DRASO: Declaratively Regularized Alternating Structural Optimization

Partha P. Talukdar, Ted Sandler, Mark Dredze, Koby Crammer

University of Pennsylvania

John Blitzer

Microsoft Research

Fernando Pereira

Google, Inc.

Learning in Text and Language Processing

- Supervised learning algorithms perform very well but labeled data generation is expensive and time consuming.
- Unlabeled data is abundant: exploited by Semi-Supervised Learning (SSL) Algorithms.
- Can we inject *prior knowledge* into SSL algorithms to make them more effective?

Alternating Structural Optimization (ASO)

- ASO (Ando & Zhang, 2005) is a semi-supervised learning algorithm.

- ASO-based algorithms have achieved impressive results:
  - Named Entity Extraction (Ando & Zhang, 2005)
  - Word Sense Disambiguation (Ando, 2006)
  - POS Adaptation (Blitzer et al, 2006)
  - Sentiment Classification Adaptation (Blitzer et al., 2007)

Supervised Training in ASO

- Standard supervised training:

$$\min_{\mathbf{w}} \sum_{i=1}^{m} L(\mathbf{w}^{'}\mathbf{x}_i, y_i) + \lambda \, ||\mathbf{w}||_2^2$$

- Supervised training in ASO:

$$\min_{\mathbf{w},\mathbf{v}} \sum_{i=1}^{m} L(\mathbf{w}^{'}\mathbf{x}_i + \mathbf{v}^{'}\Phi\mathbf{x}_i, y_i) + \lambda \, ||\mathbf{w}||_2^2$$

Learned from unlabeled data

How does ASO work?

1. Given a target problem (e.g. sentiment classification), design multiple **auxiliary problems**.

2. **Train auxiliary problems** on unlabeled data.

3. **Reduce dimension** of the weight vector matrix. Let Φ be this shared lower dimensional transformation matrix.

4. Use Φ to **generate new features** for training instances. Learn weight for these new features (along with existing features) using labeled training data.

Auxiliary Problems for Sentiment Classification

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, 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

Auxiliary Problems

Presence or absence of frequent words and bigrams:

don't_waste, horrible, suffering

Step 2: Training Auxiliary Problems

For each unlabeled instance, create a binary presence / absence label

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

(2) An **excellent** book. Once again, another wonderful novel from Grisham

Binary problem: Does "**not buy**" appear here?

- **Mask** and predict auxiliary problems using other features
- Train n **linear predictors**, one for each binary auxiliary problem

Using Prior Knowledge in ASO

- Many features have equal predictive power.
  - e.g. presence of *excellent* or *fantastic* in a document is equally predictive of it being a positive review.

- Can we constrain the model so that similar features get similar weights (not necessarily exact)?

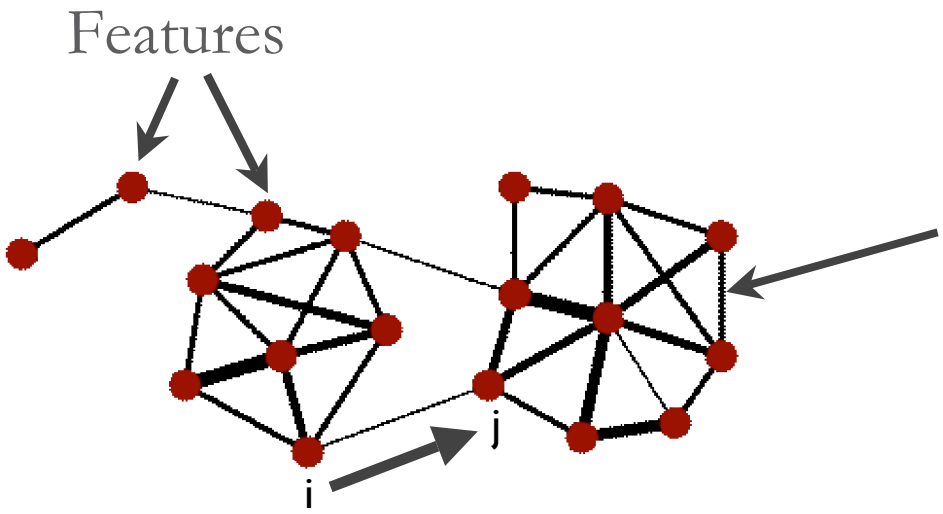
- Answer: Locally Linear Feature Regularization (LLFR) (Sandler et al., 2008)

Feature Similarity as Prior Knowledge

Domain Knowledge:

- Neighboring features in lattice-structured data (e.g., images, time series data) often provide similar information.
- lexicons: tell us which words are synonyms

Model Feature Similarities with a Feature Graph

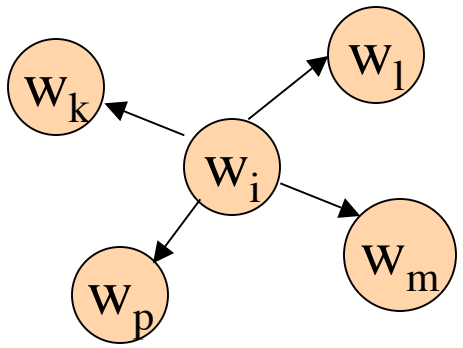


P_{ij} is similarity of
feature i to feature j

Regularization Criteria

Because we believe features are similar to neighbors, we shrink weights toward neighborhood mean.

$$R(\mathbf{w}) = \sum_i \left(\mathbf{w}_i - \sum_j P_{ij} \mathbf{w}_j \right)^2$$



Prefer each weight to be a locally linear (convex) combination of its neighbors.

Regularization in Auxiliary Problem Training

- ASO

$$\text{Loss} + \lambda \, ||\mathbf{w}||_2^2$$

- DRASO

$$\text{Loss} + \lambda \, ||\mathbf{w}||_2^2 + \gamma \, \mathbf{w}^{'} M \mathbf{w}$$

where $M = (I - P)^{'}(I - P)$

What is the effect of this new regularizer?

- Use of SVD in ASO is not just a matter of choice: *it follows from the derivation*.
- The new regularizer (in DRASO) results in a different eigenvalue problem (derivation is in the paper) which can be efficiently solved.
- The eigenvalue problem in DRASO is a generalized version of the one in ASO, the two are same when $\mathbf{M} = \mathbf{I}$.

Experimental Results

- Book reviews from Amazon.com (Blitzer et al., 2007)
- Prior knowledge was obtained from SentiWordNet (Esuli & Sebastini, 2006).
- Manually selected 31 positive and 42 negative sentiment words from ranked SentiWordNet lists.
- Each word was connected to its 10 nearest neighbors.

University of Pennsylvania

Comparing Learned Projections

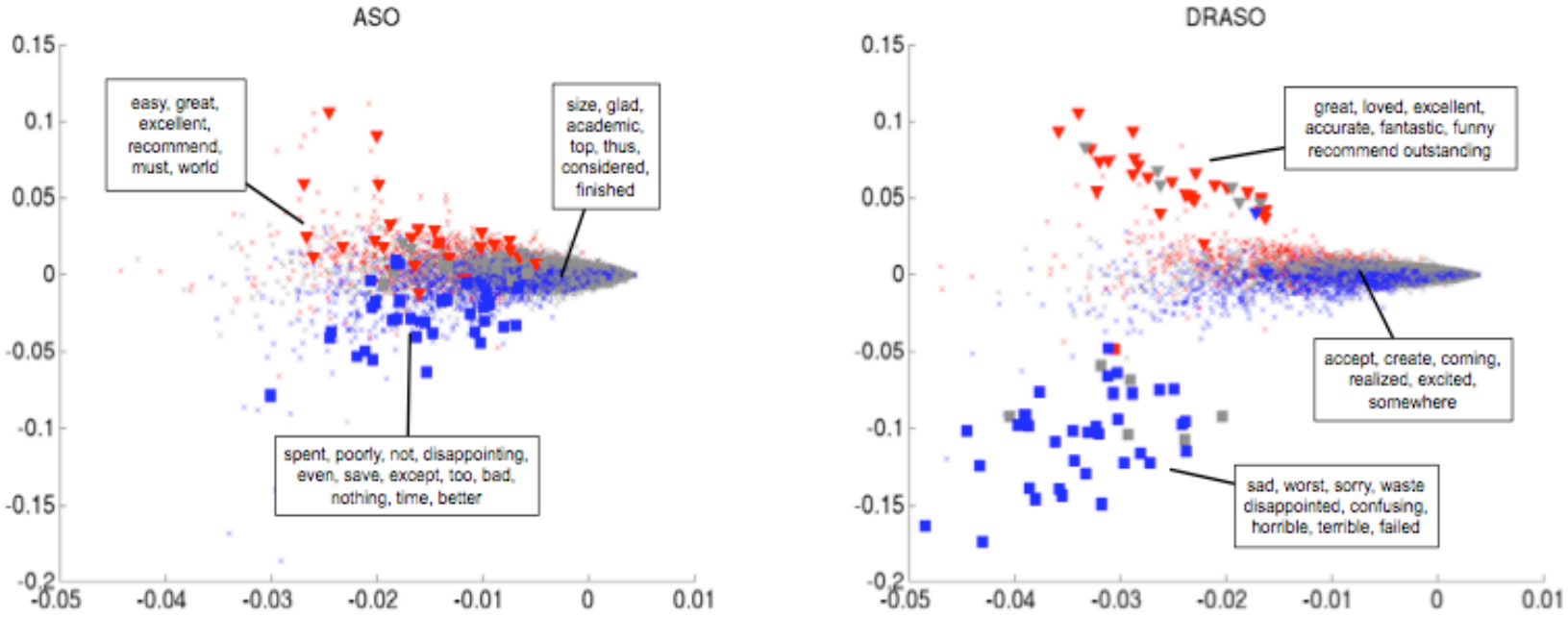


Figure 1. ASO and DRASO projections of 32,502 words into a two-dimensional space. Squares (negative) and triangles (positive) indicate prior knowledge words. Polarity of features as measured from labeled data is indicated by blue (negative), red (positive) and grey (neutral). Some of the features are annotated to demonstrate the effects of the projection.

Conclusion

- We have presented a principled way to inject prior knowledge into the ASO framework.
- Current work: Application to other problems where similar regularization can be useful (e.g. NER).

Thanks!