

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

June 7, 2015

**Proposition 1.** *Suppose two words  $w_1, w_2$  often collocate. In text snippets  $\dots, w_1, w_2, \dots$  where  $w_1$  is in the focus and  $w_2$  is in the context, and in  $\dots, w_2, w_1, \dots$ , where their roles exchange, there should be common semantic regularities. 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 focus words and context words use different representations  $\mathbf{u}_{w_i}$  and  $\mathbf{v}_{w_i}$ , and the two conditional probabilities are modeled as  $P(w_1|w_2) = f(\mathbf{u}_{w_1}^\top \mathbf{v}_{w_2}, a_{w_1}, a_{w_2})$ , and  $P(w_2|w_1) = f(\mathbf{v}_{w_1}^\top \mathbf{u}_{w_2}, a_{w_2}, a_{w_1})$ , where  $a_{w_i}$  is a non-interactive parameter. Then the respective contributions of the embeddings,  $\mathbf{u}_{w_1}^\top \mathbf{v}_{w_2}$  and  $\mathbf{v}_{w_1}^\top \mathbf{u}_{w_2}$ , are completely irrelevant, other than being indirectly linked through the corpus statistics.  $\square$

**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  $\|E_{P(s)}[\mathbf{v}_s]\|_1 < \frac{1}{3}E_{P(s)}[\|\mathbf{v}_s\|_1]$ , 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
$\ E_{P(s)}[\mathbf{v}_s]\ _1$	9.68	44.51	361.54	4.82
$E_{P(s)}[\ \mathbf{v}_s\ _1]$	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  $E_{P(s)}[\mathbf{v}_s] \approx 0$ . 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  $\text{Rdn}(y; x_1, x_2)$  are formally defined as follows.

$$I(y; x_i) = E_{P(x_i, y)}[\log \frac{P(y|x_i)}{P(y)}]$$

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

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

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

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

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

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

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

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

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

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

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

## 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 Empirical 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.