Model Comparison For Semantic Grouping

Francisco Vargas & Kamen Brestnichki
Problem statement

Given two sentences, how similar would you say they are from 0 to 5? Examples:

- The activity of learning or being trained **vs** The gradual process of acquiring knowledge - **4.0**
- The act of designating a role to someone **vs** The act of designating or identifying something - **1.8**

How do we quantify the odds of two sentences being in the same group?
Modelling (Bag of Word Embeddings)

We contrast two models — one that assumes both sentences were drawn from the same distribution, and one that assumes they were drawn from separate ones.
Examples of Similarities

- **Bayes Factor** - Integrates out Parameters

  \[
  \text{sim}(D_1, D_2) = \log \frac{p(D_1, D_2 | M_1)}{p(D_1 | M_2)p(D_2 | M_2)}.
  \]

  \[
  p(D_1, D_2 | M_1) = \int \prod_{w_k \in D_1 \cup D_2} p(w_k | \theta)p(\theta) d\theta,
  \]

  \[
  p(D_i | M_2) = \int \prod_{w_k \in D_i} p(w_k | \theta)p(\theta) d\theta,
  \]

- **Information TheoreticCriterion (ITC)** - Fits Parameters via MLE

  \[
  \text{sim}(D_1, D_2) = \alpha \left( \hat{\mathcal{L}}(\hat{\theta}_{1,2} | M_1) - (\hat{\mathcal{L}}(\hat{\theta}_1 | M_2) + \hat{\mathcal{L}}(\hat{\theta}_2 | M_2)) \right) + P
  \]

  where $P$ is some penalty for $M_2$ which has double the number of parameters.
Assumptions and Likelihoods

If word embedding length is noise, we can model unit-normed embeddings through the von Mises-Fisher (vMF) distribution.

\[
p(w|\mu, \kappa) = \frac{\kappa^{\frac{d}{2} - 1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2} - 1}(\kappa)} \exp \left( \kappa \mu^\top w \right)
\]

\[
= \frac{1}{Z(\kappa)} \exp \left( \kappa \mu^\top w \right),
\]

Alternatively, if we word embedding length brings important information we may choose to model with the Gaussian distribution.

\[
p(w|\mu, \Sigma) = \mathcal{N}(w|\mu, \Sigma)
\]
Results of our methods on STS

- Gaussian likelihood gives better results than vMF

<table>
<thead>
<tr>
<th>Embedding</th>
<th>Method</th>
<th>STS12</th>
<th>STS13</th>
<th>STS14</th>
<th>STS15</th>
<th>STS16</th>
</tr>
</thead>
<tbody>
<tr>
<td>FastText</td>
<td>vMF+TIC</td>
<td>0.5219</td>
<td>0.5147</td>
<td>0.5719</td>
<td>0.6456</td>
<td>0.6347</td>
</tr>
<tr>
<td></td>
<td>Diag+AIC</td>
<td><strong>0.6193</strong></td>
<td><strong>0.6334</strong></td>
<td><strong>0.6721</strong></td>
<td><strong>0.7328</strong></td>
<td><strong>0.7518</strong></td>
</tr>
<tr>
<td>GloVe</td>
<td>vMF+TIC</td>
<td>0.5421</td>
<td>0.5598</td>
<td>0.5736</td>
<td>0.6474</td>
<td>0.6168</td>
</tr>
<tr>
<td></td>
<td>Diag+AIC</td>
<td><strong>0.6031</strong></td>
<td><strong>0.6131</strong></td>
<td><strong>0.6445</strong></td>
<td><strong>0.7171</strong></td>
<td><strong>0.7346</strong></td>
</tr>
<tr>
<td>Word2Vec GN</td>
<td>vMF+TIC</td>
<td>0.5665</td>
<td>0.5735</td>
<td>0.6062</td>
<td>0.6681</td>
<td>0.6510</td>
</tr>
<tr>
<td></td>
<td>Diag+AIC</td>
<td><strong>0.5957</strong></td>
<td><strong>0.6358</strong></td>
<td><strong>0.6614</strong></td>
<td><strong>0.7213</strong></td>
<td><strong>0.7187</strong></td>
</tr>
</tbody>
</table>

- Outperforms SIF on
  - Glove
  - GN-Word2Vec

- Marginally underperforms SIF on
  - FastText
THANK YOU

Method details at Pacific Ballroom #219