Unsupervised Random Walk Sentence Embeddings: A Strong but Simple Baseline

Kawin Ethayarajh

University of Toronto

3rd Workshop on Representation Learning for NLP, ACL 2018

Background

Arora et al. (2017):

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

smoothed inverse frequency (SIF) (W):

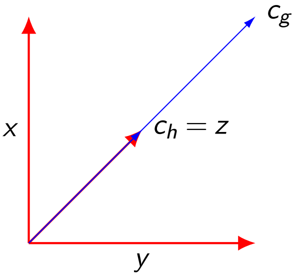
$$\widetilde{c_s} = \frac{1}{|s|} \sum_{w \in s} \frac{a}{p(w) + a} \cdot v_w$$

common component removal (R):

$$c_s = \widetilde{c_s} - \text{proj}_{c_0} \widetilde{c_s}$$

Why not SIF?

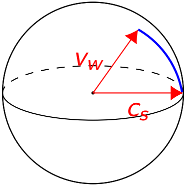
1. log-linear production model → confound of word vector length
   e.g., $h = \langle z, z \rangle$ and $g = \langle x, y \rangle$, but $p(h|c_h) \approx p(g|c_g)$:



2. tuning hyperparameter a requires labelled data

Approach

A word production model that is ~~log-linear.~~ based on angular distance.



unsupervised smoothed inverse frequency (uSIF) (U):

$$\widetilde{c_s} = \frac{1}{|s|} \sum_{w \in s} \frac{a}{p(w) + \frac{1}{2}a} \cdot v_w$$

partial common component removal (P):

$$c_s = \widetilde{c_s} - \sum_{i=1}^{m} \lambda_i \ \mathrm{proj}_{c_i'}$$

Advantages

Angular distance-based production solves both problems:

1. $p(w|c_s)$ not sensitive to $\|v_w\|$
2. can estimate a using $p(w)$, vocabulary size, and average sentence length – no labelled data required!

Results

Model	STS'12	STS'13	STS'14	STS'15	SICK14
Wieting et al. (2015) - unsupervised					
PP-XXL	61.5	58.9	73.1	77.0	72.7
skip-thought	30.8	24.8	31.4	31.0	49.8
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

Results

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
skip-thought (Kiros et al. (2015))	82.0	85.8	82.3
BYTE mLSTM (Radford et al. (2017))	91.8	79.2	-

Conclusion

Unsupervised smoothed inverse frequency (uSIF) with partial common component removal is:

1. a tough-to-beat baseline
2. simple to use
3. completely unsupervised

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.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In Advances in Neural Information Processing Systems, pages 3294–3302.

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

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