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**

**Pivot-based Approaches**

**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_0 = toughness</td>
</tr>
<tr>
<td>Previous_token</td>
<td>w_{-1} = news</td>
</tr>
<tr>
<td>Next_token</td>
<td>w_{+1} = and</td>
</tr>
<tr>
<td>Suffix_ngram</td>
<td>suffix = ness</td>
</tr>
</tbody>
</table>

Feature embeddings for domain adaptation:
- Induce low-dimensional embeddings using feature co-occurrence information as supervision
- Predict active features of other templates iteratively

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

**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

**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)

Accuracy results with different latent dimensions

**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

**Learned representations:**

\[ \mathbf{x}_{n}^{(aug)} = \mathbf{x}_{n} \oplus \tanh(\mathbf{u}_{f_{n}(1)} \oplus \cdots \oplus \mathbf{u}_{f_{n}(T)}) \]

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

**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
- 'toughness': FEMA-current, current, local, existing, international, entire
- 'and': FEMA-current, real, big, local, personal, word2vec, special, existing, newly, own
- 'and': FEMA-current, amd, announced, and, anesthetized, anguished