_{vl}\): weight of label \(l\) on node \(v\)
- \(Y_{vl}\): seed weight for label \(l\) on node \(v\)
- \(S\): diagonal matrix, nonzero for seed nodes
- \(R_{vl}\): regularization target for label \(l\) on node \(v\)
A Propagation Algorithm

Inputs $Y, R : |V| \times (|L| + 1), W : |V| \times |V|, S : |V| \times |V|$ diagonal

$\hat{Y} \leftarrow Y$

$M = W + W^\top$

$Z_v \leftarrow S_{vv} + \mu_1 \sum_{u \neq v} M_{vu} + \mu_2 \quad \forall v \in V$

repeat

for all $v \in V$ do

$\hat{Y}_v \leftarrow \frac{1}{Z_v} \left( (SY)_v + \mu_1 M_v \cdot \hat{Y} + \mu_2 R_v \right)$

end for

until convergence

• Some details of the construction of the input matrices omitted for simplicity

• Converges under reasonable assumptions

• Many variants, alternative objectives (Subramanya and Bilmes 2008)
Good News

• Label propagation can combine multiple information sources effectively

• Useful coverage of class-instance relations, much bigger than in previous work

• Embarrassingly parallel algorithms

• Graph representation can encode a variety of useful constraints
Limitations

• Can’t express “few classes per instance”
• How to classify instances in context?
  ‣ Whistler paintings vs Whistler skiers
• Propagation is additive, averaging
• Algorithmically nice (convexity, convergence)
• But it can’t “push back” to express incompatibilities between classes
  ‣ artist ⊃ painter ⊄ ski-resort
Few Classes Per Instance

- If a classification is to be informative, it must have limited ambiguity
- **Posterior regularization**: constrain the ambiguity of final labeling
- POS tagging pilot (Graça et al. 09): motivated by this work, but easier to test
- Penalize the $\ell_1/\ell_\infty$ norm of the posterior distribution of classes given instances
Contextual Classification

- Graph-based approach:
  - One node per mention (lots of nodes!)
  - Link mention nodes that have similar contexts

- Language modeling approach:
  - Class labels are also terms
  - Model probability of class given context
  - Find most likely class for each mention given its context
• Where do edge weights come from?
• These experiments: heuristic scoring functions drawn from language modeling and information retrieval
• Can we learn edge scoring functions?
  • Minkov & Cohen: learning from random walks
  • McCallum’s group: learning factor potentials by M-H sampling
• Learn from user feedback (Talukdar et al., SIGMOD 2010)
Summary

• First steps in inferring broad-coverage semantic relationships from the actions of Web users:
  • What they write
  • How they interact with search results
• Multiple correlated sources provide a wealth of indirect supervision
• (Some) graph-based algorithms scale up
• Related work:
  • Wang & Cohen’s SEAL and follow-ups
  • Weikum lab’s SOFIE
Current Work

- Web-scale contextual semantic annotation
  - is-a, co-reference
- Combining multiple relationships
  - The distinct senses of “Whistler“ belong to disjoint co-reference classes
- Probabilistic interpretations
  - Relational factor graphs
  - Non-parametric Bayesian models