Domain Adaptation for Sentiment Classification

ErikAnthony Harté
Sentiment Analysis
Ling 575 Spring Quarter

Introduction

- "Biographies, Bollywood, Boom-boxes and Blenders" (Blitzer, Dredze, and Pereira)
- Motivation
  - We want to classify many different domains…
  - But, most corpora are unlabeled and ideally we would like to only annotate a few of them
  - Idea: Train classifiers on a few select corpora and apply them to similar domains

Introduction

- But…
  - Problem 1: We know that classification accuracy falls off when using a classifier on a domain different than that it was trained on.
  - Problem 2: What the heck does "similar" mean?

Structured Correspondence Learning (SCL)

- Developed by authors in previous work (Blitzer et al. 2006)
- Originally used for POS tagging
- Our intuition is that even when words are distinct between domains, they may be able to serve the same role for classification.

SCL

Example

Computers		Cell Phones
✔	excellent	✔
✔	awful	✔
	good reception	✔
✔	dual-core	

SCL

Feature Correlation

- If "dual-core" and "good reception" are both highly correlated with "excellent", we can align them
- Classifier trained on Computer domain:
  - f("dual-core") = positive
  - f("good reception") = positive
- Unlabeled data

SCL

Pivot Features

- First choose *m* pivot features in both domains
  - labeled and unlabeled
- Pivot Features:
  - binary function such as, {appears within *n* words of <token>}
  - a POS pattern: {PRP VBP PRP$ NN}, as in “I love my KitchenAid”
  - single word such as “excellent”

SCL

Mutual Information

- SCL originally used in POS tagging. Frequently occurring words are often function words (determiners, prepositions) and as such are good POS indicators

- For SC, need more: Needs to be a good predictor of the source label
  - Mutual Information with the source label.

- For example: “excellent” appears in Computers and Cell_Phones and most likely has high MI with positive sentiment label.

SCL

Example – Pivot Selection (books/kitchen)

SCL, not SCL-MI	SCL-MI, not SCL
book	a_must
one	a_wonderful
so	loved_it
very	weak
about	don’t_waste
good	highly_recommended

SCL
Dataset

Labeled (1000 positive / 1000 negative)

- Books
- DVDs
- Electronics
- Kitchen

Unlabeled (also balanced pos/neg)

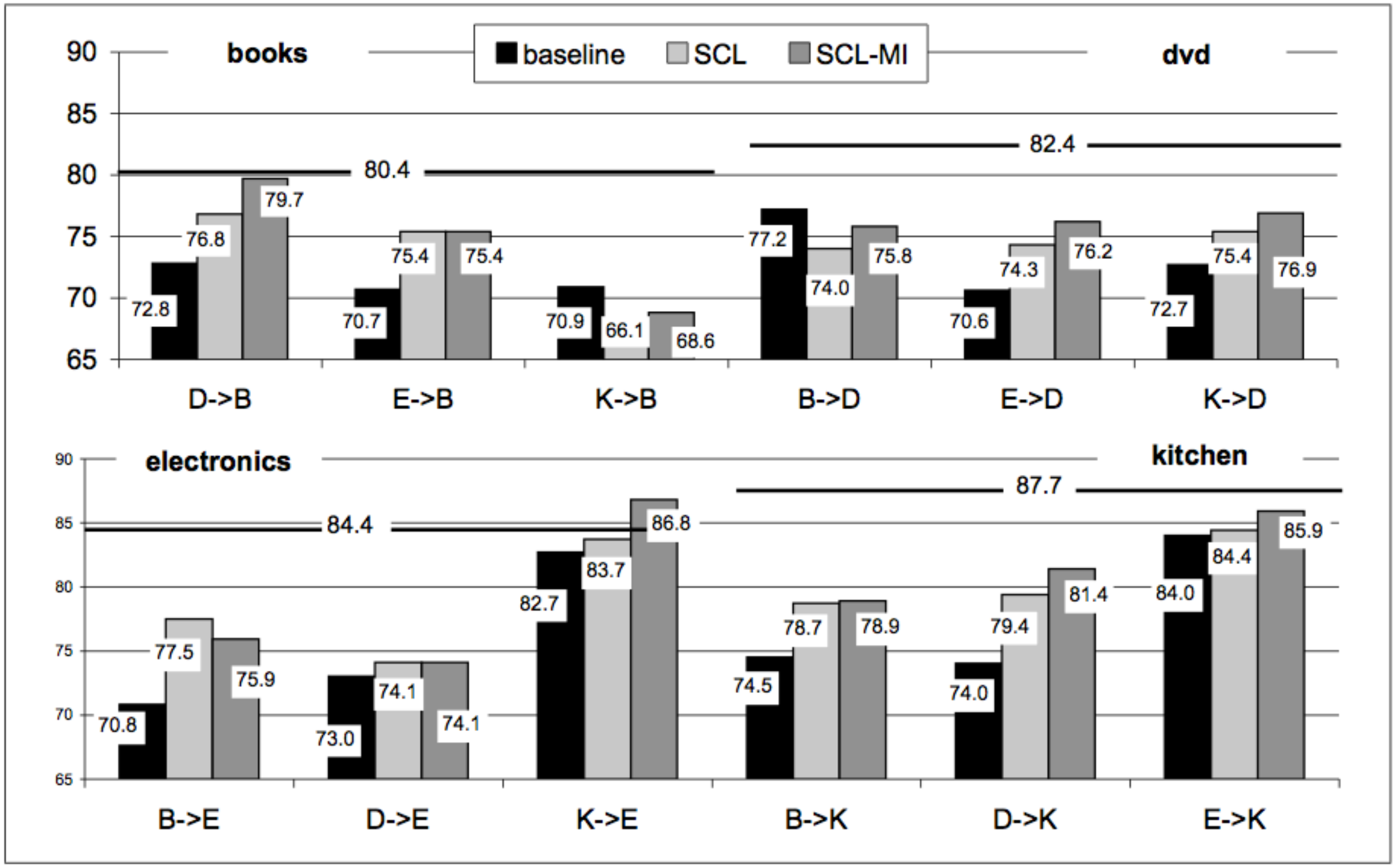
- 3685 DVDs
- 5945 Kitchen

SCL
Dataset

- Dataset 1600 training / 400 test
- Baseline is unigram/bigram classifier with no adaptation
- Gold standard is in-domain classifier trained on same domain as tested

SCL

Results



books
dvd
baseline
SCL
SCL-MI
90
85
80
75
70
65
80.4
82.4
72.8
76.8
79.7
70.7
75.4
75.4
70.9
66.1
68.6
77.2
74.0
75.8
70.6
74.3
76.2
72.7
75.4
76.9
D->B
E->B
K->B
B->D
E->D
K->D
electronics
kitchen
84.4
87.7
70.8
77.5
75.9
73.0
74.1
74.1
82.7
83.7
86.8
74.5
78.7
78.9
74.0
79.4
81.4
84.0
84.4
85.9
B->E
D->E
K->E
B->K
D->K
E->K

SCL

Interpretation

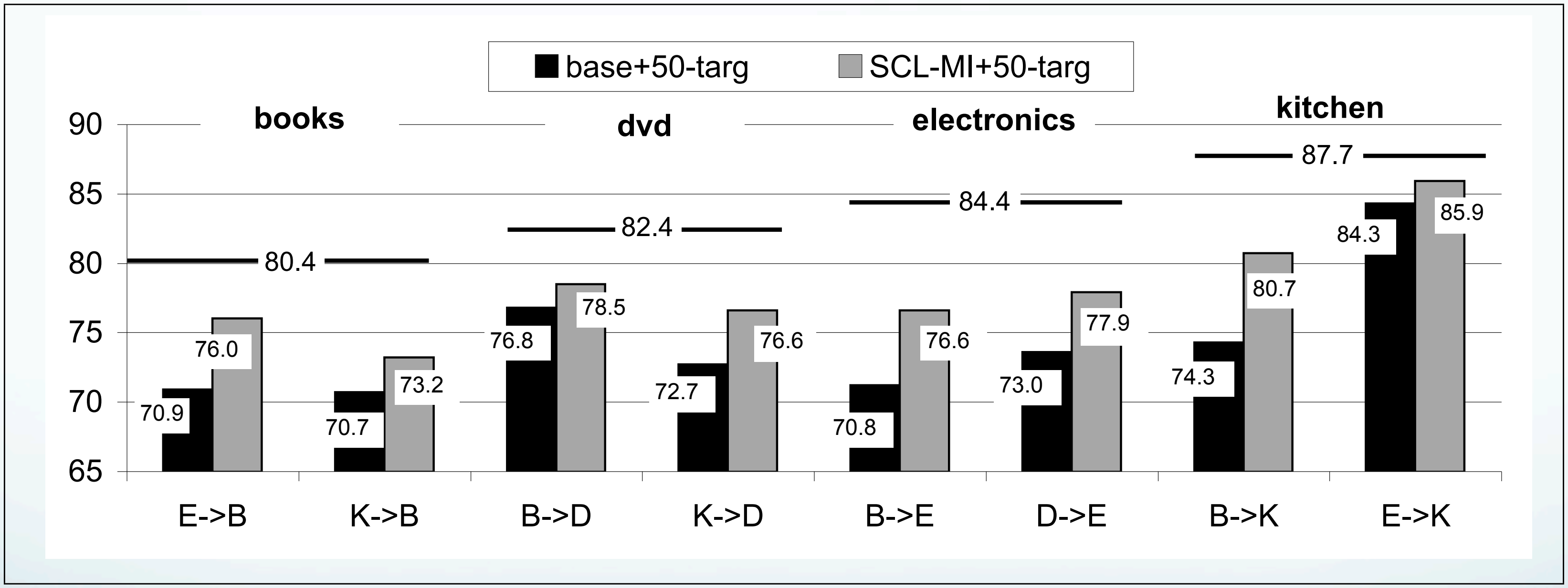
- Adaptation loss (DVDs to Books)
  - baseline = 7.6%
  - SCL-MI = 0.7%
  - reduction in relative error = 90.8%
- Books domain similar to DVDs domain
- Kitchen similar to Electronics
- Books/DVDs NOT similar to Kitchen/Electronics

SCL

Misalignments

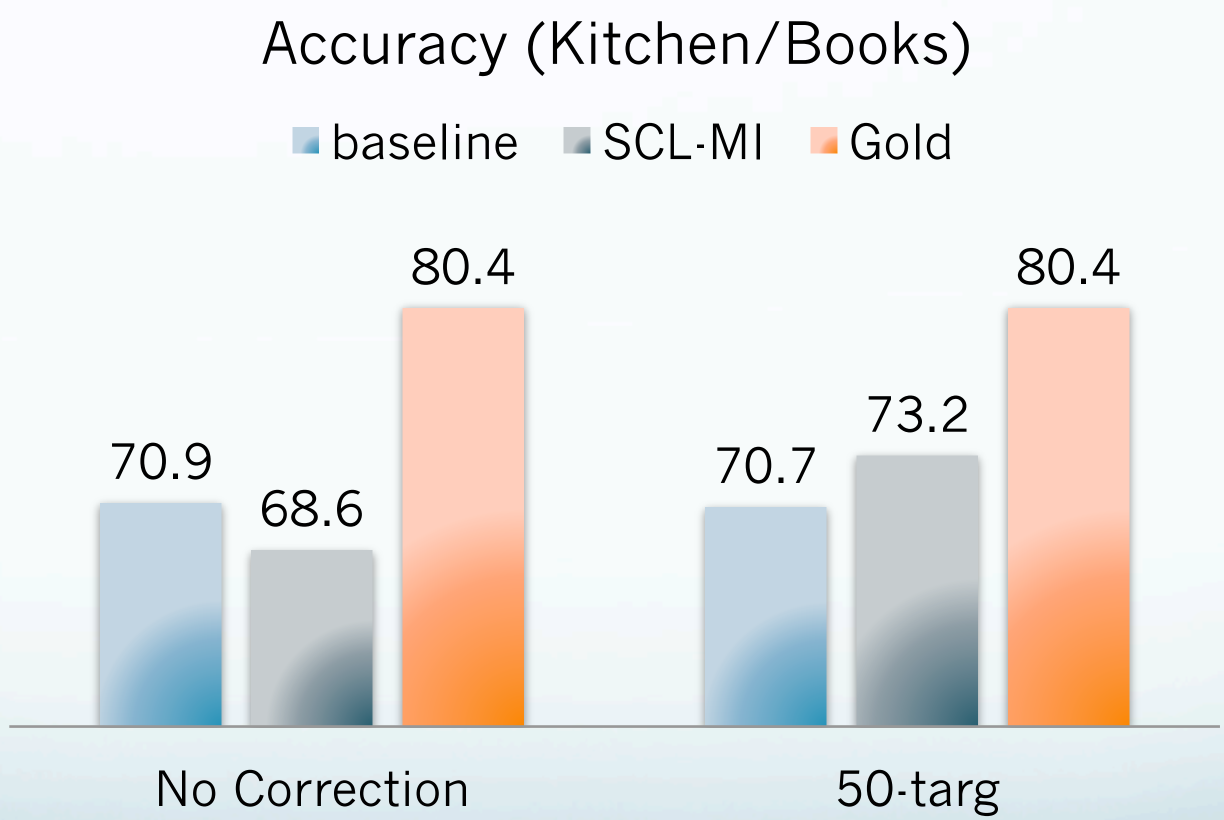
- Problem:
  - Kitchen => Books adaptation shows really poor performance
  - Likely due to feature misalignment
  - Books domain is richer than Kitchen domain
  - SCL matrix results in projections that are uninformative for labeling Kitchen instances
- Solution:
  - Hand label a small selection of target data and re-train classifier to adjust (correct) weights
  - Improves performance of classifier

SCL
Misalignment Correction



SCL

Misalignment Correction



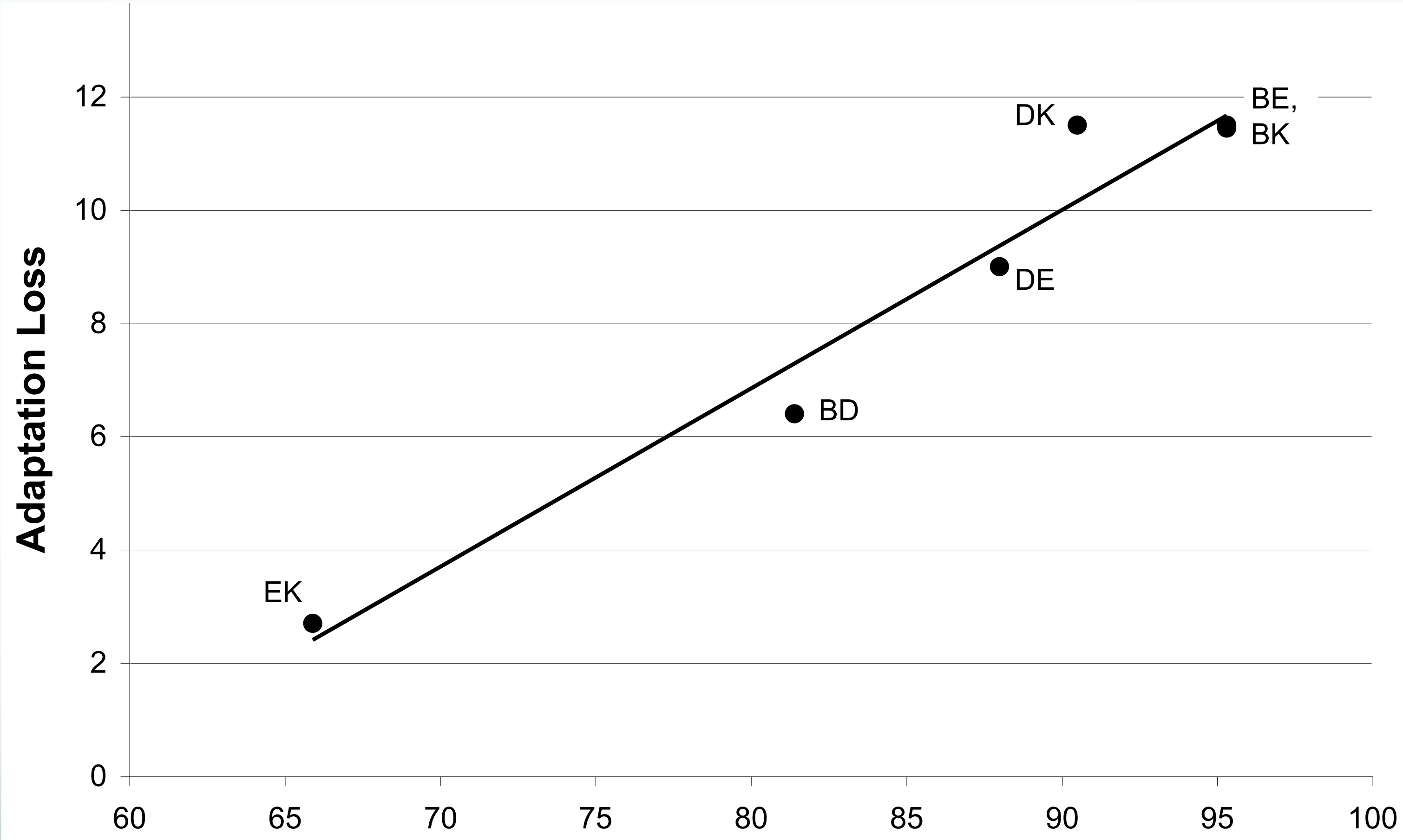
SCL

Similarity

- Need a way to measure difference in probability distributions
  - A-distance
- Previous work Ben-David et al. (2006) showed computing A-distance between two domain is same as *minimizing empirical risk of a classifier* that selects between them.
- So, A-distance can be used to determine similarity between two domains
- Use Huber loss as a proxy for A-distance (also per previous work by Ben-David et al. (2006)

SCL

Similarity – Proxy A-distance



Thoughts

- What is the effort involved to determine the Huber loss or other proxy for the A-distance?
- Is there a better way of selecting the pivots to prevent misalignments?

References

- John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain adaptation for sentiment classification. *In Association for Computational Linguistics.*
- Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. In *Neural Information Processing Systems.*
- John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain Adaptation with Structural Correspondence Learning. *In Empirical Methods in Natural Language Processing.*