Unsupervised Random Walk Sentence Embeddings: A Strong but Simple Baseline

Kawin Ethayarajh[1]

[1]*University of Toronto, Canada.*

Computer Science
UNIVERSITY OF TORONTO

Motivation

Smoothed Inverse Frequency (SIF)

- Arora et al. (2017) proposed a sentence embedding based on the idea that words are generated by the random walk of a "discourse vector". This proved to be a strong baseline.
- They replaced the sequence of discourse vectors $\{c_t\}$ with a single vector c_s. Words could also be produced by chance or by a "common discourse vector" c_0 responsible for frequent words:

$$\langle c_0, c_s, p(w), c_s, c_s \rangle \longrightarrow \textit{The quick brown fox jumps.}$$

- The MAP estimate of a sentence embedding c_s for a sentence s with words $\{w\}$ is calculated in two stages, **SIF weighting** (W) and **common component removal** (R):

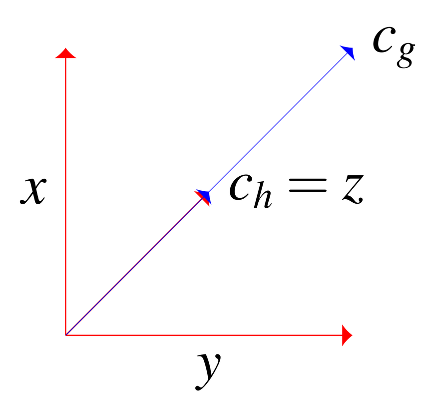
$$\text{W:} \quad \tilde{c}_s = \frac{1}{|s|} \sum_{w \in s} \frac{a}{p(w) + a} \cdot v_w$$

$$\text{R:} \quad c_s = \tilde{c}_s - \text{proj}_{c_0} \tilde{c}_s$$

where a is a hyperparameter, $p(w)$ is the word frequency, and the first singular vector of all $\{\tilde{c}_s\}$ is used as the estimate for c_0.

Shortcomings of SIF

- Due to the log-linear word production model (i.e., $p(w|c_t) \propto \exp(\langle v_w, c_t \rangle)$), word vector length has a confounding effect.



For example, despite $h = \langle z, z \rangle$ and $g = \langle x, y \rangle$, $p(h|c_h) \approx p(g|c_g)$, simply because $\|x\| = \|y\| > \|z\|$.

- There is a hyperparameter a that needs tuning, which requires labelled data.

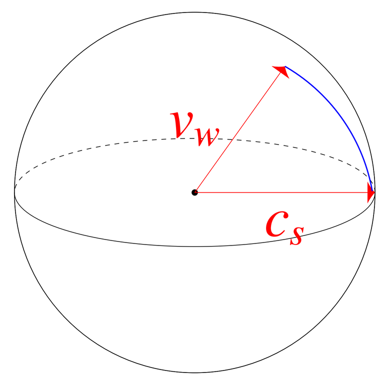
Approach

Angular Distance-based Word Production

- We replace the underlying log-linear word production model with an angular distance-based one:

$$p(w|c_t) \propto 1 - \frac{\arccos(\cos(v_w, c_t))}{\pi}$$

- The angular distance between two vectors is equivalent to the geodesic distance between them on the unit sphere:



Unsupervised Smoothed Inverse Frequency (uSIF)

- The MAP estimate of c_s is calculated in two stages, **uSIF weighting** (U) and **partial common component removal** (P):

$$\text{U:} \quad \tilde{c}_s = \frac{1}{|s|} \sum_{w \in s} \frac{a}{p(w) + \frac{1}{2}a} \cdot v_w$$

$$\text{P:} \quad c_s = \tilde{c}_s - \sum_{i=1}^{m} \lambda_i \, \text{proj}_{c_i'} \tilde{c}_s$$

where $a, \{\lambda_1, ..\lambda_m\}$ are hyperparameters and $\{c_1', \ldots, c_m'\}$ are m common discourse vectors.

- $\{c_1', \ldots, c_m'\}$ are estimated as the first m singular vectors of all $\{\tilde{c}_s\}$ and λ_i is estimated as the proportion of variance explained by its corresponding singular vector c_i'.
- Hyperparameter a can also be estimated directly, using the word frequency, average sentence length n, and vocabulary size $|\mathcal{V}|$:

$$\alpha = \frac{\sum_{w \in \mathcal{V}} \mathbb{1}\left[p(w) > 1 - \left(1 - \frac{1}{|\mathcal{V}|}\right)^n\right]}{|\mathcal{V}|}$$

$$a = \frac{2(1-\alpha)}{\alpha |\mathcal{V}|}$$

Results

- Average results on textual similarity (Pearson's $r \times 100$), sentiment classification, and entailment tasks. The highest score in each column is in bold.

Model	STS'12	STS'13	STS'14	STS'15	SICK14
Wieting et al. (2016) - unsupervised					
PP	58.7	55.8	70.9	75.8	71.6
PP-XXL	61.5	58.9	73.1	77.0	72.7
Arora et al. (2017) - weakly supervised					
GloVe+WR	56.2	56.6	68.5	71.7	72.2
PSL+WR	59.5	61.8	73.5	76.3	72.9
Conneau et al. (2017) - unsupervised (transfer learning)					
InferSent (AllSNLI)	58.6	51.5	67.8	68.3	-
InferSent (SNLI)	57.1	50.4	66.2	65.2	-
Wieting and Gimpel (2017) - unsupervised					
ParaNMT BiLSTM Avg.	67.4	60.3	76.4	79.7	-
ParaNMT Trigram-Word	67.8	62.7	77.4	**80.3**	-
Our Approach - unsupervised					
GloVe+UP	64.9	63.6	74.4	76.1	73.0
PSL+UP	65.8	65.2	75.9	77.6	72.3
ParaNMT+UP	**68.3**	**66.1**	**78.4**	79.0	**73.5**

Model	SST	SICK-R	SICK-E
ParaNMT BiLSTM AVG (Wieting and Gimpel (2017))	82.8	85.9	83.8
ParaNMT+WR (Arora et al. (2017))	80.5	83.9	80.9
ParaNMT+UP (ours)	80.7	83.8	81.1
BiLSTM-Max (on AllNLI) (Conneau et al. (2017))	84.6	**88.4**	**86.3**
BYTE mLSTM (Radford et al. (2017))	**91.8**	79.2	-

Conclusion

- uSIF with partial common component removal is a tough-to-beat, simple, and completely unsupervised baseline for sentence embeddings.
- Future work may involve using better hyperparameter estimations and incorporating more information into the embedding (e.g., word order).

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *International Conference on Learning Representations*.

John Wieting and Kevin Gimpel. 2017. Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.