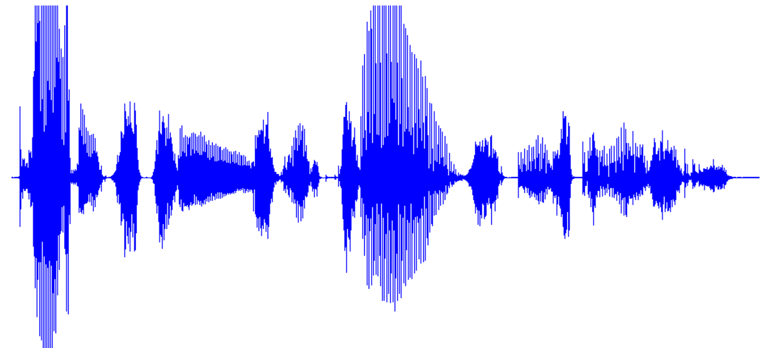
Domain Adaptation with Structural Correspondence Learning

John Blitzer

Joint work with

Shai Ben-David, Koby Crammer, Mark Dredze, Ryan McDonald, Fernando Pereira

Statistical models, multiple domains

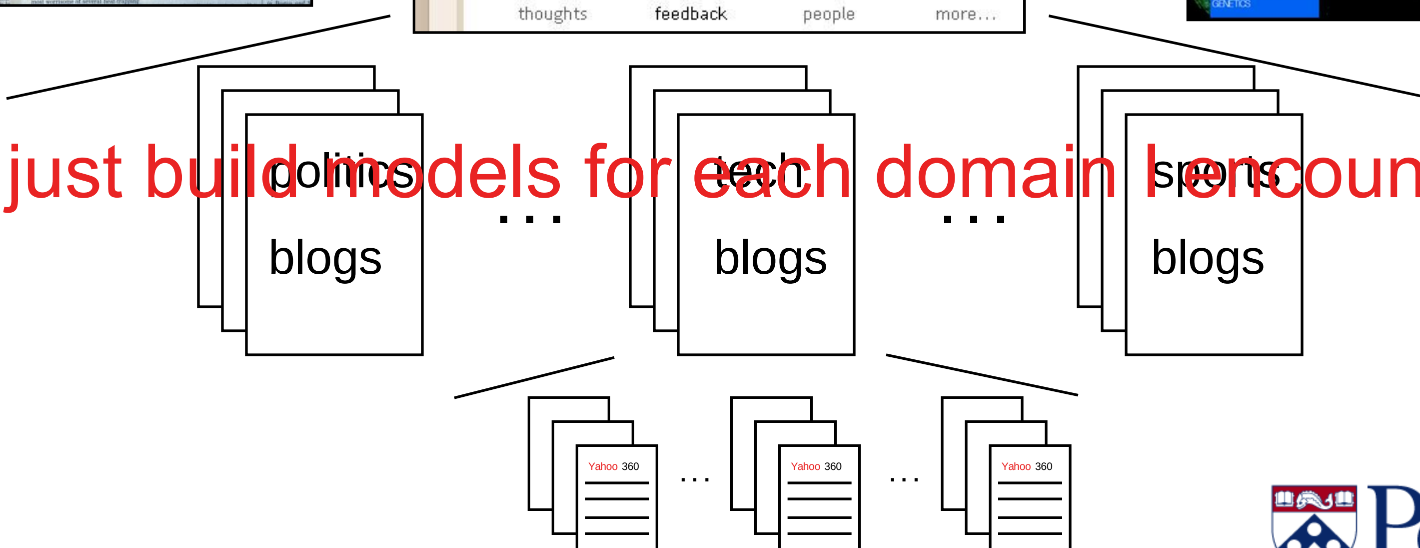


Different Domains of Text

- Huge variation in vocabulary & style

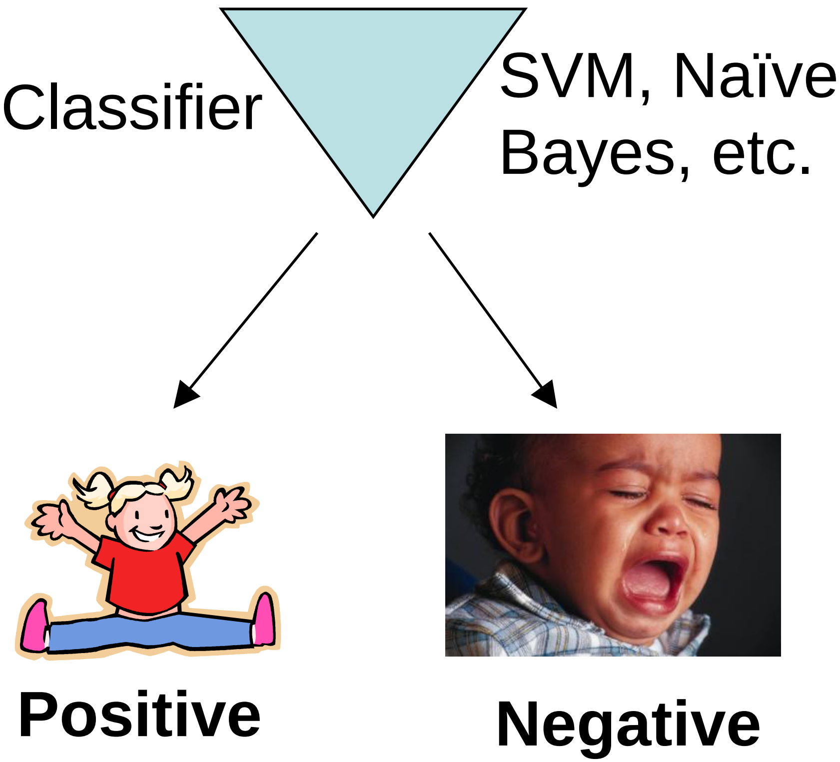

Cracking the Neanderthal Code
Our closest cousin's DNA may unravel secrets of evolution
MARKETPLACE | A9
A Mystery Multiplies Over a Math Prize
Why isn't a reclusive Russian pursuing a $1 million award?
SCIENCE JOURNAL | A9
THE WALL STREET JOURNAL.
WSJ.com
FRIDAY, JULY 21, 2006
What's News—
Business and Finance
World-Wide
Burning Debate
As Emission Restrictions Loom, Texas Utility Bets Big on Coal
TXU to Build 11 New Plants As Rivals Try Alternatives; Global Warming Is Concern
Mr. Wilder Cites Demand
Power Source

"Ok, I'll just build models for each domain I encounter"

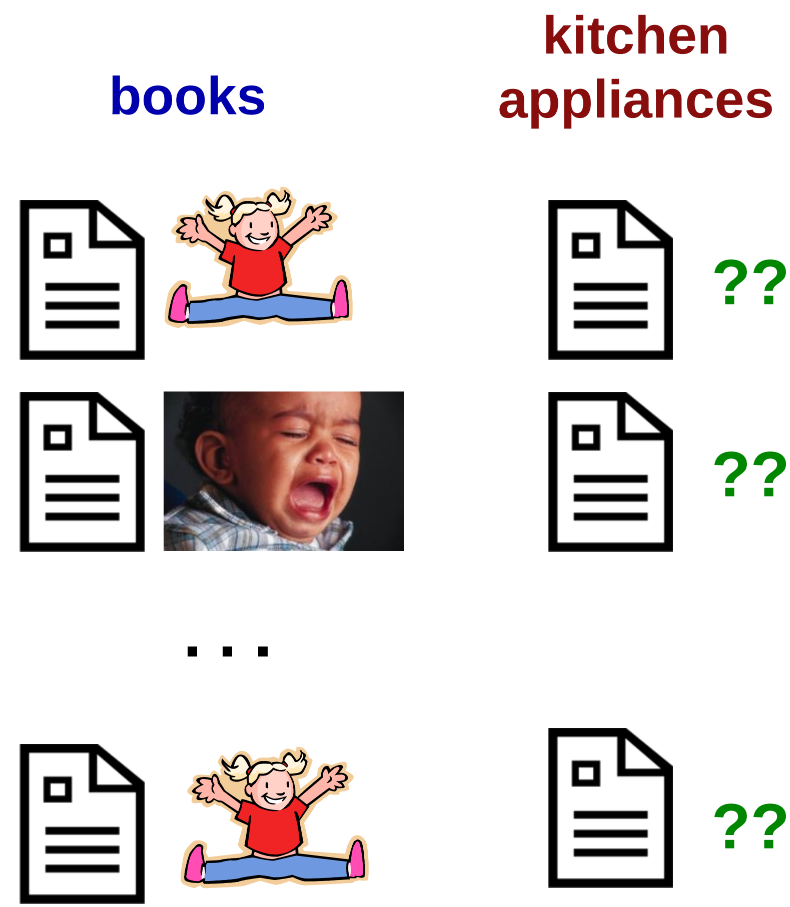


Sentiment Classification for Product Reviews

Product Review



Multiple Domains



books & kitchen appliances

Running with Scissors: A Memoir

Title: Horrible book, horrible.

This book was horrible. I **read half** of it, **suffering from a headache** the entire time,

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

Avante Deep Fryer, Chrome & Black

Title: lid does not work well...

I love the way the Tefal deep fryer cooks, however, I am **returning** my

closure. The lid may close initially, but after a few uses it no longer stays closed. I **will not be purchasing** this one again.

Error increase: 13% → 26%

Part of Speech Tagging

Wall Street Journal (WSJ)

DT	NN	VBZ	DT	NN	IN	DT	JJ	NN	CC
The	**clash**	is	a	sign	of	a	new	**toughness**	and

NN	IN	NNP	POS	JJ	JJ	NNS	.
divisiveness	in	Japan	's	**once-cozy**	**financial**	circles	.

Error increase: 3% → 12%

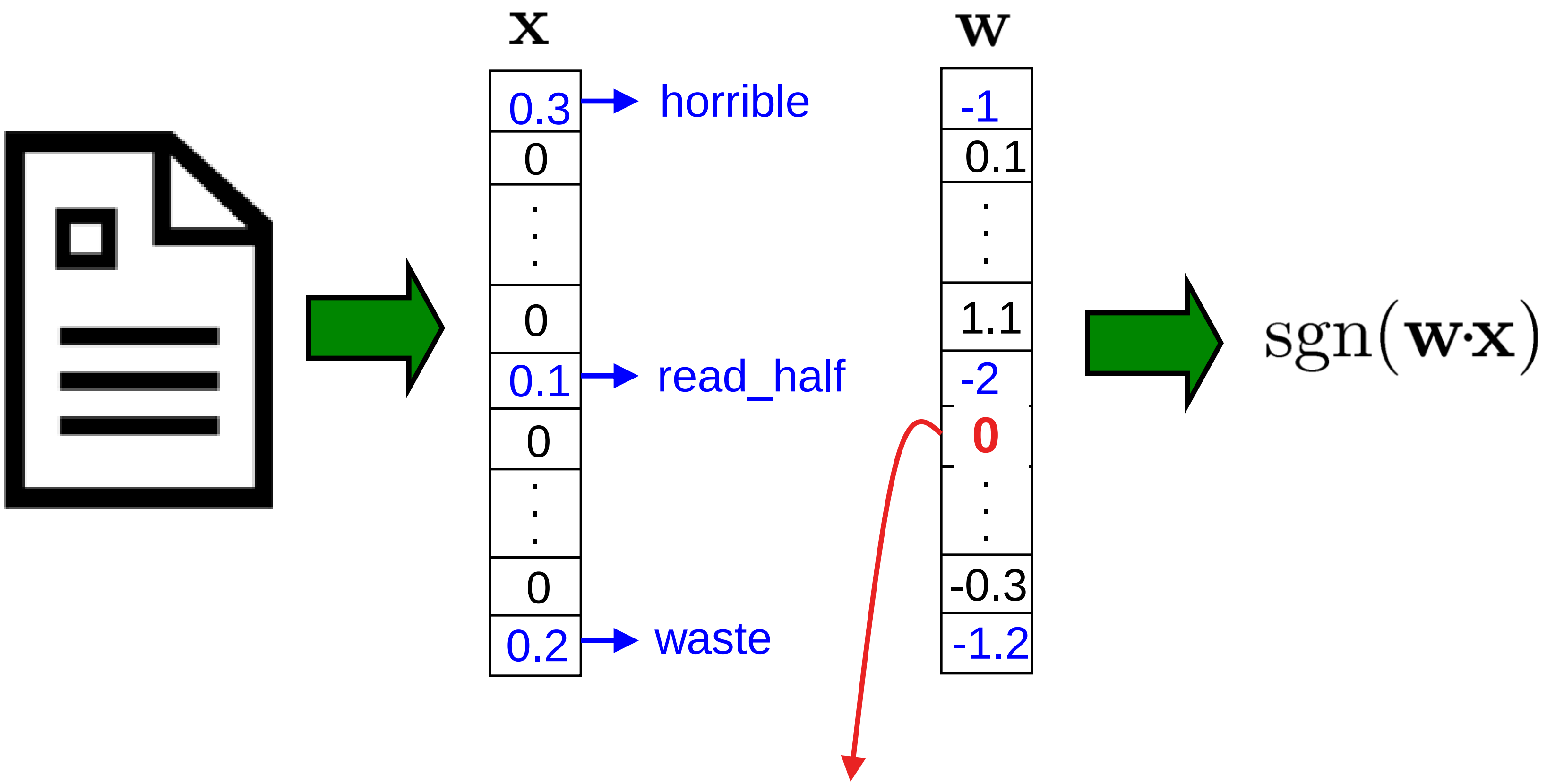
MEDLINE Abstracts (biomed)

DT	JJ	VBN	NNS	IN	DT	NN	NNS	VBP
The	**oncogenic**	mutated	forms	of	the	**ras**	proteins	are

RB	JJ	CC	VBP	IN	JJ	NN
constitutively	active	and	interfere	with	normal	signal

NN	.
transduction	.

Features & Linear Models



Problem: If we've only trained on book reviews, then **w(defective) = 0**

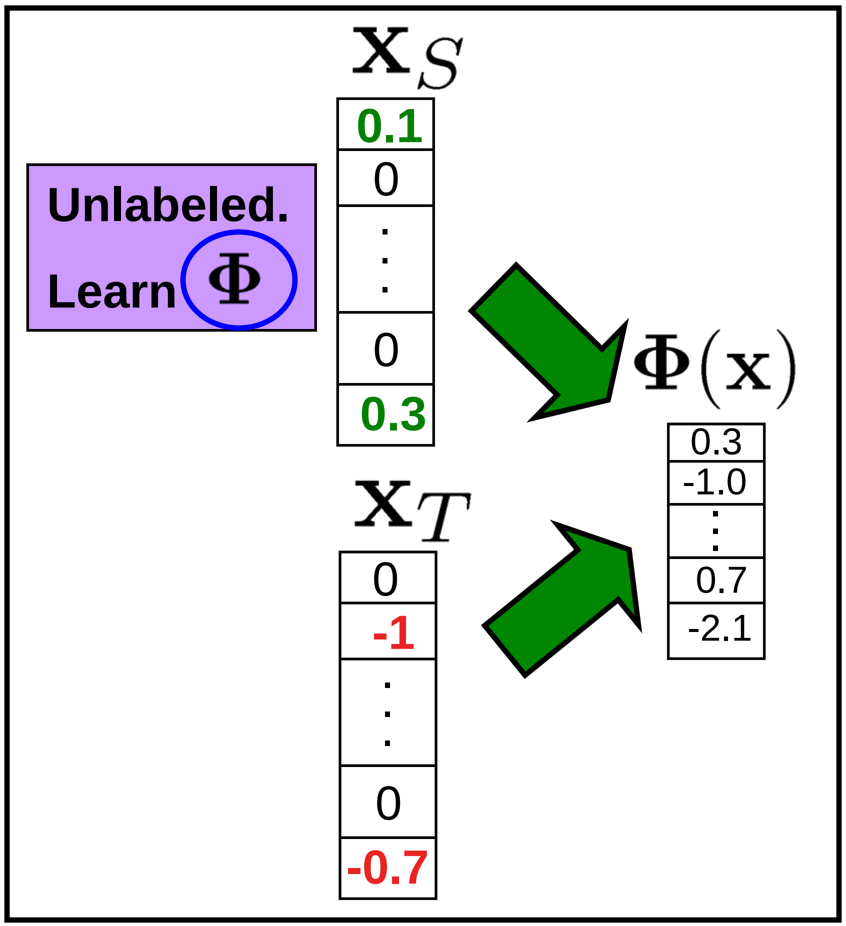
Structural Correspondence Learning (SCL)

- **Cut adaptation error by more than 40%**
- Use **unlabeled** data from the target domain
- Induce correspondences among different features
- **read-half, headache** ⟷ **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)

SCL: 2-Step Learning Process

Step 1: Unlabeled – Learn correspondence mapping



Step 2: Labeled – Learn weight vector

Labeled. Learn $\mathbf{v}$

$$\mathbf{\Phi}(\mathbf{x}) \Rightarrow \mathrm{sgn}\left(\mathbf{v} \cdot \mathbf{\Phi}(\mathbf{x})\right)$$

- **$\mathbf{\Phi}$ should make the domains look as similar as possible**
- **But $\mathbf{\Phi}$ should also allow us to classify well**

Penn
University of Pennsylvania

SCL: Making Domains Look Similar

Incorrect classification of kitchen review **defective** lid

Unlabeled **kitchen** contexts	Unlabeled **books** contexts
• Do **not buy** the Shark portable steamer …. Trigger mechanism is **defective**.	• The book is so **repetitive** that I found myself yelling …. I will definitely **not buy** another.
• the very nice lady assured me that I must have a **defective** set …. What a **disappointment**!	• A **disappointment** …. Ender was talked about for **<#> pages** altogether.
• Maybe mine was **defective** …. The directions were **unclear**	• it's **unclear** …. It's repetitive and **boring**

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

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

book one <num> so all very about they like good when	**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

Penn
University of Pennsylvania

SCL: Dimensionality Reduction

$$\begin{bmatrix} | & & | & & | \\ \mathbf{w_1} & \cdots & \mathbf{w_i} & \cdots & \mathbf{w_N} \\ | & & | & & | \end{bmatrix}$$

- $\mathbf{W}^T\mathbf{x}$ 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** Φ

Latent Semantic Indexing (LSI), (Deerwester et al. 1990)

Latent Dirichlet Allocation (LDA), (Blei et al. 2003)

Back to Linear Classifiers

$$\mathbf{x} = \begin{bmatrix} 0.3 \\ 0 \\ \vdots \\ 0 \\ 0.1 \end{bmatrix}$$

Classifier $\operatorname{sgn}\left[\mathbf{w}\cdot\mathbf{x} + \mathbf{v}\cdot\mathbf{\Phi}^T\mathbf{x}\right]$

- **Source** training: Learn $\mathbf{w}$ & $\mathbf{v}$ together

$$\mathbf{\Phi}^T\mathbf{x} = \begin{bmatrix} 0.3 \\ -1.0 \\ \vdots \\ 0.7 \\ -2.1 \end{bmatrix}$$

- **Target** testing: First apply $\mathbf{\Phi}$, then apply $\mathbf{w}$ and $\mathbf{v}$

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

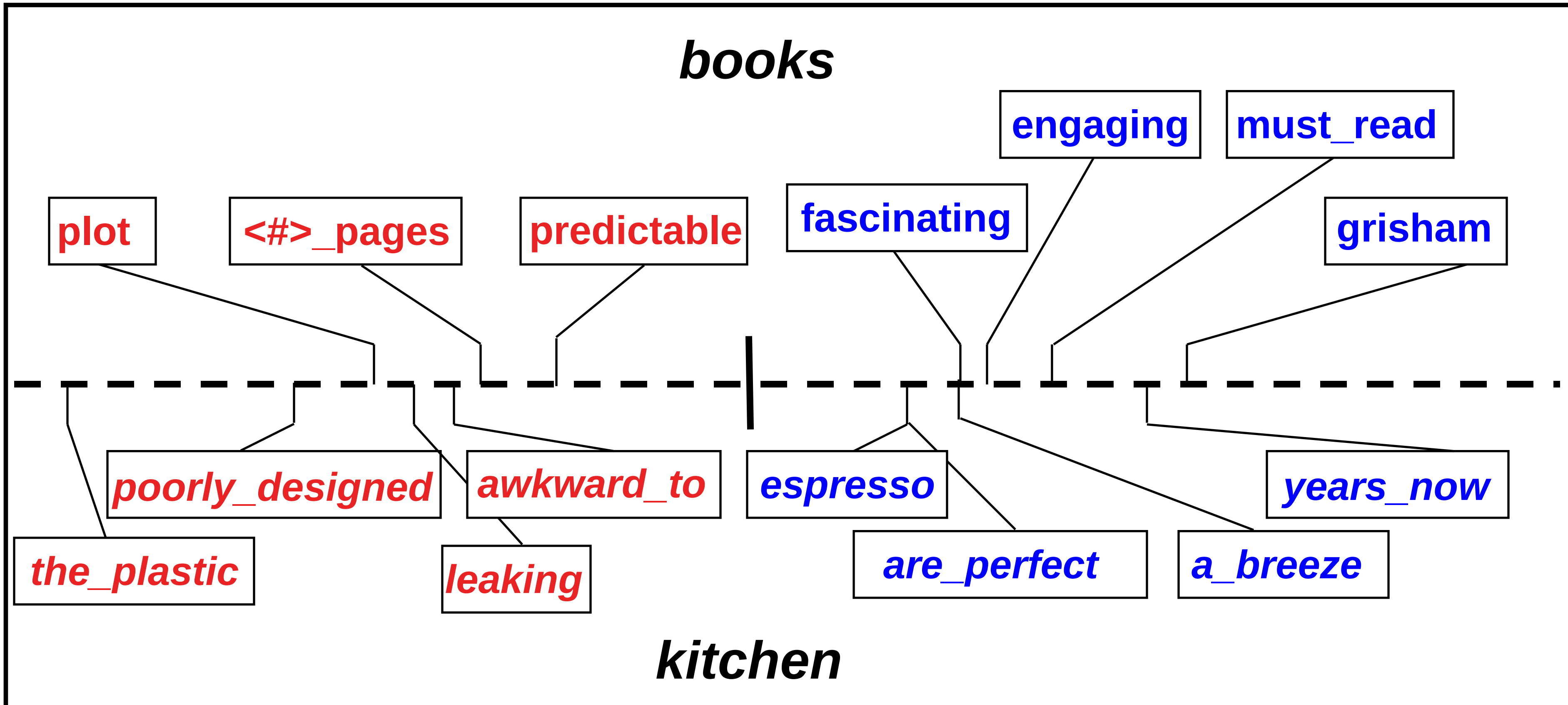
Sentiment Classification Data

- **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)

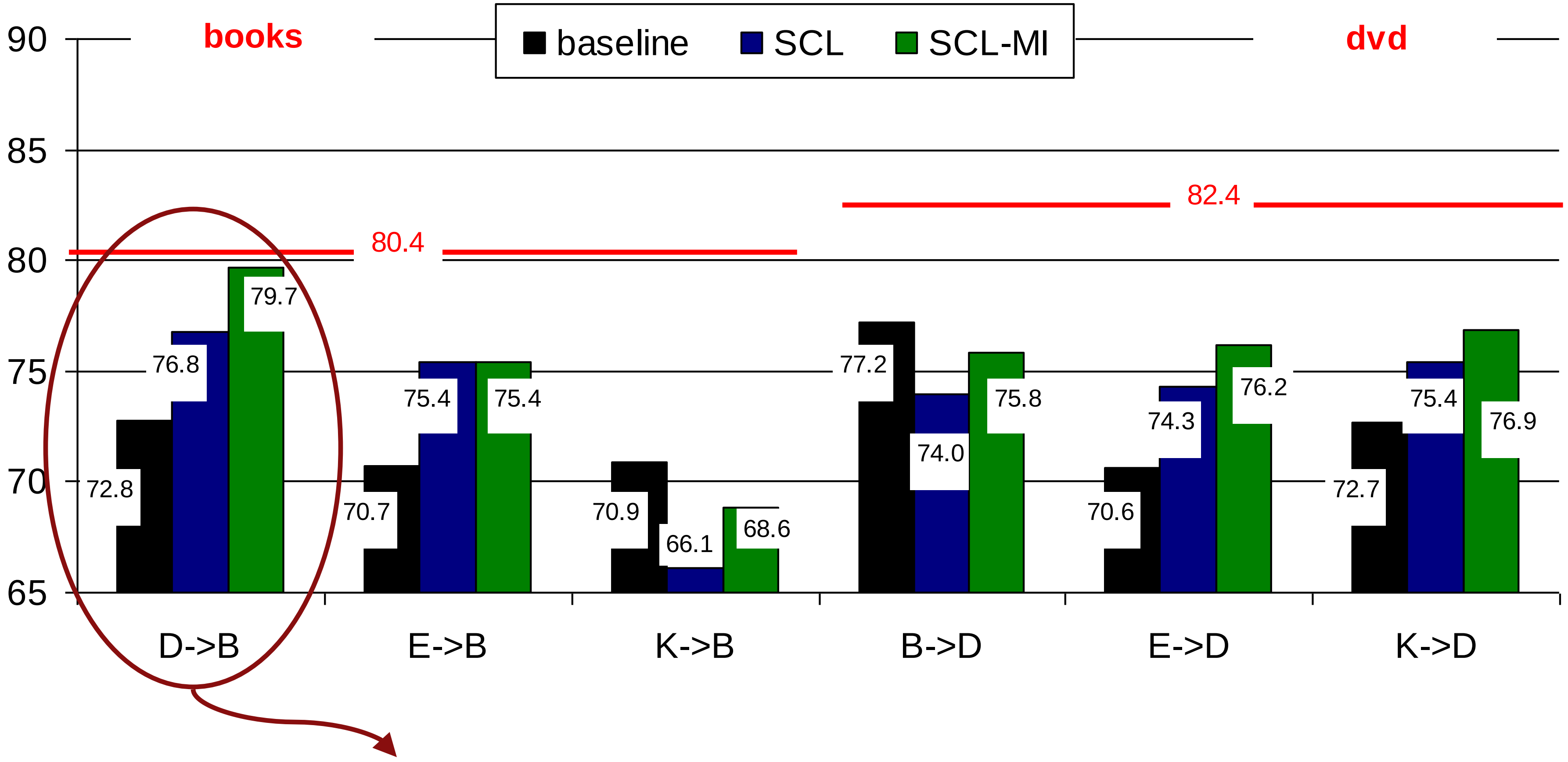
Penn
University of Pennsylvania

Visualizing Φ (books & kitchen)

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



Empirical Results: books & DVDs

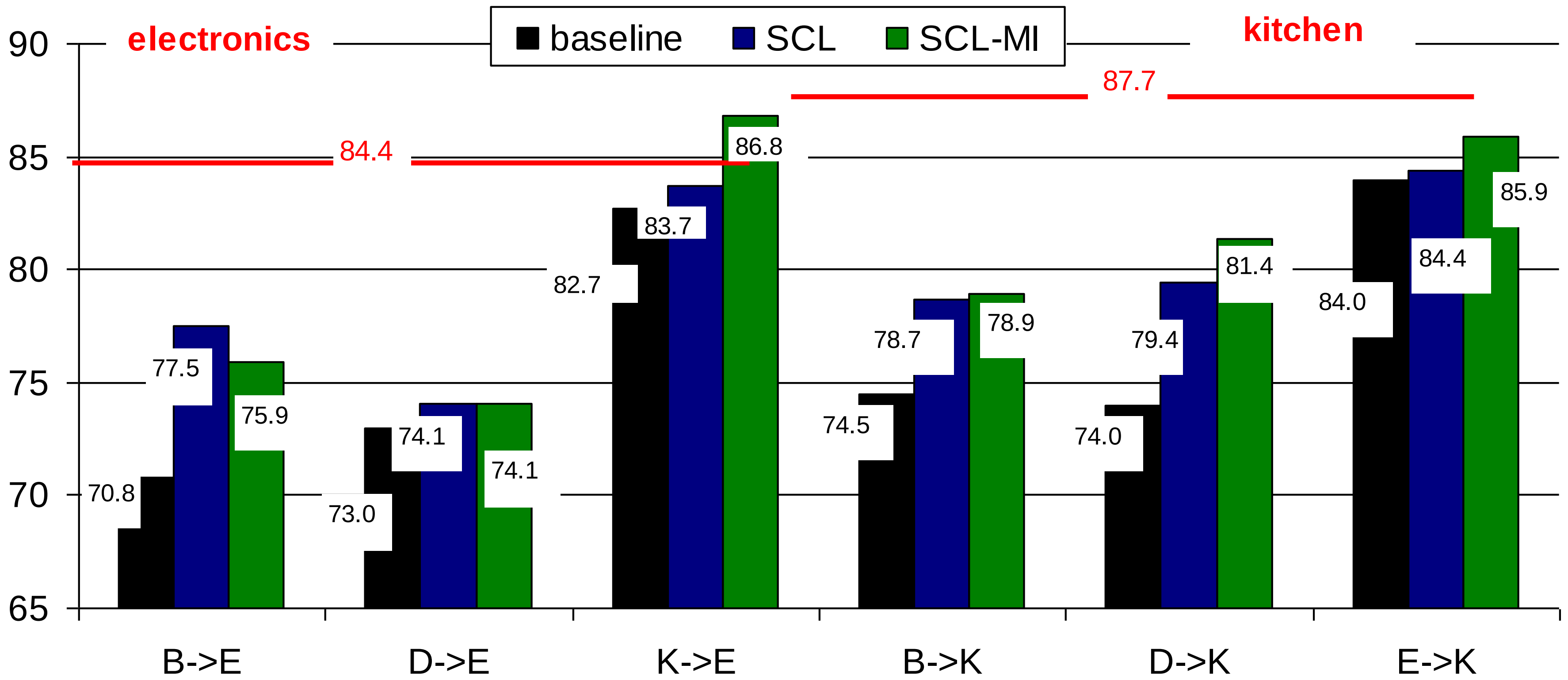


baseline loss due to adaptation: 7.6%

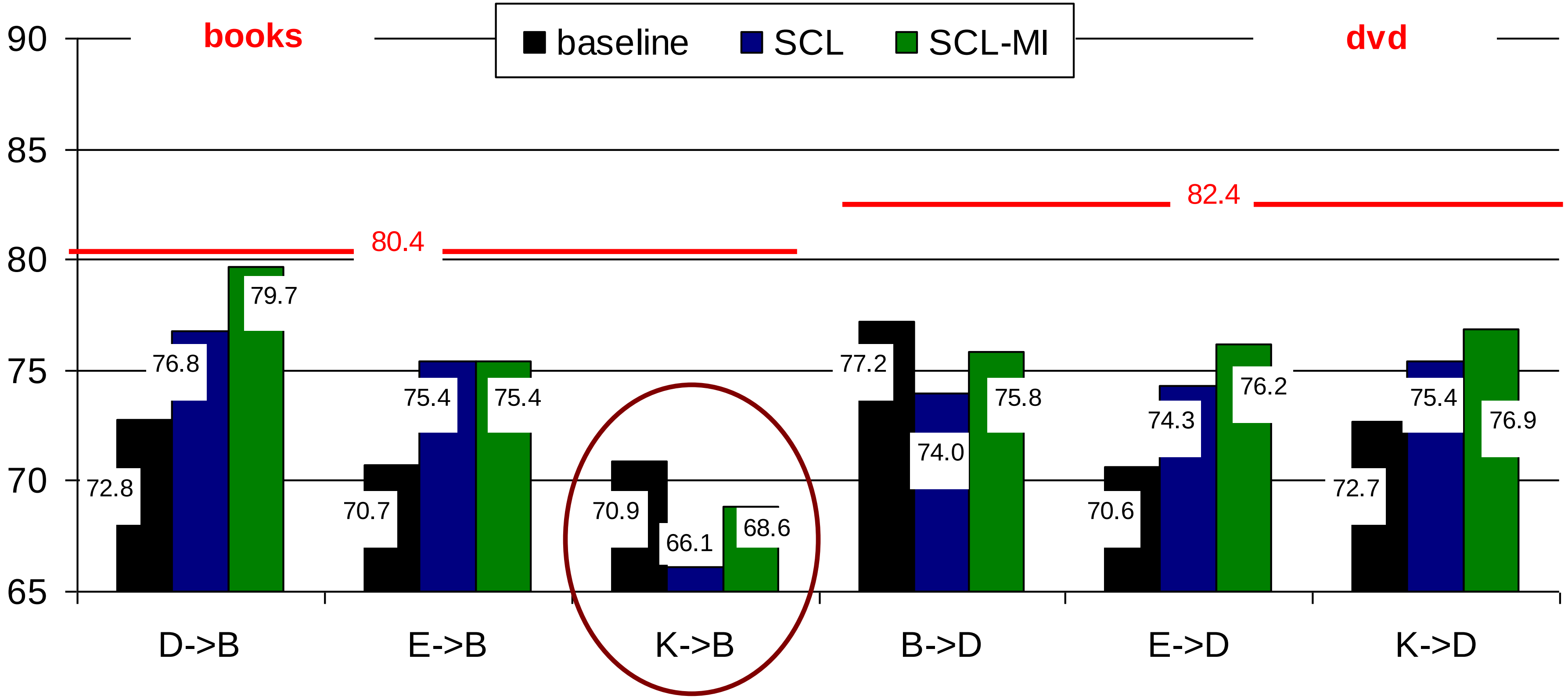
SCL-MI loss due to adaptation: 0.7%

Penn
University of Pennsylvania

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 $\mathbf{v}_s$

Target data, regularize weight vector to be close to $\mathbf{v}_s$

Chelba & Acero, EMNLP 2004

$$\min_{\mathbf{w},\mathbf{v}} \sum_j L(\ldots \mathbf{x}_j \ldots \mathbf{x}_j, y_j) +$$

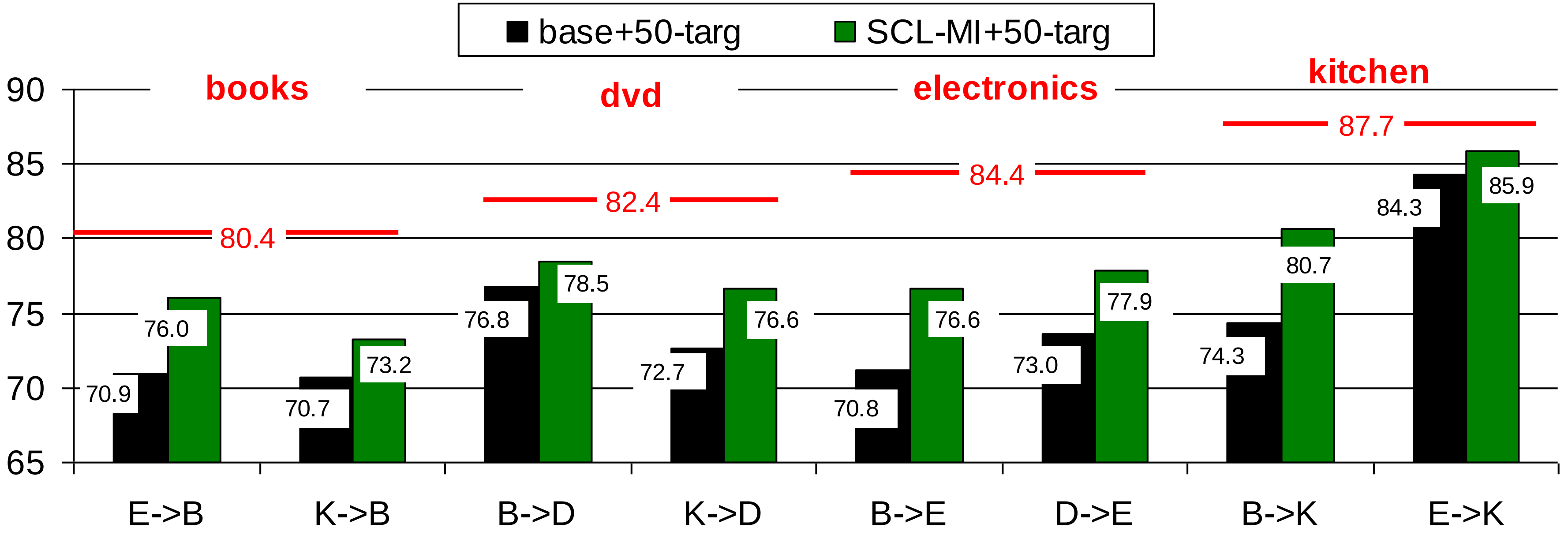
$$\lambda||\mathbf{w}||^2 + \mu||\mathbf{v} - \mathbf{v}_s||^2$$

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

model	base	base +targ	scl	scl-mi	scl-mi +targ
Avg Adaptation Loss	9.1	9.1	7.1	5.8	**4.9**

- **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%**

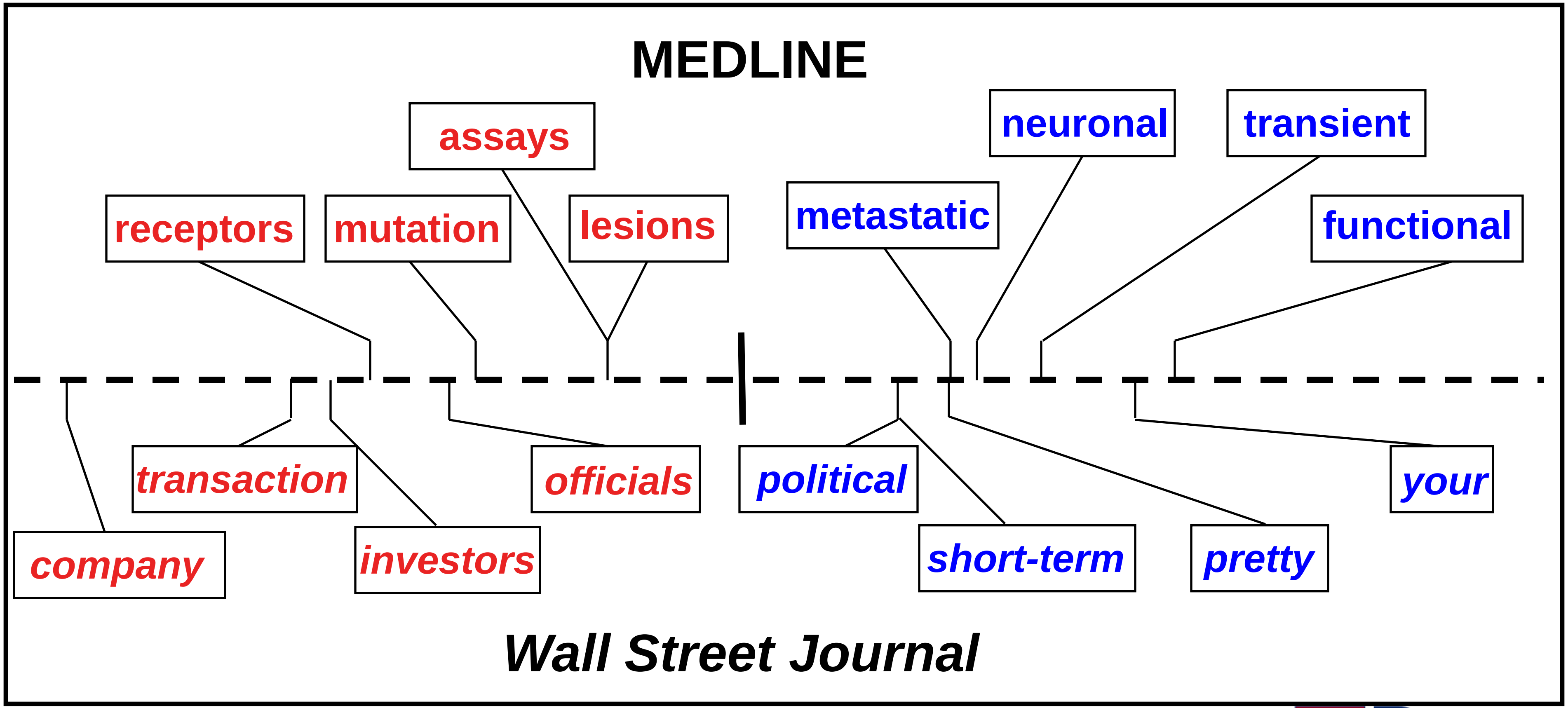
PoS Tagging: Data & Model

- **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

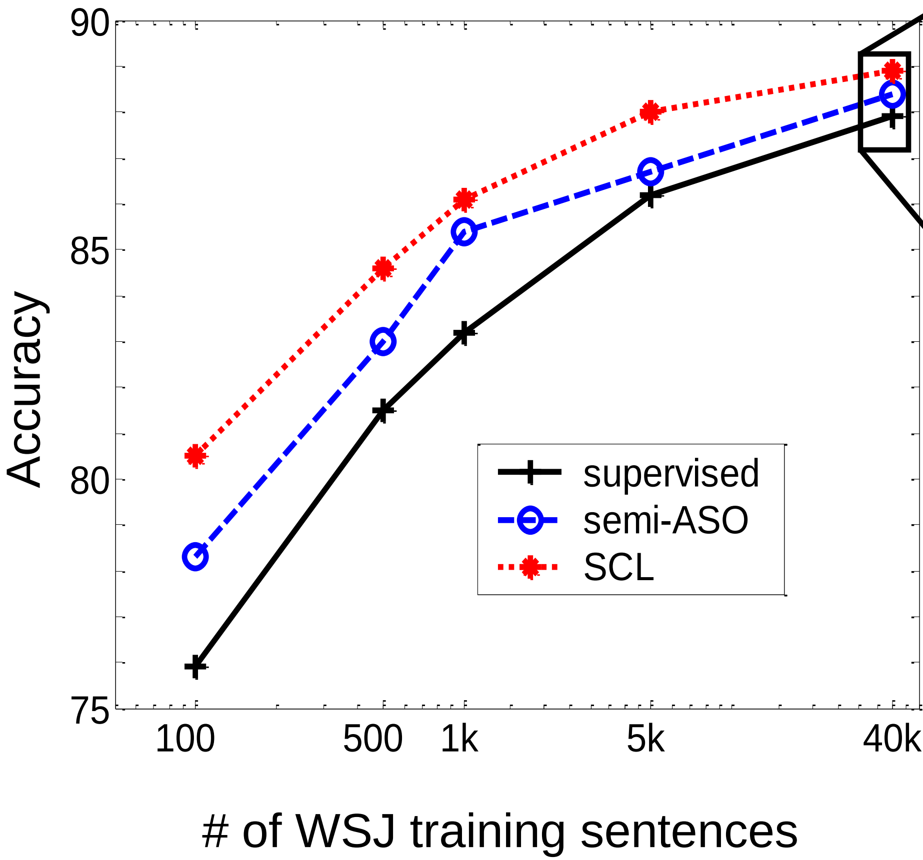
Visualizing Φ PoS Tagging

nouns vs. **adjs & dets**



Empirical Results

561 MEDLINE test sentences

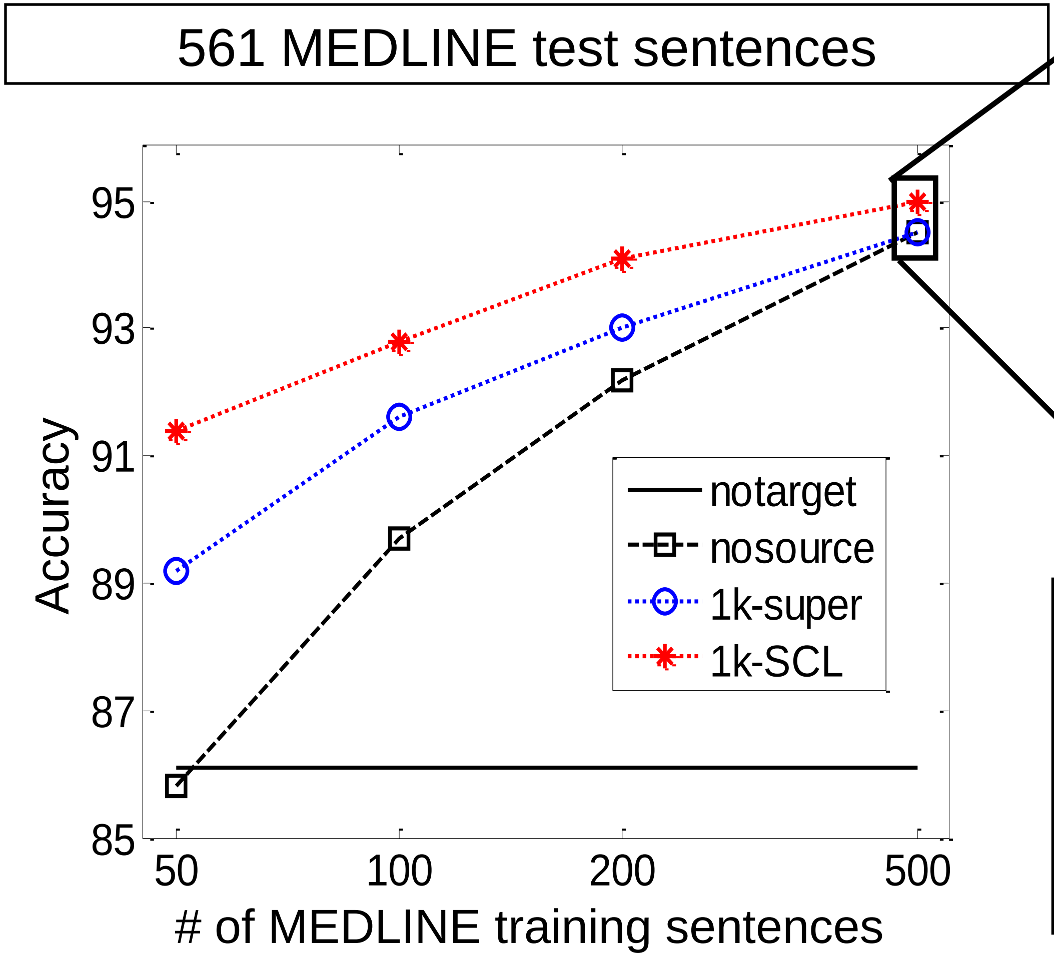


Model	All Words	Unk words
MXPOST	87.2	65.2
super	87.9	68.4
semi-ASO	88.4	70.9
SCL	**88.9**	**72.0**

McNemar's test

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



Model	Accuracy
1k-SCL	95.0
1k-super	94.5
Nosource	94.5

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

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**

Learning Theory & Adaptation

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.

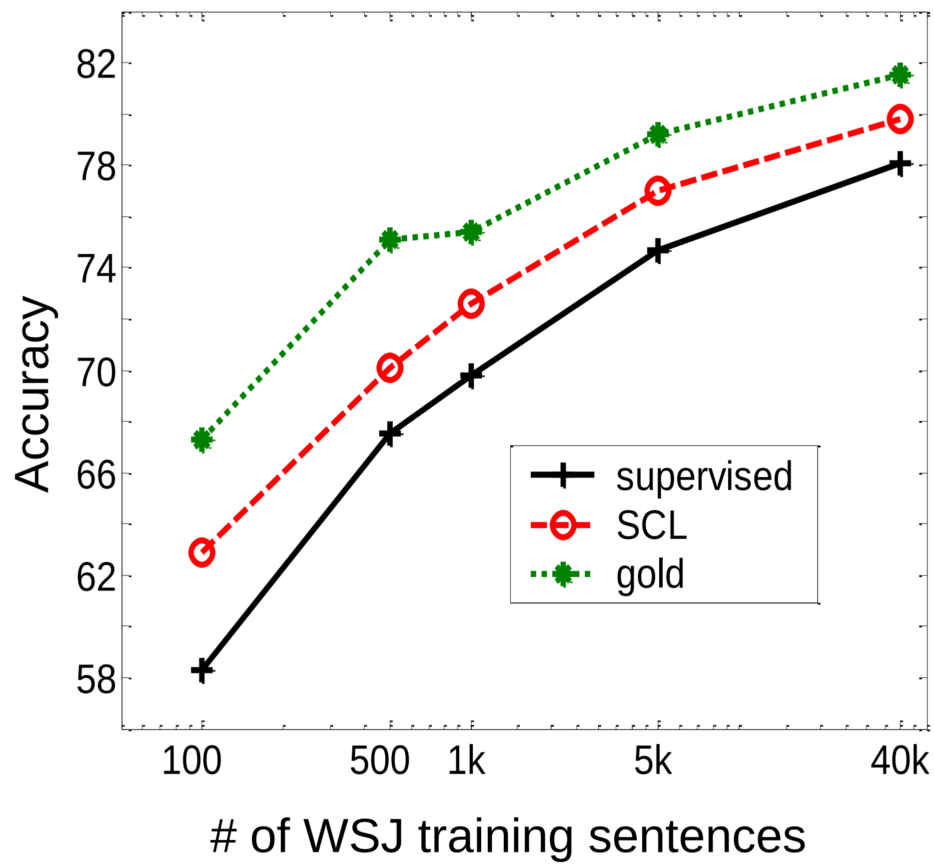
NIPS 2007 (To Appear).

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

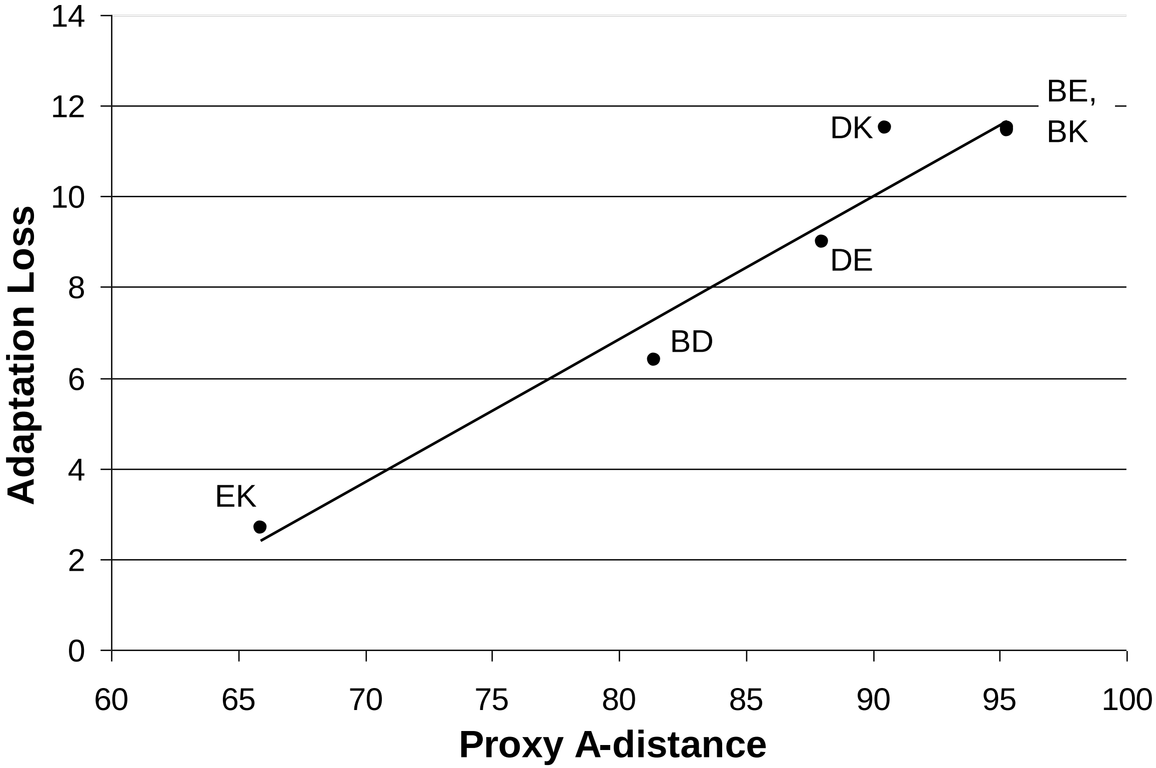
Accuracy for different tagger inputs



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**

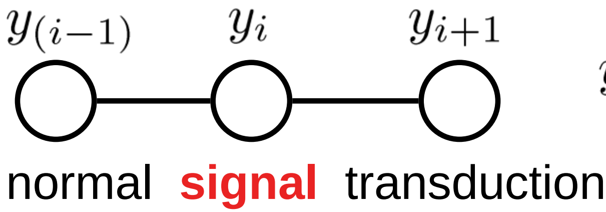
A-distance vs Adaptation loss



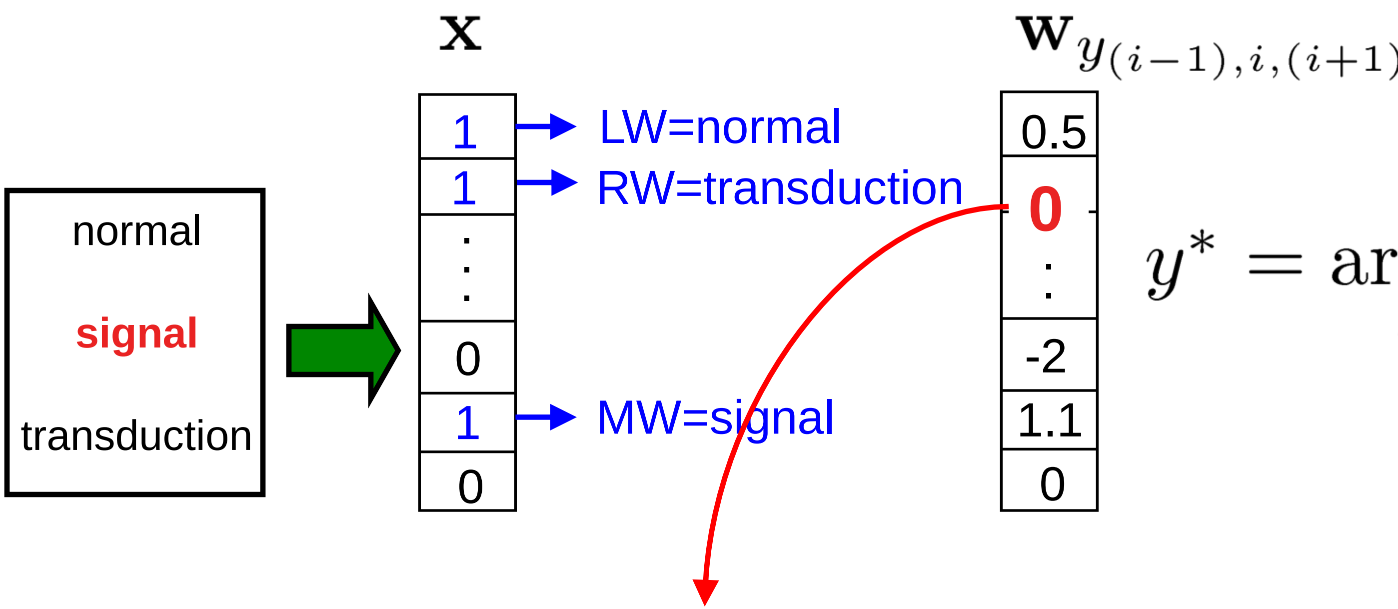
Suppose we can afford to label 2 domains

Then we should label 1 of electronics/kitchen and 1 of books/DVDs

Features & Linear Models



$$y_{(i-1),i,(i+1)} = \text{JJ-NN-NN}$$



$$y^* = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \left[\mathbf{w}_y \cdot \mathbf{x} \right]$$

Problem: If we've only trained on financial news, then
w(RW=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**