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 \( W \) by reconstructing pivot features (Chen et al., 2012; Yang and Eisenstein, 2014)

**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

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

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

**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**

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

<table>
<thead>
<tr>
<th>Feature template</th>
<th>Feature value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current_token</td>
<td>( w_u = \text{toughness} )</td>
</tr>
<tr>
<td>Previous_token</td>
<td>( w_{u-1} = \text{new} )</td>
</tr>
<tr>
<td>Next_token</td>
<td>( w_{u+1} = \text{and} )</td>
</tr>
<tr>
<td>Suffix_4gram</td>
<td>( w_s = \text{less} )</td>
</tr>
</tbody>
</table>

Objective function:

\[
\ell_u = \frac{1}{2} \sum_{i=1}^{T} \sum_{j=1}^{n_u} \log \sigma(u_{j}^u)h_i + \frac{k}{2} \sum_{i=1}^{T} \sum_{j=1}^{n_u} \log \sigma(-u_{j}^u)v_i
\]

Learned representations:

\[
x_u^{(\text{aug})} = x_u + \tanh(u_{f_u(1)} + \cdots + u_{f_u(T)})
\]

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

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

**Label consistency of the \( Q \)-most similar words:**
- 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: and, announced, and, anesthetized, anguished
  - FEMA-prev: or, but, as, when
  - FEMA-next: or, but, without, since, when
  - WORD2VEC: but, while, which, because, practically