Model Comparison For Semantic Grouping

Francisco Vargas & Kamen Brestnichki

Problem statement

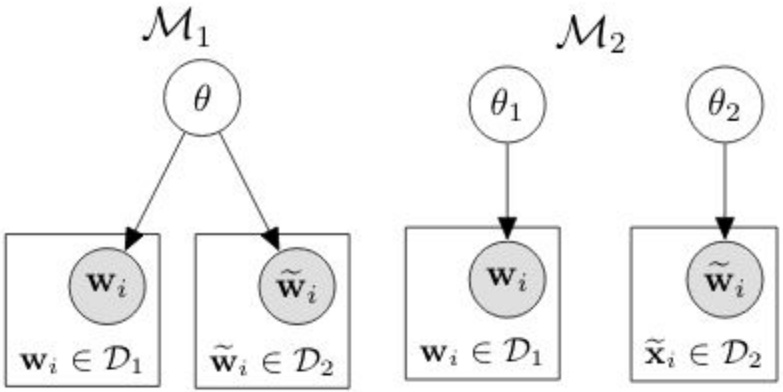
Given two sentences, how similar would you say they are from **0** to **5**? Examples:

- 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**

$$\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) = \log \frac{p(\mathcal{D}_1, \mathcal{D}_2 | \mathcal{M}_1)}{p(\mathcal{D}_1 | \mathcal{M}_2) p(\mathcal{D}_2 | \mathcal{M}_2)}.$$

$$p(\mathcal{D}_1, \mathcal{D}_2 | \mathcal{M}_1) = \int \prod_{\boldsymbol{w}_k \in \mathcal{D}_1 \bigoplus \mathcal{D}_2} p(\boldsymbol{w}_k | \theta) p(\theta) d\theta,$$

$$p(\mathcal{D}_i | \mathcal{M}_2) = \int \prod_{\boldsymbol{w}_k \in \mathcal{D}_i} p(\boldsymbol{w}_k | \theta) p(\theta) d\theta,$$

- Information Theoretic Criterion (ITC) - **Fits Parameters via MLE**

$$\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) = \alpha \left(\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}_{1,2} | \mathcal{M}_1) - (\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}_1 | \mathcal{M}_2) + \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}_2 | \mathcal{M}_2)) \right) + P$$

where P is some penalty for $\mathcal{M}_2$ 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.

$$
\begin{aligned}
p(\boldsymbol{w}|\boldsymbol{\mu}, \kappa) &= \frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa)} \exp\left(\kappa \boldsymbol{\mu}^\top \boldsymbol{w}\right) \\
&= \frac{1}{Z(\kappa)} \exp\left(\kappa \boldsymbol{\mu}^\top \boldsymbol{w}\right),
\end{aligned}
$$

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

$$
p\left(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \mathcal{N}\left(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)
$$

Results of our methods on STS

- Gaussian likelihood gives better results than vMF

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

- Outperforms SIF on
  - Glove
  - GN-Word2Vec
- Marginally underperforms SIF on
  - FastText

Embedding	Method	STS12	STS13	STS14	STS15	STS16
FastText	Diag+AIC	**0.6193**	0.6334	**0.6721**	0.7328	**0.7518**
	SIF	0.6079	**0.6989**	**0.6777**	**0.7436**	0.7135
	MWV	0.5994	0.6494	0.6473	0.7114	0.6814
	WMD	0.5576	0.5146	0.5915	0.6800	0.6402
GloVe	Diag+AIC	**0.6031**	0.6131	**0.6445**	**0.7171**	**0.7346**
	SIF	0.5774	**0.6319**	0.6135	0.6740	0.6589
	MWV	0.5526	0.5643	0.5625	0.6314	0.5804
	WMD	0.5516	0.5007	0.5811	0.6704	0.6246
Word2Vec GN	Diag+AIC	**0.5957**	0.6358	**0.6614**	**0.7213**	**0.7187**
	SIF	0.5697	**0.6594**	**0.6669**	**0.7261**	0.6952
	MWV	0.5744	0.6330	0.6561	0.7040	0.6617
	WMD	0.5554	0.5250	0.6074	0.6730	0.6399

THANK YOU

Method details at **Pacific Ballroom #219**