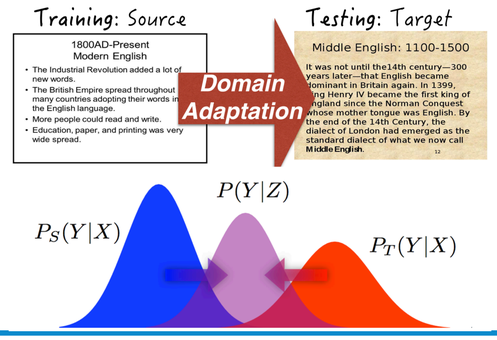
Fast Easy Unsupervised Domain Adaptation with Marginalized Structured Dropout

Yi Yang and Jacob Eisenstein

Domain Adaptation



Example: Part-of-speech Tagging

Source:
And God said, Let ...
CC NNP VBD, VB ...

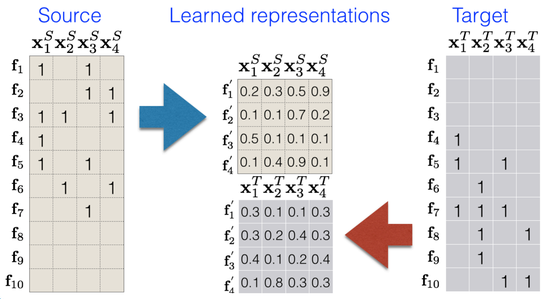
Features:
- Mid_said source spec
- Prev_God cross domain
- Next_Let source spec
- ...

Target:
And God seide, Liyt ...
CC NNP VBD, VB ...

Features:
- Mid_seide target spec
- Prev_God cross domain
- Next_Liyt target spec
- ...

Representation Learning

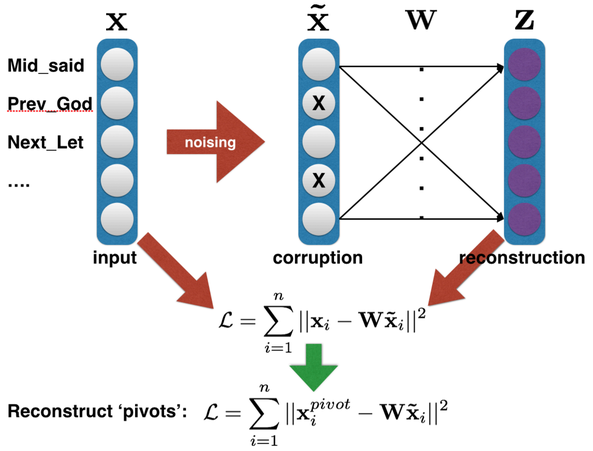
Learn new sets of dense features:



Representation learning for domain adaptation:

- Structural Corresponding Learning (SCL) [1]
- Brown Clustering
- (marginalized) Stacked Denoising Autoencoders (SDA/mSDA) [2,3]
- Latent Variable Models
- Neural Probabilistic Language Model

Denoising Autoencoders



$$\mathcal{L} = \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{W}\tilde{\mathbf{x}}_i||^2$$

Reconstruct 'pivots': $\mathcal{L} = \sum_{i=1}^{n} ||\mathbf{x}_i^{pivot} - \mathbf{W}\tilde{\mathbf{x}}_i||^2$

Closed-form solution: $\mathbf{W} = \mathbf{P}\mathbf{Q}^{-1}$,
with $\mathbf{P} = \sum_{i=1}^{n} \mathbf{x}_i\tilde{\mathbf{x}}_i^\top$ and $\mathbf{Q} = \sum_{i=1}^{n} \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top$

Marginalized Denoising Autoencoders [3]:
$\mathbf{P} = \sum_{i=1}^{n} E[\mathbf{x}_i\tilde{\mathbf{x}}_i^\top]$, and $\mathbf{Q} = \sum_{i=1}^{n} E[\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top]$

Learned representations: $\tanh(\mathbf{W}\mathbf{X})$

Compute P and Q under dropout noise:
For each feature of an instance, remove it with probability p.

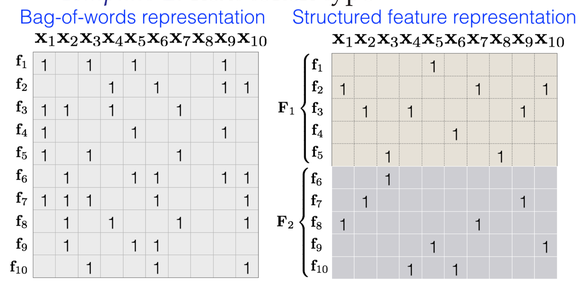
$$\mathbf{Q}_{\alpha,\beta} = \begin{cases} (1-p)^2\mathbf{S}_{\alpha,\beta} & \text{if } \alpha \neq \beta \\ (1-p)\mathbf{S}_{\alpha,\beta} & \text{if } \alpha = \beta \end{cases},$$

$$\mathbf{P}_{\alpha,\beta} = (1-p)\mathbf{S}_{\alpha,\beta},$$

where $\mathbf{S} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top$ is the scatter matrix, α and β index two features.

Structured Dropout

Feature templates: And God said, let ...
- *Template 1*: previous token type
- *Template 2*: next token type



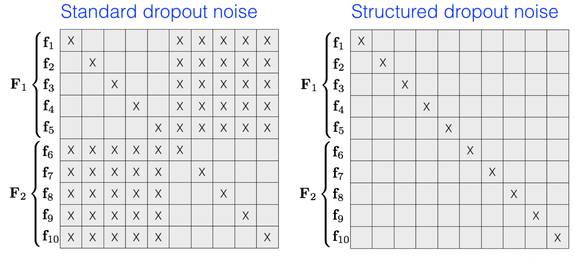
Compute P and Q under structured dropout:
Randomly choose one active feature (type) to keep, dropout all other features.

$$\mathbf{Q}_{\alpha,\beta} = \begin{cases} 0 & \text{if } \alpha \neq \beta \\ \frac{1}{K}\mathbf{S}_{\alpha,\beta} & \text{if } \alpha = \beta \end{cases}$$

$$\mathbf{P}_{\alpha,\beta} = \frac{1}{K}\mathbf{S}_{\alpha,\beta},$$

where K is the number of feature types. There is no free hyperparameter.

Shape of Q under different noises:



Eliminate matrix inverse for $\mathbf{W} = \mathbf{P}\mathbf{Q}^{-1}$!

Evaluation: Same accuracy, 25x faster!

Datasets: Tycho Brahe corpus (historical Portuguese texts with 383 tags)

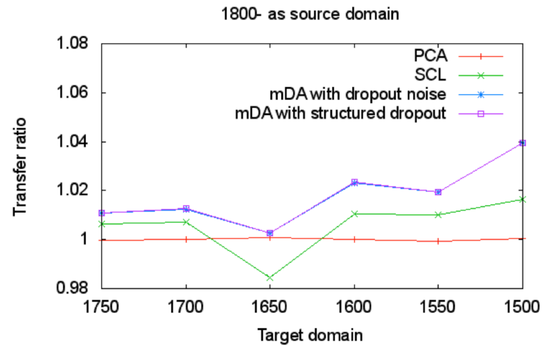
Dataset	# of Tokens				
	Total	Narrative	Letters	Dissertation	Theatre
1800-1849	125719	91582	34137	0	0
1750-1799	202346	57477	84465	0	60404
1700-1749	278846	0	130327	148519	0
1650-1699	248194	83938	115062	49194	0
1600-1649	295154	117515	115252	62387	0
1550-1599	148061	148061	0	0	0
1500-1549	182208	126516	0	55692	0
Overall	1480528	625089	479243	315792	60404

Experiment setup:
- *CRF tagger*: 16 feature types, 372,902 features, and 1572 pivots.
- *Methods*: baseline, PCA, SCL
- *Parameters*: decided with development data on the training set.

Representation learning time:

Method	PCA	SCL	mDA	
			dropout	structured
Time (sec)	7,779	38,849	8,939	**339**

Results:



Task	baseline	PCA	SCL	mDA	
				dropout	structured
from 1800-1849					
→ 1750	89.12	89.09	89.69	**90.08**	**90.08**
→ 1700	90.43	90.43	91.06	91.56	**91.57**
→ 1650	88.45	88.52	87.09	88.69	**88.70**
→ 1600	87.56	87.58	88.47	89.60	**89.61**
→ 1550	89.66	89.61	90.57	**91.39**	**91.39**
→ 1500	85.58	85.63	86.99	**88.96**	88.95
from 1750-1849					
→ 1700	94.64	94.62	94.81	**95.08**	**95.08**
→ 1650	**91.98**	90.97	90.37	90.83	90.84
→ 1600	92.95	92.91	93.17	**93.78**	**93.78**
→ 1550	93.27	93.21	93.75	**94.06**	94.05
→ 1500	89.80	89.75	90.59	**91.71**	**91.71**

References

[1] John Blitzer et al. Domain Adaptation with Structural Correspondence Learning. In *EMNLP'06*.
[2] Xavier Glorot et al. Domain Adaptation for Large-Scale Sentiment ClassiïňĄcation: A Deep Learning Approach. In *ICML '11*
[3] Minmin Chen et al. Marginalized Denoising Autoencoders for Domain Adaptation. In *ICML '12*

Acknowledgments

This research was supported by National Science Foundation award 1349837. The first author was also supported by National Science Foundation ACL travel award.