**Mixed Membership Word Embeddings for Computational Social Science**

James Foulds
University of Maryland, Baltimore County

---

**Overview**

- **Word embeddings** find similarities between words, leading to improved performance for many NLP tasks: translation, part-of-speech tagging, chunking, NER ...
- **Allow NLP** to "scratch," without feature engineering
- **Typically trained w-big data setting**
- **Have not yet been widely adopted for computational social science** research due to the following limitations:
  - Target corpus of interest is often not big data
  - It is important for the model to be interpretable
- I propose a method for training interpretable word embeddings without big data, for computational social science, leveraging insights from topic models

**Contributions**

- **Interpretable, statistically efficient embedding model**
  - **Key insight:** Mixed membership representation for parameter sharing while retaining model flexibility
- **Efficient training algorithm**, using recent advances from both topic models and word embeddings:
  - **Noise-contrastive estimation** (Mikolov et al., 2013)
- Proposed training algorithm is (amortized) sublinear time in the vocabulary size and number of topics
- **Extensive quantitative experimental results**
- Computational social science case studies
- **Practical recommendations** and insights based on these results, especially in the use of generic big data embeddings, which is a very common practice in NLP

**Background: Word Embeddings**

- **Represent dictionary words with vectors.** Similar words have similar vectors.
- **Simple model, scales easily to large data sets**
- **Can beat deep neural network models**

**Amortized Sublinear Time Training for MM Skip-Gram**

- **Bayesian inference via collapsed Gibbs sampling**
  \[
  p(z_t = k) \propto \exp\left( \sum_{w \in V} (v(w) \cdot u_k + b_k) \right) 
  \]
- **Scale to many topics:** Metropolis-Hastings-Walker
- **Alias table data structure, amortized O(1) sampling**
- **"Mixture of experts" proposal**
- **Simulated annealing to escape early local maxima**

**Inference for MM Skip-Gram Topic Model**

- **Online EM impractical - O(KV) updates**
- **Key insight:** MMSG topic model equivalent to word embedding model *(up to the Dirichlet prior)*
- **Pre-solve E-step via topic model CGS MWMH algorithm**
- **Apply noise-contractive estimation to solve M-step**

**Connections to Topic Models, and Mixed Membership Extension to the Skip-Gram**

<table>
<thead>
<tr>
<th>Skip-gram</th>
<th>Mixed membership model</th>
</tr>
</thead>
<tbody>
<tr>
<td>For each word in the corpus ( w ), ( p(w</td>
<td>\theta) \propto \exp\left( \sum_{k \in K} \alpha_k u_k + b_k \right) )</td>
</tr>
</tbody>
</table>

**Machine learning topic representation - belief learning**

- **Contrast:** We used an SVM when learning to predict the class above
- **Word embeddings are convex combinations of topic embeddings**
- **Words have mixed membership distributions over topics**
- **Topics have embeddings \( \theta_t \) words don't. Resolves polysemy
- **Fewer vectors than words:** Statistical efficiency on small data
- **Word embeddings recovered as prior \( \pi_t \) or posterior mean \( \tilde{\theta}_t \): convex combinations of topic embeddings**
- **Interpretable:** Topics can be interpreted via top words lists, word embeddings are defined in terms of topic embeddings
- **Context can be leveraged to improve embeddings \( \tilde{T}_{\tilde{w}} \) via the posterior distribution over topics for a word token \( w \)****

**Experimental Results**

**Top Words in Topics**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cat</td>
<td>1246</td>
</tr>
<tr>
<td>2</td>
<td>dog</td>
<td>1136</td>
</tr>
<tr>
<td>3</td>
<td>man</td>
<td>432</td>
</tr>
<tr>
<td>4</td>
<td>king</td>
<td>216</td>
</tr>
<tr>
<td>5</td>
<td>queen</td>
<td>108</td>
</tr>
</tbody>
</table>

**Downstream Tasks: Classification and Regression**

- **Document categorization (classification accuracy; larger is better)**
- **Reciprocal rank for documents (except for SOTU, which is very small)**
- **Target corpora beats generic big-data section (except for SOTU dataset)**
- **MWS MSG, generic design vectors show complementary information**

**Vector Composition in Topic Space**

- Nearest topic after composition of mess vectors for words
  - object + recognition
  - character + recognition
  - speech + recognition
  - computer + vision
  - computer + science
- **topics visual object recognition model**
- **character recognition**
- **speech recognition**
- **hmm system hybrid computer vision lese image pattern university science colorado**
- **error training set data performance and computational distribution model matrix**

**Data Visualization: Document, Topic, and Author Embeddings on State of the Union Addresses and NIPS Articles (t-SNE Projections)**

**NIPS Authors**

- **Blue = authors**
- **Red = authors**
- **Gray = topics**

**NIPS Documents**

- **Orange = topics**
- **Blue = authors**
- **Gray = topics**

---

**State of the Union Addresses**

- **Democrats (blue), blended topics**
- **Republicans (gray), blended topics**

**NIPS Documents**

- **Orange = topics**
- **Blue = authors**
- **Gray = topics**

---

**Experimental Results**

**Predicting Held-Out Words**

<table>
<thead>
<tr>
<th>Prediction task</th>
<th>Model</th>
<th>Prediction accuracy (higher is better)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict context words</td>
<td>Using full context help (posterior over topic or summarizing vectors)</td>
<td></td>
</tr>
<tr>
<td>Mixed membership models (w/posterior)</td>
<td>best value Bayes model</td>
<td></td>
</tr>
<tr>
<td>Topic models best embedding models</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Document categorization**

- **Accuracy, larger is better**
- **Reciprocal rank for documents**
- **Target corpora beat generic big-data section (except for SOTU, which is very small)**
- **Input vectors**
- **Output vectors**
- **Bias term**

- **Amortized training**
- **Topic model embedding models**

**Vector Composition in Topic Space**

- Nearest topic after composition of mess vectors for words
  - object + recognition
  - character + recognition
  - speech + recognition
  - computer + vision
  - computer + science
- **topics visual object recognition model**
- **character recognition**
- **speech recognition**
- **hmm system hybrid computer vision lese image pattern university science colorado**
- **error training set data performance and computational distribution model matrix**

---

**Background: Word Embeddings**

- **Represent dictionary words with vectors.** Similar words have similar vectors.
- **Simple model, scales easily to large data sets**
- **Can beat deep neural network models**

**Inference for MM Skip-Gram Topic Model**

- **Bayesian inference via collapsed Gibbs sampling**
  \[
  p(z_t = k) \propto \exp\left( \sum_{w \in V} (v(w) \cdot u_k + b_k) \right) 
  \]
- **Scale to many topics:** Metropolis-Hastings-Walker
- **Alias table data structure, amortized O(1) sampling**
- **"Mixture of experts" proposal**
- **Simulated annealing to escape early local maxima**

**Amortized Sublinear Time Training for MM Skip-Gram**

- **Online EM impractical - O(KV) updates**
- **Key insight:** MMSG topic model equivalent to word embedding model *(up to the Dirichlet prior)*
- **Pre-solve E-step via topic model CGS MWMH algorithm**
- **Apply noise-contractive estimation to solve M-step**

---

**Connections to Topic Models, and Mixed Membership Extension to the Skip-Gram**

<table>
<thead>
<tr>
<th>Skip-gram</th>
<th>Mixed membership model</th>
</tr>
</thead>
<tbody>
<tr>
<td>For each word in the corpus ( w ), ( p(w</td>
<td>\theta) \propto \exp\left( \sum_{k \in K} \alpha_k u_k + b_k \right) )</td>
</tr>
</tbody>
</table>

---

**Experimental Results**

**Top Words in Topics**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cat</td>
<td>1246</td>
</tr>
<tr>
<td>2</td>
<td>dog</td>
<td>1136</td>
</tr>
<tr>
<td>3</td>
<td>man</td>
<td>432</td>
</tr>
<tr>
<td>4</td>
<td>king</td>
<td>216</td>
</tr>
<tr>
<td>5</td>
<td>queen</td>
<td>108</td>
</tr>
</tbody>
</table>

**Downstream Tasks: Classification and Regression**

- **Document categorization (classification accuracy; larger is better)**
- **Reciprocal rank for documents (except for SOTU, which is very small)**
- **Target corpora beats generic big-data section (except for SOTU)**
- **Input vectors**
- **Output vectors**
- **Bias term**

- **Amortized training**
- **Topic model embedding models**

---

**Background: Word Embeddings**

- **Represent dictionary words with vectors.** Similar words have similar vectors.
- **Simple model, scales easily to large data sets**
- **Can beat deep neural network models**

**Inference for MM Skip-Gram Topic Model**

- **Bayesian inference via collapsed Gibbs sampling**
  \[
  p(z_t = k) \propto \exp\left( \sum_{w \in V} (v(w) \cdot u_k + b_k) \right) 
  \]
- **Scale to many topics:** Metropolis-Hastings-Walker
- **Alias table data structure, amortized O(1) sampling**
- **"Mixture of experts" proposal**
- **Simulated annealing to escape early local maxima**

**Amortized Sublinear Time Training for MM Skip-Gram**

- **Online EM impractical - O(KV) updates**
- **Key insight:** MMSG topic model equivalent to word embedding model *(up to the Dirichlet prior)*
- **Pre-solve E-step via topic model CGS MWMH algorithm**
- **Apply noise-contractive estimation to solve M-step**

---

**Connections to Topic Models, and Mixed Membership Extension to the Skip-Gram**

<table>
<thead>
<tr>
<th>Skip-gram</th>
<th>Mixed membership model</th>
</tr>
</thead>
<tbody>
<tr>
<td>For each word in the corpus ( w ), ( p(w</td>
<td>\theta) \propto \exp\left( \sum_{k \in K} \alpha_k u_k + b_k \right) )</td>
</tr>
</tbody>
</table>