Supplementary Materials to “A Generative Word Embedding Model and its Low Rank Positive Semidefinite Solution"

June 14, 2015

Proposition 1. *Suppose two words w_1, w_2 often collocate. In text snippets $\cdots, w_1, w_2, \cdots$ where w_1 is in the focus and w_2 is in the context, and in $\cdots, w_2, w_1, \cdots$, where their roles exchange, there should be common semantic regularities. In the language of statistics, there should be positive correlation between $P(w_1|w_2)$ and $P(w_2|w_1)$. When using different representations for focus words and context words, this correlation is lost.*

Proof. This correlation is guaranteed by the Bayes theorem:

$$\log P(w_1|w_2) = \log P(w_2|w_1) + \log P(w_1) - \log P(w_2),$$

where $\log P(w_1) - \log P(w_2)$ is the non-interactive part.

Suppose fous words and context words use different representations $\boldsymbol{u}_{w_i}$ and $\boldsymbol{v}_{w_i}$, and the two conditional probabilities are modeled as $P(w_1|w_2) = f(\boldsymbol{u}_{w_1}^\top \boldsymbol{v}_{w_2}, a_{w_1}, a_{w_2})$, and $P(w_2|w_1) = f(\boldsymbol{v}_{w_1}^\top \boldsymbol{u}_{w_2}, a_{w_2}, a_{w_1})$, where a_{w_i} is a non-interactive parameter. Then the respective contributions of the embeddings, $\boldsymbol{u}_{w_1}^\top \boldsymbol{v}_{w_2}$ and $\boldsymbol{v}_{w_1}^\top \boldsymbol{u}_{w_2}$, are completely irrelevant, other than being indirectly linked through the corpus statistics. □

Fact 2. *Observations from the trained embeddings of various methods show that, the expectation of the embeddings is close to* 0.

Table 1 lists the expectations of the embeddings of popular methods. The expectation of the embeddings usually satisfies $||E_{P(s)}[\boldsymbol{v}_s]||_1 < \frac{1}{3} E_{P(s)}[||\boldsymbol{v}_s||_1]$, i.e. the magnitude of the expected embedding is much shorter than the expected

Table 1: The expectations of the embeddings of word2vec, GloVe, Forest, PSD.

	word2vec	GloVe	Forest	PSD
$\|\|E_{P(s)}[\boldsymbol{v}_s]\|\|_1$	9.68	44.51	361.54	4.82
$E_{P(s)}[\|\|\boldsymbol{v}_s\|\|_1]$	31.38	113.46	1711.4	39.02

embedding magnitude. This shows that the embeddings of different words point at "random" directions in the embedding space, and cancel each other to a large extent. This observation is natural since in the optimization objective, there is no explicit constraint on the directions of the embeddings.

In "word2vec" [1], there is a normalizing constant for each focus word s_j in the embedding function. This is equivalent to adopting a constant residue r_{s_j}. In GloVe [2], the residue of s_i, s_j is a linear combination of the residues of words s_i and s_j. However we will prove that a constant residue, or even the linear combination of the residues of s_i and s_j, could not satisfy Bayes's theorem, given the assumption that $E_{P(s)}[\boldsymbol{v}_s] \approx 0$. Therefore we have to generalize the residue to a residue for each bigram s_i, s_j, denoted as $a_{s_i s_j}$.

Definition 3. Definitions of mutual information, redundant information, and interaction information.

The mutual information $I(y; x_i)$ and the redundant information $\mathrm{Rdn}(y; x_1, x_2)$ are formally defined as follows.

$$I(y; x_i) = E_{P(x_i,y)}[\log \frac{P(y|x_i)}{P(y)}]$$
$$\mathrm{Rdn}(y; x_1, x_2) = E_{P(y)}\left[\min_{x_1,x_2} E_{P(x_i|y)}[\log \frac{P(y|x_i)}{P(y)}]\right]$$

The synergistic information $\mathrm{Syn}(y; x_1, x_2)$ is defined as the PI-function in [4], skipped here.

The interaction information $\mathrm{Int}(x_1, x_2, y)$ measures the relative strength of $\mathrm{Rdn}(y; x_1, x_2)$ and $\mathrm{Syn}(y; x_1, x_2)$ [**?**]:

$$\begin{aligned}
&\mathrm{Int}(x_1, x_2, y) \\
=&\mathrm{Syn}(y; x_1, x_2) - \mathrm{Rdn}(y; x_1, x_2) \\
=&I(y; x_1, x_2) - I(y; x_1) - I(y; x_2) \\
=&E_{P(x_1,x_2,y)}[\log \frac{P(x_1)P(x_2)P(y)P(x_1, x_2, y)}{P(x_1, x_2)P(x_1, y)P(x_2, y)}]
\end{aligned}$$

The pointwise counterparts of these types of information are obtained by simply dropping the expectation operator.

Definition 4. The normalizing function $\mathcal{Z}(\boldsymbol{A}, \boldsymbol{V})$.

$\mathcal{Z}(\boldsymbol{A}, \boldsymbol{V})$ is the normalizing function of $\mathcal{N}_{\mathrm{Fea}(\boldsymbol{G},N)}(\boldsymbol{A}; 0, \boldsymbol{H}) \cdot \mathrm{U}(\mathrm{Sol}(\boldsymbol{V}; \boldsymbol{A}))$:

$$\begin{aligned}
&\mathcal{Z}(\boldsymbol{A}, \boldsymbol{V}) \\
=&\int_{\mathrm{Fea}(\boldsymbol{G},N)} \exp\{-||\boldsymbol{A}||^2_{HF}\} \cdot \lambda(\mathrm{Sol}(\boldsymbol{V}; \boldsymbol{A}))d\boldsymbol{A},
\end{aligned}$$

where $\lambda(\mathrm{Sol}(\boldsymbol{V};\boldsymbol{A}))$ is a Lebesgue measure of $\mathrm{Sol}(\boldsymbol{V};\boldsymbol{A})$. Note $\lambda(\mathrm{Sol}(\boldsymbol{V};\boldsymbol{A}))$ changes with $\boldsymbol{A}$.

Theorem 5. *Eq. (10), the reduction of learning objective in the manuscript, after integrating $p(\boldsymbol{D},\boldsymbol{A},\boldsymbol{V})$ over Sol$(\boldsymbol{V};\boldsymbol{A})$:*

$$\begin{aligned}
&\boldsymbol{A}^*, \mathrm{Sol}(\boldsymbol{V};\boldsymbol{A}^*)\\
&= \arg\max_{\boldsymbol{A},\boldsymbol{V}} p(\boldsymbol{D},\boldsymbol{A},\mathrm{Sol}(\boldsymbol{V};\boldsymbol{A}))\\
&= \arg\max_{\boldsymbol{A},\boldsymbol{V}} \int_{\mathrm{Sol}(\boldsymbol{V};\boldsymbol{A})} \frac{C_2 \cdot \exp\{-||\boldsymbol{A}||_{HF}^2\}}{\lambda(\mathrm{Sol}(\boldsymbol{V};\boldsymbol{A}))} d\boldsymbol{V}\\
&= \arg\max_{\boldsymbol{A},\boldsymbol{V}} \exp\{-||\boldsymbol{A}||_{HF}^2\}\\
&= \arg\min_{\boldsymbol{A},\boldsymbol{V}} ||\boldsymbol{G} - \boldsymbol{V}^\top \boldsymbol{V}||_{HF}
\end{aligned}$$

References

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, *Distributed representations of words and phrases and their compositionality*, in Proceedings of NIPS 2013, 2013, pp. 3111–3119.

[2] J. Pennington, R. Socher, and C. D. Manning, *Glove: Global vectors for word representation*, Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014), 12 (2014).

[3] N. Timme, W. Alford, B. Flecker, and J. M. Beggs, *Synergy, redundancy, and multivariate information measures: an experimentalist's perspective*, Journal of computational neuroscience, 36 (2014), pp. 119–140.

[4] P. L. Williams and R. D. Beer, *Nonnegative decomposition of multivariate information*, CoRR, abs/1004.2515 (2010).

[5] D. Yogatama, M. Faruqui, C. Dyer, and N. A. Smith, *Learning word representations with hierarchical sparse coding*, in Proceedings of ICML, 2015.