University of Stuttgart
Germany

Second-order Co-occurrence Sensitivity of Skip-Gram with Negative Sampling

November 5, 2020

Dominik Schlechtweg, Cennet Oguz, Sabine Schulte im Walde
Institute for Natural Language Processing, University of Stuttgart, Germany

Second-order co-occurrence

First-order co-occurrence vectors

represent a word *w* by a vector of the **counts of context words** it directly co-occurs with

Second-order co-occurrence vectors (Schütze, 1998)

represent a word *w* by a **count vector of the context words of the context words**, i.e., the second-order context words of *w*

Second-order co-occurrence vectors (Schütze, 1998)

- **less sparse** and more **robust** than first-order vectors
- helpful where first-order information is a rare or biased
- can be seen as a way of **generalization**

Example

(1) As far as the Soviet Communist **Party** and the Comintern were concerned ...

(2) ... this is precisely the approach taken by the British **Government**.

(3) The Communist authorities hated rock culture ...

(4) ... rather than risk deportation to British authorities.

Second-order co-occurrence vectors (Schütze, 1998)

- **less sparse**, more **robust**, **generalization**

$\rightarrow$ capturing second-order information improves performance

Vector Space Models

Traditional Count

✘ Count, PPMI do not capture second-order co-occurrence information, but can be modified to do so (✓)

Traditional Embeddings

✓ Truncated SVD does capture second-order co-occurrence information (Kontostathis & Pottenger, 2002)

Modern Embeddings

? SGNS, GloVe, FastText

We compare

- ✘ Positive Pointwise Mutual Information (PPMI)
- ✓ Truncated Singular Value Decomposition (SVD)
- ? Skip-Gram with Negative Sampling (SGNS)

PPMI

Pointwise Mutual Information

$$pmi(w;c) = \log \frac{p(w,c)}{p(w)p(c)}$$

Truncated SVD

$$
\begin{aligned}
M^{\mathrm{PPMI}} &= U \Sigma V^{\top} \\
M^{\mathrm{SVD}} &= U_d \Sigma_d
\end{aligned}
$$

SGNS

Training objective

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c)\in D'} \log \sigma(-v_c \cdot v_w)$$

Training

a c
c a
c b
b c
b d
d e

Training pairs

Experiment 1: Simulating context overlap

1. **first-order overlap (1st)**:
   $=$ same context words in first, $\neq$ distinct context words in second order
2. **2nd-order overlap (2nd)**:
   $\neq$ distinct context words in first, $=$ same context words in second order
3. **no overlap (none)**:
   $\neq$ distinct context words in first, $\neq$ distinct context words in second order

Experiment 1: Simulating first/second-order context overlap

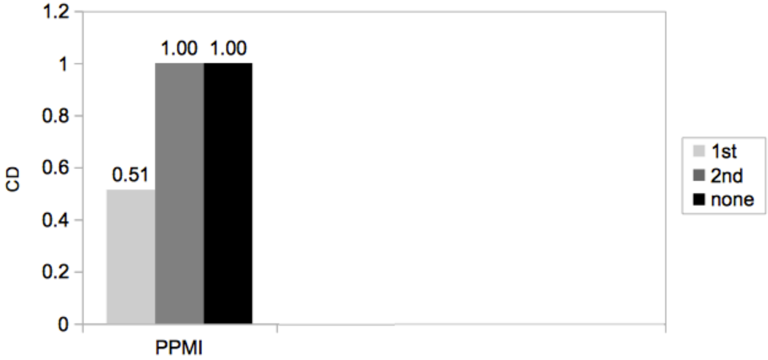
order	1st	2nd	none
C1	a **c**	a c	a c
	a **d**	a d	a d
	b **c**	b e	b e
	b **d**	b f	b f
C2	c u	c **u**	c u
	c v	c **v**	c v
	d w	d **u**	d w
	d x	d **v**	d x

Experiment 1: Simulating context overlap

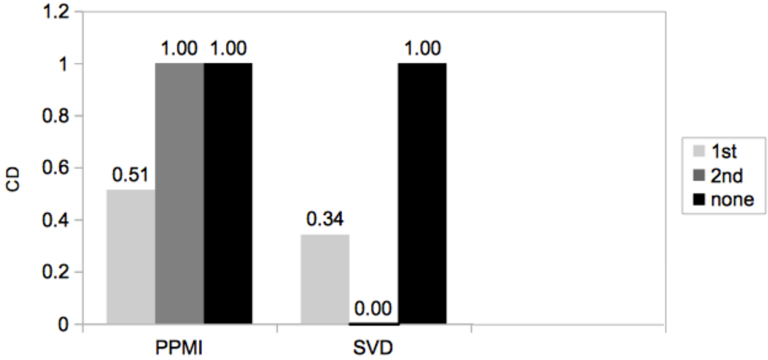
Hypothesis

SGNS and SVD will predict target words from the **2nd-group to be more similar on average than target words from the none-group** (although both groups have no first-order context overlap), while PPMI will predict similar averages for both groups.

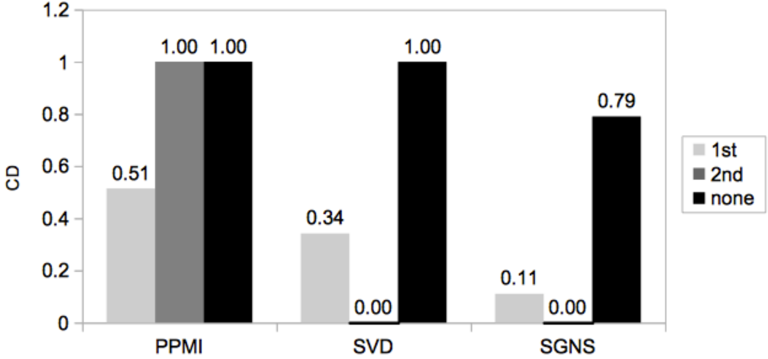
Results



Results



Results



1.2
1
0.8
0.6
0.4
0.2
0
CD
1.00
1.00
0.51
0.34
0.00
1.00
0.11
0.00
0.79
PPMI
SVD
SGNS
1st
2nd
none

Experiment 2: Propagating second-order co-occurrence information

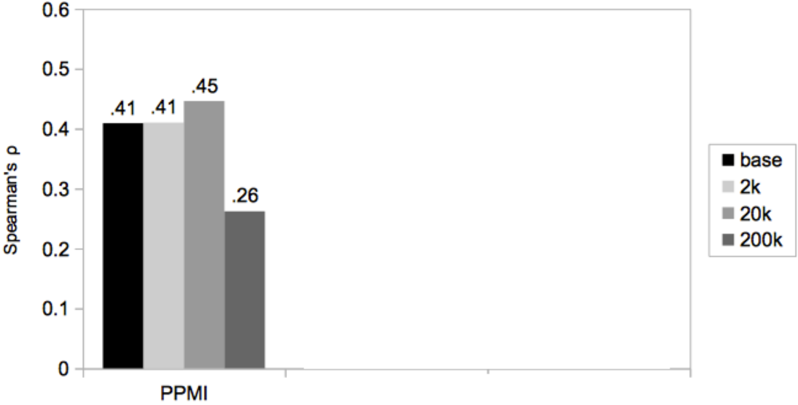
1. create very small corpus (10M tokens from ukWaC)
2. extract first-and second-order word-context pairs
3. add second to first-order pairs for low-frequency words
4. compare performance (WordSim353) on first-order vs. mixed training pairs

Experiment 2: Propagating second-order co-occurrence information

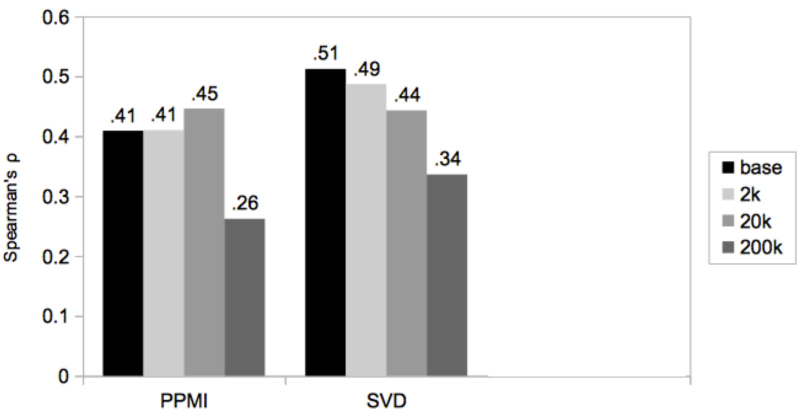
Hypothesis

Additional second-order information will **impact PPMI representations positively and stronger than SVD and SGNS**, because the latter already capture second-order information.

Results

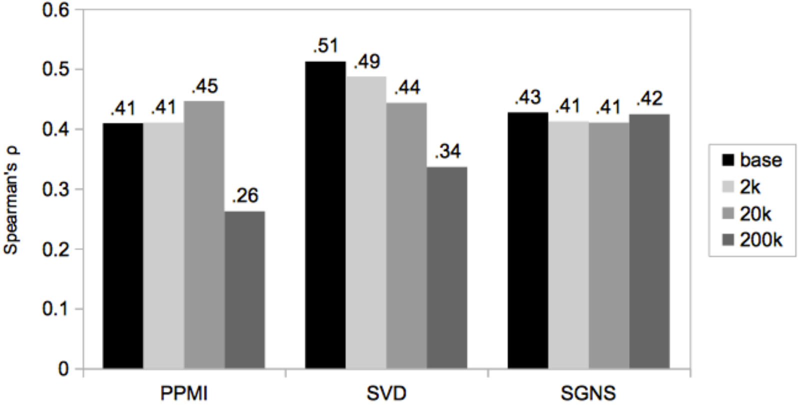


Results



0.6
0.5
0.4
0.3
0.2
0.1
0
Spearman's ρ
.41
.41
.45
.26
.51
.49
.44
.34
PPMI
SVD
base
2k
20k
200k

Results


0.6
0.5
0.4
0.3
0.2
0.1
0
Spearman's ρ
.41
.41
.45
.26
.51
.49
.44
.34
.43
.41
.41
.42
PPMI
SVD
SGNS
base
2k
20k
200k

Explanation: SGNS

$$W = \begin{array}{lrrr} \textbf{Banana} & 1 & -2 & 3 \\ \textbf{Watermelon} & -3 & 2 & 1 \\ \ldots & & & \end{array} \qquad C = \begin{array}{lrrr} \ldots & & & \\ \ldots & & & \\ \textbf{eat} & 2 & 3 & -1 \end{array}$$

SGNS

$$W = \begin{array}{lccc} \textbf{Banana} & 2 & -1 & 2 \\ \textbf{Watermelon} & -3 & 2 & 1 \\ \ldots & & & \end{array} \qquad C = \begin{array}{lccc} & \ldots & & \\ & \ldots & & \\ \textbf{eat} & 2 & 2 & 0 \end{array}$$

SGNS

$$W = \begin{array}{lccc} \textbf{Banana} & 2 & -1 & 2 \\ \boxed{\textbf{Watermelon}} & -2 & 2 & 1 \\ \ldots & & & \end{array} \qquad C = \begin{array}{cccc} \ldots & & & \\ \ldots & & & \\ \textbf{eat} & 1 & 2 & 1 \end{array}$$

SGNS

$$W = \begin{array}{cccc} \textbf{Banana} & 1 & 0 & 1 \\ \textbf{Watermelon} & -2 & 2 & 1 \\ \dots & & & \end{array} \qquad C = \begin{array}{cccc} & \dots & & \\ & \dots & & \\ \textbf{eat} & 1 & 1 & 1 \end{array}$$

SGNS

$$W = \begin{array}{cccc} \textbf{Banana} & 1 & 0 & 1 \\ \boxed{\textbf{Watermelon} \quad -1 \quad 1 \quad 1} \\ \dots \end{array} \qquad C = \begin{array}{c} \dots \\ \dots \\ \boxed{\textbf{eat} \quad 1 \quad 1 \quad 1} \end{array}$$

SGNS

$$W = \begin{array}{cccc} \textbf{Banana} & 1 & 1 & 1 \\ \textbf{Watermelon} & -1 & 1 & 1 \\ \dots & & & \end{array} \qquad C = \begin{array}{cccc} & \dots & & \\ & \dots & & \\ \textbf{eat} & 1 & 1 & 1 \end{array}$$

SGNS

$$W = \begin{array}{cccc} \textbf{Banana} & 1 & 1 & 1 \\ \boxed{\textbf{Watermelon} \quad 0 \quad 1 \quad 1} \\ \ldots \end{array} \qquad C = \begin{array}{cccc} \ldots \\ \ldots \\ \boxed{\textbf{eat} \quad 1 \quad 1 \quad 1} \end{array}$$

SGNS

$$W = \begin{array}{cccc} \textbf{Party} & 1 & -2 & 3 \\ \textbf{Government} & 3 & 2 & -1 \\ \dots & & & \\ \dots & & & \\ \textbf{authorities} & 2 & -3 & 1 \end{array} \qquad C = \begin{array}{cccc} \dots & & & \\ \dots & & & \\ \textbf{Communist} & -1 & 2 & 3 \\ \textbf{British} & 3 & 2 & -1 \\ \dots & & & \end{array}$$

SGNS

$$W = \begin{array}{cccc} \textbf{Party} & 1 & -2 & 3 \\ \textbf{Government} & 3 & 2 & -1 \\ \ldots & & & \\ \ldots & & & \\ \textbf{authorities} & 1 & -2 & 2 \end{array} \qquad C = \begin{array}{cccc} \ldots & & & \\ \ldots & & & \\ \textbf{Communist} & 0 & 1 & 2 \\ \textbf{British} & 3 & 2 & -1 \\ \ldots & & & \end{array}$$

SGNS

$$W = \begin{array}{cccc} \textbf{Party} & 1 & -2 & 3 \\ \textbf{Government} & 3 & 2 & -1 \\ \dots & & & \\ \dots & & & \\ \textbf{authorities} & 2 & -1 & 1 \end{array} \qquad C = \begin{array}{cccc} \dots & & & \\ \dots & & & \\ \textbf{Communist} & 0 & 1 & 2 \\ \textbf{British} & 2 & 1 & 0 \\ \dots & & & \end{array}$$

SGNS

$$W = \begin{array}{cccc} \textbf{Party} & 1 & -2 & 3 \\ \textbf{Government} & 3 & 2 & -1 \\ \dots & & & \\ \dots & & & \\ \textbf{authorities} & 1 & 0 & 1 \end{array} \qquad C = \begin{array}{cccc} \dots & & & \\ \dots & & & \\ \textbf{Communist} & 1 & 0 & 2 \\ \textbf{British} & 2 & 1 & 1 \\ \dots & & & \end{array}$$

SGNS

$$W = \begin{array}{cccc} \textbf{Party} & 1 & -2 & 3 \\ \textbf{Government} & 3 & 2 & -1 \\ \dots & & & \\ \dots & & & \\ \textbf{authorities} & 1 & 0 & 1 \end{array} \qquad C = \begin{array}{cccc} \dots & & & \\ \dots & & & \\ \textbf{Communist} & 1 & 0 & 2 \\ \textbf{British} & 1 & 0 & 1 \\ \dots & & & \end{array}$$

SGNS

$$W = \begin{array}{lrrr} \textbf{Party} & 1 & -1 & 2 \\ \textbf{Government} & 3 & 2 & -1 \\ \ldots & & & \\ \ldots & & & \\ \textbf{authorities} & 1 & 0 & 1 \end{array} \qquad C = \begin{array}{lrrr} \ldots & & & \\ \ldots & & & \\ \textbf{Communist} & 1 & -1 & 2 \\ \textbf{British} & 1 & 0 & 1 \\ \ldots & & & \end{array}$$

SGNS

$$W = \begin{array}{cccc} \textbf{Party} & 1 & -1 & 2 \\ \textbf{Government} & 2 & 1 & 0 \\ \ldots & & & \\ \ldots & & & \\ \textbf{authorities} & 1 & 0 & 1 \end{array} \qquad C = \begin{array}{cccc} \ldots & & & \\ \ldots & & & \\ \textbf{Communist} & 1 & -1 & 2 \\ \textbf{British} & 2 & 1 & 0 \\ \ldots & & & \end{array}$$

SGNS

$$W = \begin{array}{cccc} \textbf{Party} & 1 & -1 & 2 \\ \textbf{Government} & 2 & 1 & 0 \\ \ldots & & & \\ \ldots & & & \\ \textbf{authorities} & 1 & 0 & 1 \end{array} \qquad C = \begin{array}{cccc} \ldots & & & \\ \ldots & & & \\ \textbf{Communist} & 1 & 0 & 1 \\ \textbf{British} & 2 & 1 & 0 \\ \ldots & & & \end{array}$$

SGNS

$$W = \begin{array}{lccc} \textbf{Party} & 1 & -1 & 2 \\ \textbf{Government} & 2 & 1 & 0 \\ \dots & & & \\ \dots & & & \\ \textbf{authorities} & 1 & 0 & 1 \end{array} \qquad C = \begin{array}{lccc} \dots & & & \\ \dots & & & \\ \textbf{Communist} & 1 & 0 & 1 \\ \textbf{British} & 1 & 0 & 1 \\ \dots & & & \end{array}$$

SGNS

$$W = \begin{array}{cccc} \boxed{\textbf{Party}} & \boxed{1} & \boxed{0} & \boxed{1} \\ \textbf{Government} & 2 & 1 & 0 \\ \ldots & & & \\ \ldots & & & \\ \textbf{authorities} & 1 & 0 & 1 \end{array} \qquad C = \begin{array}{cccc} \ldots & & & \\ \ldots & & & \\ \boxed{\textbf{Communist}} & \boxed{1} & \boxed{0} & \boxed{1} \\ \textbf{British} & 1 & 0 & 1 \\ \ldots & & & \end{array}$$

SGNS

$$W = \begin{array}{lccc} \textbf{Party} & 1 & 0 & 1 \\ \textbf{Government} & 1 & 1 & 1 \\ \ldots & & & \\ \ldots & & & \\ \textbf{authorities} & 1 & 0 & 1 \end{array} \qquad C = \begin{array}{lccc} \ldots & & & \\ \ldots & & & \\ \textbf{Communist} & 1 & 0 & 1 \\ \textbf{British} & 1 & 1 & 1 \\ \ldots & & & \end{array}$$

Relation between SVD and SGNS

- show similar results (Levy et al., 2015)
- their training objectives have been related to each other (Levy & Goldberg, 2014)
- their correspondence in the low-dimensional case has not been shown yet

→ **if SGNS is implicit SVD, it should be second-order co-occurrence sensitive**

Does this show that SGNS is implicit SVD?

- no
- it just shows that in the low-dimensional case they share **one** fundamental property
- there is evidence that vector spaces learned by low-dimensional SGNS and SVD have other different properties (Shin et al., 2018)

Conclusion

- SGNS captures second-order co-occurrence information, a property it shares with SVD and distinguishes it from PPMI
- variety of algorithms with SGNS architecture
- SGNS became the "traditional model" this year
- so, what about GloVe, ELMo, BERT?
- **how does second-order sensitivity relate to performance?** (Artetxe, Labaka, Lopez-Gazpio, & Agirre, 2018)

Bibliography I

Artetxe, M., Labaka, G., Lopez-Gazpio, I., & Agirre, E. (2018). Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *Proceedings of the 22nd conference on computational natural language learning* (pp. 282–291). Brussels, Belgium: Association for Computational Linguistics.

Kontostathis, A., & Pottenger, W. M. (2002). Detecting patterns in the LSI term-term matrix. In *Proceedings of the workshop on foundations of data mining and discovery.*

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th international conference on neural information processing systems - volume 2* (pp. 2177–2185). Cambridge, MA, USA: MIT Press.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225.

Schütze, H. (1998, March). Automatic word sense discrimination. *Computational Linguistics*, *24*(1), 97–123.

Bibliography II

Shin, J., Madotto, A., & Fung, P. (2018). Interpreting word embeddings with eigenvector analysis. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018), IRASL Workshop.* Montréal, Canada.

Results GloVe

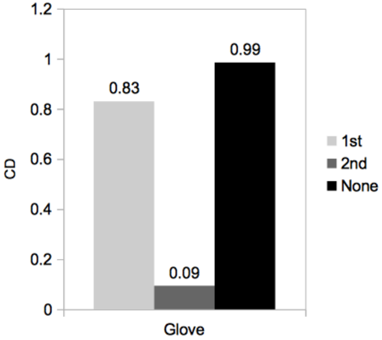


Figure 1: Results of simulation experiment with GloVe embeddings.