Domain Adaptation with Structural Correspondence Learning

John Blitzer

Joint work with

Shai Ben-David, Koby Crammer, Mark Dredze, Ryan McDonald, Fernando Pereira
Statistical models, multiple domains
Different Domains of Text

- Huge variation in vocabulary & style

"Ok, I’ll just build models for each domain I encounter"
Sentiment Classification for Product Reviews

Product Review
Classifier → SVM, Naïve Bayes, etc.
Positive
Negative

Multiple Domains
books
kitchen appliances
??
??
??
## books & kitchen appliances

<table>
<thead>
<tr>
<th>Running with Scissors: A Memoir</th>
<th>Avante Deep Fryer, Chrome &amp; Black</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title:</strong> Horrible book, horrible.</td>
<td><strong>Title:</strong> lid does not work well...</td>
</tr>
<tr>
<td>This book was horrible. I read half of it, suffering from a headache the entire time,</td>
<td>I love the way the Tefal deep fryer cooks, however, I am returning my</td>
</tr>
<tr>
<td>and eventually I lit it on fire. One less copy in the world... don't waste your money. I wish I had the time spent reading this book back so I could use it for better purposes. This book wasted my life</td>
<td>second one due to a defective lid closure. The lid may close initially, but after a few uses it no longer stays closed. I will not be purchasing this one again.</td>
</tr>
</tbody>
</table>

**Error increase:** 13% → 26%
The clash is a sign of a new toughness and divisiveness in Japan’s once-cozy financial circles.

The oncogenic mutated forms of the ras proteins are constitutively active and interfere with normal signal transduction.
Features & Linear Models

Problem: If we’ve only trained on book reviews, then \(w(\text{defective}) = 0\)
Structural Correspondence Learning (SCL)

- Cut adaptation error by more than 40%
- Use *unlabeled* data from the target domain
- Induce correspondences among different features
- *read-half, headache* $\leftrightarrow$ *defective, returned*
- Labeled data for *source* domain will help us build a good classifier for *target* domain

Maximum likelihood linear regression (MLLR) for speaker adaptation ([Leggetter & Woodland, 1995](#))
Step 1: Unlabeled – Learn correspondence mapping

\[ \Phi(X_S) \]

\[ \begin{pmatrix} 0.1 \\ 0 \\ \vdots \\ 0 \\ 0.3 \end{pmatrix} \]

\[ \Phi(x) \]

\[ \begin{pmatrix} 0.3 \\ -1.0 \\ \vdots \\ 0.7 \\ -2.1 \end{pmatrix} \]

Step 2: Labeled – Learn weight vector

\[ \Phi(x) \rightarrow \text{sgn}(v \cdot \Phi(x)) \]

- \( \Phi \) should make the domains look as similar as possible
- But \( \Phi \) should also allow us to classify well
**SCL: Making Domains Look Similar**

Incorrect classification of kitchen review: **defective lid**

<table>
<thead>
<tr>
<th>Unlabeled <em>kitchen</em> contexts</th>
<th>Unlabeled <em>books</em> contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Do <strong>not buy</strong> the Shark portable steamer .... Trigger mechanism is <strong>defective</strong>.</td>
<td>• The book is so <strong>repetitive</strong> that I found myself yelling .... I will definitely <strong>not buy</strong> another.</td>
</tr>
<tr>
<td>• the very nice lady assured me that I must have a <strong>defective</strong> set .... What a <strong>disappointment</strong>!</td>
<td>• A <strong>disappointment</strong> .... Ender was talked about for &lt;#&gt; <strong>pages</strong> altogether.</td>
</tr>
<tr>
<td>• Maybe mine was <strong>defective</strong> .... The directions were <strong>unclear</strong></td>
<td>• it’s <strong>unclear</strong> .... It’s repetitive and <strong>boring</strong></td>
</tr>
</tbody>
</table>
SCL: Pivot Features

**Pivot Features**

- Occur frequently in both domains
- Characterize the task we want to do
- Number in the hundreds or thousands
- Choose using labeled **source**, unlabeled **source** & **target** data

### SCL: words & bigrams that occur frequently in both domains

<table>
<thead>
<tr>
<th>book</th>
<th>one</th>
<th>&lt;num&gt;</th>
<th>so</th>
<th>all</th>
<th>very</th>
<th>about</th>
<th>they</th>
<th>like</th>
<th>good</th>
<th>when</th>
</tr>
</thead>
</table>

### SCL-MI: SCL but also based on mutual information with labels

| a_must | a_wonderful | loved_it | weak | don’t_waste | awful | highly_recommended | and_easy |
SCL Unlabeled Step: Pivot Predictors

Use **pivot features** to align other features

(1) The book is so **repetitive** that I found myself yelling .... I will definitely **another**.

(2) Do **the Shark portable steamer** .... Trigger mechanism is **defective**.

**Binary problem:** Does “not buy” appear here?

- **Mask** and predict pivot features using other features
- Train N **linear predictors**, one for each binary problem
- Each pivot predictor implicitly aligns non-pivot features from **source & target** domains
SCL: Dimensionality Reduction

\[ \begin{bmatrix}
  \mathbf{w}_1 \\
  \vdots \\
  \mathbf{w}_i \\
  \vdots \\
  \mathbf{w}_N 
\end{bmatrix} \]

- \( \mathbf{W}^T \mathbf{x} \) gives \( N \) new features
- Value of \( i^{th} \) feature is the propensity to see “not buy” in the same document

- We still want fewer new features (1000 is too many)
- Many pivot predictors give similar information
  - “horrible”, “terrible”, “awful”
- Compute SVD & use top left singular vectors \( \Phi \)

Latent Semantic Indexing (LSI), (Deerwester et al. 1990)
Latent Dirichlet Allocation (LDA), (Blei et al. 2003)
Back to Linear Classifiers

\[
\text{Classifier} \quad \text{sgn} \left[ w \cdot x + v \cdot \Phi^T x \right]
\]

- **Source** training: Learn \( w \) & \( v \) together

- **Target** testing: First apply \( \Phi \), then apply \( w \) and \( v \)
Inspirations for SCL

1. **Alternating Structural Optimization (ASO)**
   - **Ando & Zhang** (JMLR 2005)
   - Inducing structures for semi-supervised learning

2. **Correspondence Dimensionality Reduction**
   - **Verbeek, Roweis, & Vlassis** (NIPS 2003).
   - **Ham, Lee, & Saul** (AISTATS 2003).
   - Learn a low-dimensional representation from high-dimensional correspondences
Sentiment Classification Data

- **Product reviews from Amazon.com**
  - Books, DVDs, Kitchen Appliances, Electronics
  - 2000 labeled reviews from each domain
  - 3000 – 6000 unlabeled reviews

- **Binary classification problem**
  - Positive if 4 stars or more, negative if 2 or less

- **Features**: unigrams & bigrams

- **Pivots**: SCL & SCL-MI

- **At train time**: minimize Huberized hinge loss (Zhang, 2004)
Visualizing $\Phi$ (books & kitchen)

**negative** vs. **positive**

**books**
- plot
- $<$#$_pages$
- predictable
- fascinating
- engaging
- must_read
- grisham
- poorly_designed
- awkward_to
- espresso
- years_now
- are_perfect
- a_breeze

**kitchen**
- the_plastic
- leaking
Empirical Results: books & DVDs

Baseline loss due to adaptation: 7.6%
SCL-MI loss due to adaptation: 0.7%
Empirical Results: electronics & kitchen
Empirical Results: books & DVDs

- Sometimes SCL can cause increases in error
- With only unlabeled data, we misalign features
Using Labeled Data

50 instances of labeled target domain data

Source data, save weight vector for SCL features $V_S$

Target data, regularize weight vector to be close to $V_S$

Chelba & Acero, EMNLP 2004

$\lambda \|w\|^2 + \mu \|v - V_S\|^2$

Huberized hinge loss

Keep SCL weights close to source weights

Avoid using high-dimensional features
Empirical Results: labeled data

- With 50 labeled target instances, SCL-MI **always** improves over baseline
### Average Improvements

<table>
<thead>
<tr>
<th>model</th>
<th>base</th>
<th>base + targ</th>
<th>scl</th>
<th>scl-mi</th>
<th>scl-mi + targ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg Adaptation Loss</td>
<td>9.1</td>
<td>9.1</td>
<td>7.1</td>
<td>5.8</td>
<td>4.9</td>
</tr>
</tbody>
</table>

- scl-mi reduces error due to transfer by 36%
- adding 50 instances [Chelba & Acero 2004] without SCL does not help
- scl-mi + targ reduces error due to transfer by 46%
PoS Tagging: Data & Model

- **Data**
  - 40k Wall Street Journal (WSJ) training sentences
  - 100k unlabeled biomedical sentences
  - 100k unlabeled WSJ sentences

- **Supervised Learner**
  - MIRA CRF: Online max-margin learner
  - Separate correct label from top k=5 incorrect labels
  - Crammer et al. JMLR 2006

- **Pivots:** Common left/middle/right words
Visualizing PoS Tagging

nouns vs. adjs & dets

MEDLINE

receptors
mutation
lesions
assays
metastatic
neuronal
transient
functional

company
investors
transaction
officials
political
short-term
pretty
your

Wall Street Journal
Empirical Results

561 MEDLINE test sentences

<table>
<thead>
<tr>
<th>Model</th>
<th>All Words</th>
<th>Unk words</th>
</tr>
</thead>
<tbody>
<tr>
<td>MXPOST</td>
<td>87.2</td>
<td>65.2</td>
</tr>
<tr>
<td>super</td>
<td>87.9</td>
<td>68.4</td>
</tr>
<tr>
<td>semi-ASO</td>
<td>88.4</td>
<td>70.9</td>
</tr>
<tr>
<td>SCL</td>
<td>88.9</td>
<td>72.0</td>
</tr>
</tbody>
</table>

Null Hyp | p-value  |
---------|----------|
semi vs. super | <0.0015  |
SCL vs. super   | <10^{-12}|
SCL vs. semi    | <0.0003  |
Results: Some labeled target domain data

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1k-SCL</td>
<td>95.0</td>
</tr>
<tr>
<td>1k-super</td>
<td>94.5</td>
</tr>
<tr>
<td>Nosource</td>
<td>94.5</td>
</tr>
</tbody>
</table>

• Use source tagger output as a feature (Florian et al. 2004)
• Compare SCL with supervised source tagger

561 MEDLINE test sentences

Accuracy

# of MEDLINE training sentences

notarget
nosource
1k-super
1k-SCL
Adaptation & Machine Translation

- **Source:** Domain specific parallel corpora (news, legal text)
- **Target:** Similar corpora from the web (i.e. blogs)
- **Learn translation rules / language model parameters for the new domain**
- **Pivots:** common contexts
Adaptation & Ranking

- **Input:** query & list of top-ranked documents
- **Output:** Ranking
- Score documents based on editorial or click-through data
- **Adaptation:** Different markets or query types
- **Pivots:** common relevant features
Bounds on the error of models in new domains

Analysis of Representations for Domain Adaptation.
Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira.
NIPS 2006.

Learning Bounds for Domain Adaptation.
John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, Jenn Wortman.
NIPS 2007 (To Appear).
Pipeline Adaptation: Tagging & Parsing

**Dependency Parsing**

- McDonald et al. 2005
- Uses part of speech tags as features
- Train on WSJ, test on MEDLINE
- Use different taggers for MEDLINE input features

### Accuracy for different tagger inputs

- **Supervised**
- **SCL**
- **Gold**

---

**# of WSJ training sentences**

- 100
- 500
- 1k
- 5k
- 40k

- **Accuracy**

- 58
- 62
- 66
- 70
- 74
- 78
- 82
Measuring Adaptability

• Given limited resources, which domains should we label?

• Idea: Train a classifier to distinguish instances from different domains

• Error of this classifier is an estimate of loss due to adaptation
Suppose we can afford to label 2 domains
Then we should label 1 of electronics/kitchen and 1 of books/DVDs
Problem: If we've only trained on financial news, then $w(RW=\text{transduction}) = 0$
Future Work

• SCL for other problems & modalities
  – named entity recognition
  – vision (aligning SIFT features)
  – speaker / acoustic environment adaptation

• Learning low-dimensional representations for multi-part prediction problems
  – natural language parsing, machine translation, sentence compression