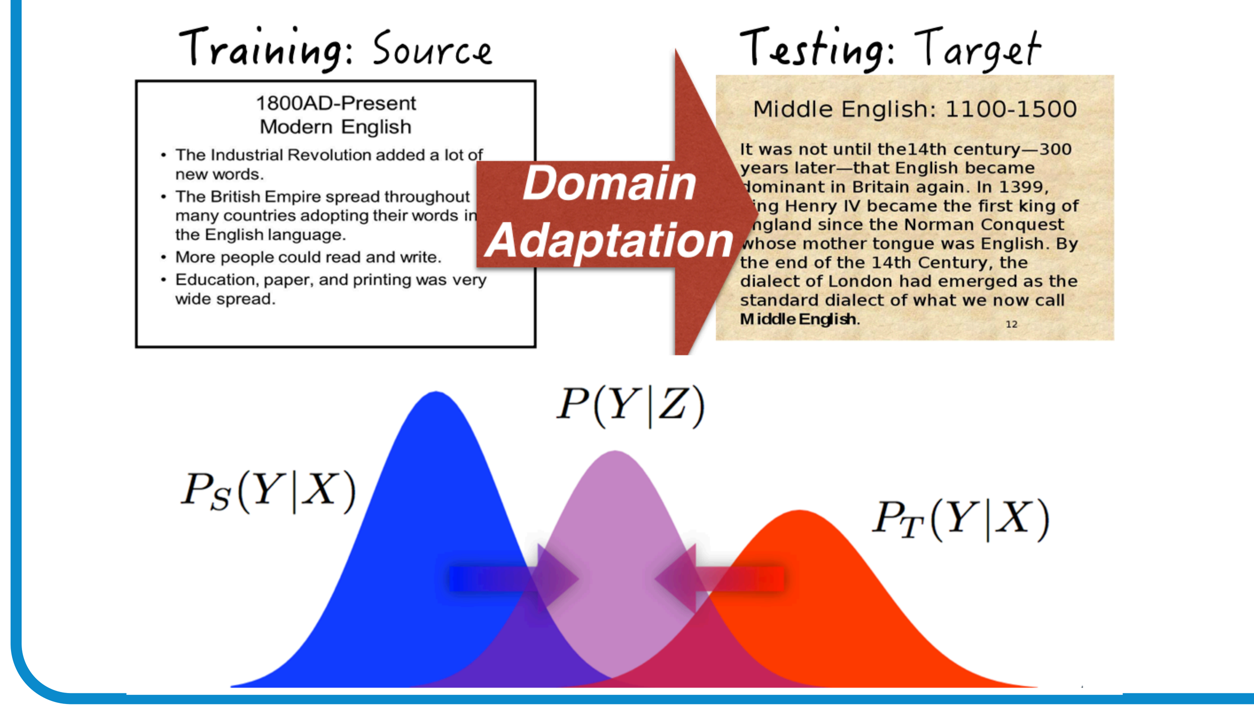
Unsupervised Multi-Domain Adaptation with Feature Embeddings

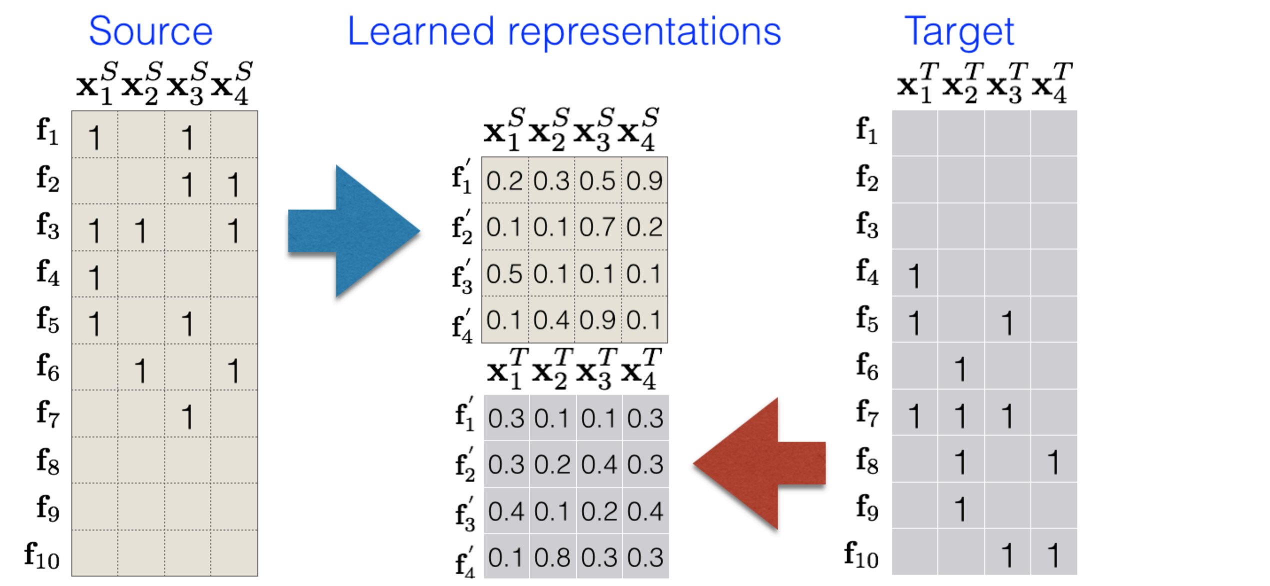
Yi Yang and Jacob Eisenstein

*This research was supported by National Science Foundation award 1349837.

Domain Adaptation and Representation Learning

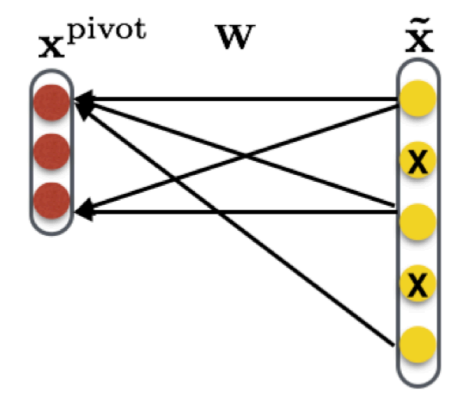


Overcome domain shift (Ben-David et al., 2010)



Pivot-Based Approaches

Denoising Autoencoders: Learning a projection matrix $\mathbf{W}$ by reconstructing pivot features (Chen et al., 2012; Yang and Eisenstein, 2014)



Pivot features:

- A small number of cross-domain features
- Each pivot leads to a new feature

Drawbacks of pivot-based approaches:

- Selection of pivots often requires task-specific heuristics
- Pivots correspond to a small subspace of the full feature co-occurrence matrix
- They are computationally expensive for learning the transformations or downstream training
- Not clear how to adapt the approaches to multi-domain adaptation tasks

Feature Embeddings

Structured feature representation:

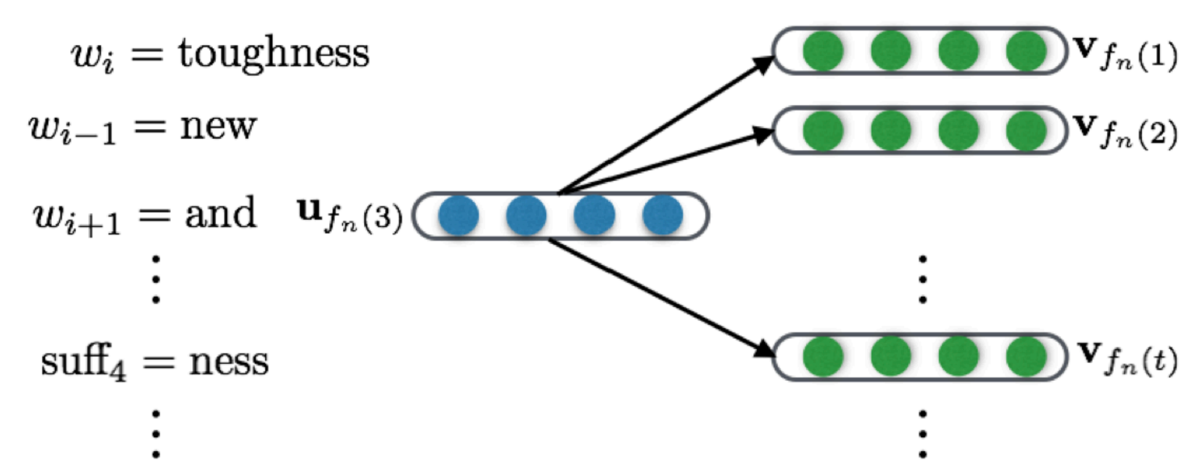
- Many core NLP tasks (e.g. POS tagging, NER, Chunking) exploit feature templates for extracting features
- There is exactly one active feature per template in each instance

...	DT	NN	IN	DT	JJ	NN	CC	NN	...
...	a	sign	of	a	new	toughness	and	divisiveness	...

Feature template	Feature value
Current_token	$w_i =$ toughness
Previous_token	$w_{i-1} =$ new
Next_token	$w_{i+1} =$ and
Suffix_4gram	$\text{suff}_4 =$ ness
...	...

Feature embeddings for domain adaptation:

- Induce low-dimension embeddings using feature co-occurrence information as supervision
- Predict active features of other templates iteratively



Objective function:

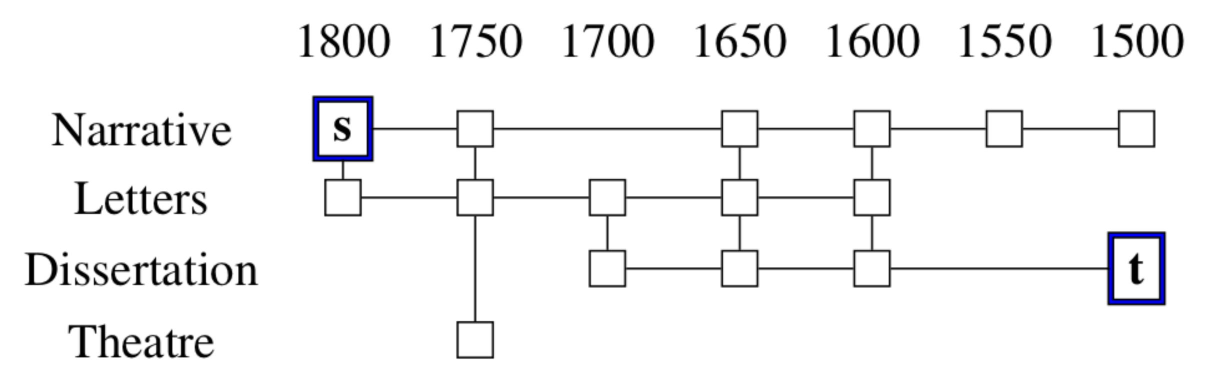
$$\ell_n = \frac{1}{T}\sum_{t=1}^{T}\sum_{t'\neq t}^{T}\left[\log\sigma(\mathbf{u}_{f_n(t)}^\top \mathbf{v}_{f_n(t')}) + k\mathbb{E}_{i\sim P_{t'}^{(n)}}\log\sigma(-\mathbf{u}_{f_n(t)}^\top \mathbf{v}_i)\right]$$

Learned representations:

$$\mathbf{x}_n^{(\text{aug})} = \mathbf{x}_n \oplus \tanh[\mathbf{u}_{f_n(1)} \oplus \cdots \oplus \mathbf{u}_{f_n(T)}]$$

Multi-Domain Adaptation

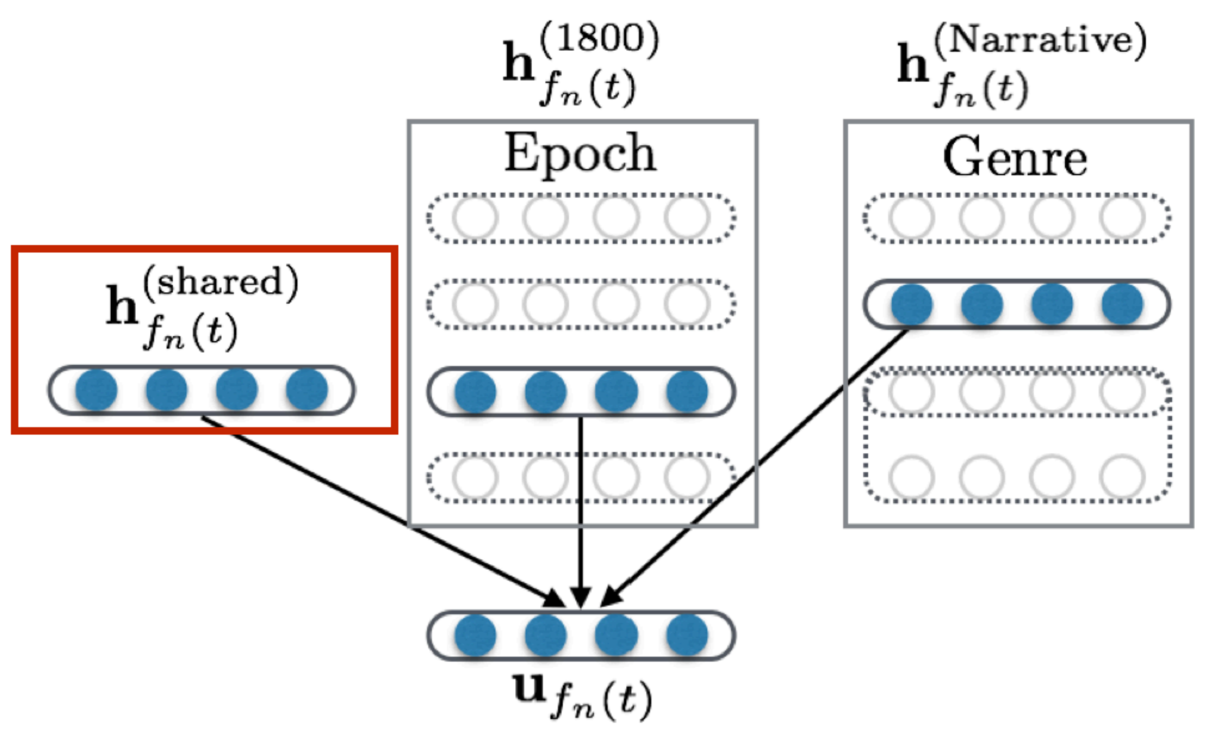
Can we leverage unlabeled data from multiple domains to improve performance in the target domain?



- Prior unsupervised domain adaptation work assumes single source and target domains
- There exist valuable metadata (e.g. genres, epochs) associated with multiple domains
- Previous multi-domain adaptation work focused on supervised setting

Feature embeddings across domains:

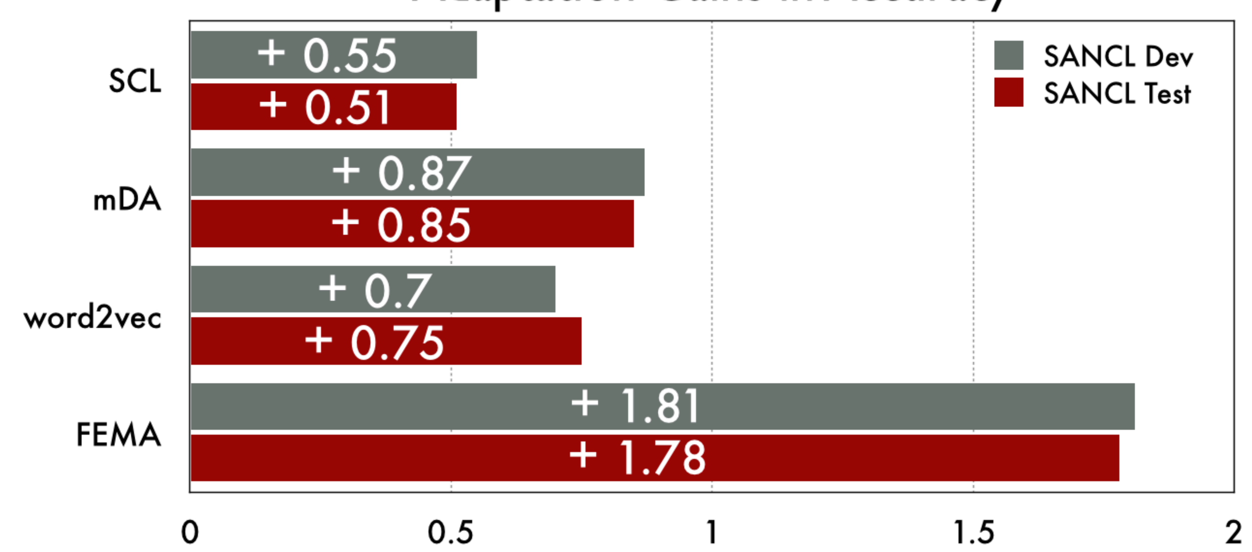
- Aggregating multiple embeddings



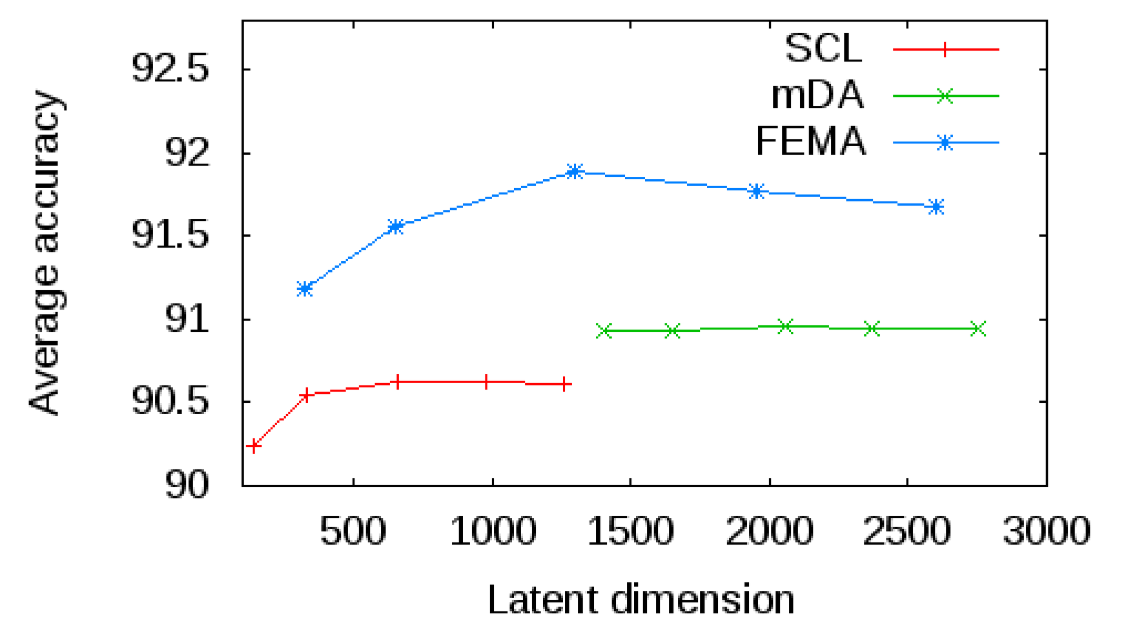
This "subtracts out" domain specific effects, leaving out more robust representations.

Evaluation

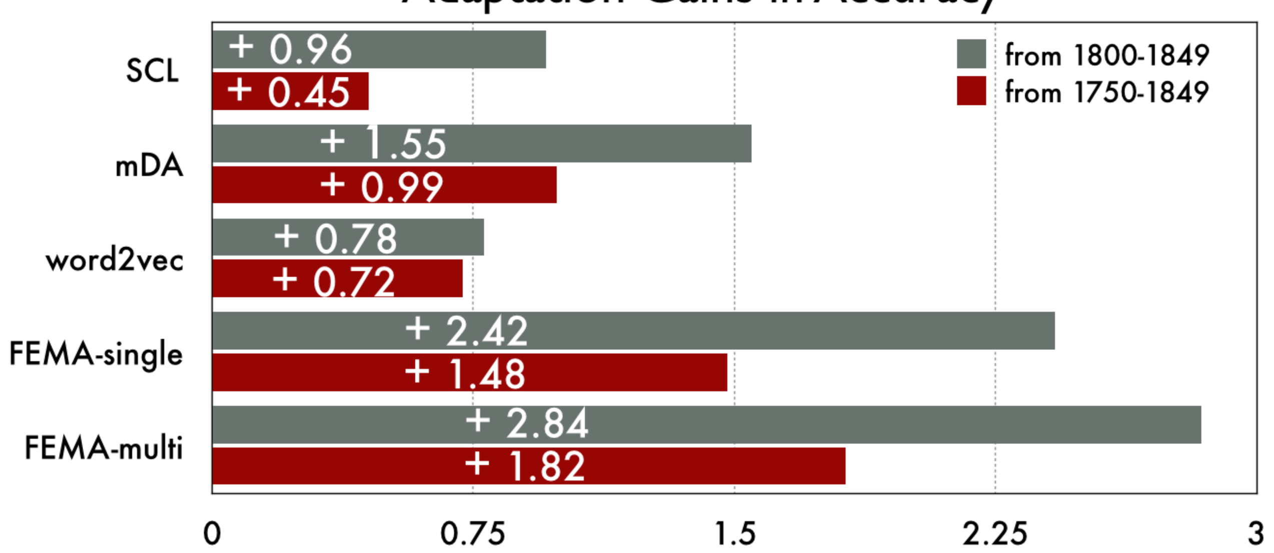
Evaluation 1: POS tagging on SANCL datasets (WSJ to Web text)



Accuracy results with different latent dimensions



Evaluation 2: POS tagging on Tycho Brahe corpus (historical Portuguese texts)



Label consistency of the Q-most similar words:

Embedding	$Q = 10$
WORD2VEC	46.17
FEMA-current	66.93
FEMA-prev	54.18
FEMA-next	55.78
FEMA-all	69.60

- FEMA captures more syntactic regularities than word2vec
- Words with the same most common POS tags are similar in the embedding space

Most similar words in the embedding space:

'new'	
FEMA-current	nephew, news, newlywed, newer, newspaper
FEMA-prev	current, local, existing, international, entire
FEMA-next	real, big, basic, local, personal
WORD2VEC	current, special, existing, newly, own
'toughness'	
FEMA-current	tightness, trespass, topless, thickness, tenderness
FEMA-prev	underside, firepower, buzzwords, confiscation, explorers
FEMA-next	aspirations, anguish, pointers, organisation, responsibilities
WORD2VEC	parenting, empathy, ailment, rote, nerves
'and'	
FEMA-current	amd, announced, afnd, anesthetized, anguished
FEMA-prev	or, but, as, when, although
FEMA-next	or, but, without, since, when
WORD2VEC	but, while, which, because, practically