Rotated Word Vector Representations and their Interpretability

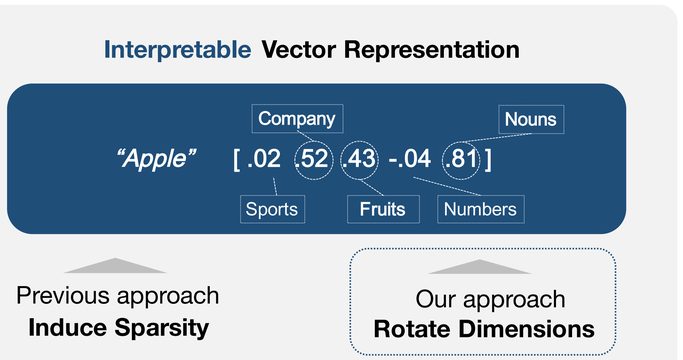
Sungjoon Park, JinYeong Bak, Alice Oh {sungjoon.park, jy.bak}@kaist.ac.kr, alice.oh@kaist.edu

USERS & INFORMATION

KAIST

Introduction

Applying the matrix rotation algorithms from psychometric analysis to word vector representations to improve the interpretability



Benefits of interpretable word vectors

- Understanding semantic / syntactic compositionality of words
- Increasing efficiency of storage
- Reducing complexity of higher-level models

Factor Rotation

Crawford-Ferguson Rotation Family

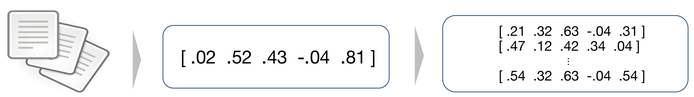
- To compute $\Lambda = A T$, satisfying: $T'T = I$ or $\mathrm{diag}(T^{-1}T^{-1\prime}) = I$ (Λ: rotated, A: original, orthogonal, oblique)
- Minimize: $f(\lambda) = (1-\kappa)\sum_{i=1}^{p}\sum_{j=1}^{m}\sum_{l\neq j,l=1}^{m}\lambda_{ij}^2\lambda_{il}^2 + \kappa\sum_{j=1}^{m}\sum_{i=1}^{p}\sum_{l\neq i,l=1}^{m}\lambda_{ij}^2\lambda_{lj}^2$ (Row complexity) (Column complexity)
- κ :

Quartimax	Varimax	Parsimax	Factor Parsimony
0	1/p	m-1/p+m-2	1

- Algorithm: Gradient Projection (Jennrich, 2001)

source: https://github.com/SungjoonPark/factor_rotation (TensorFlow)

Experimental Settings



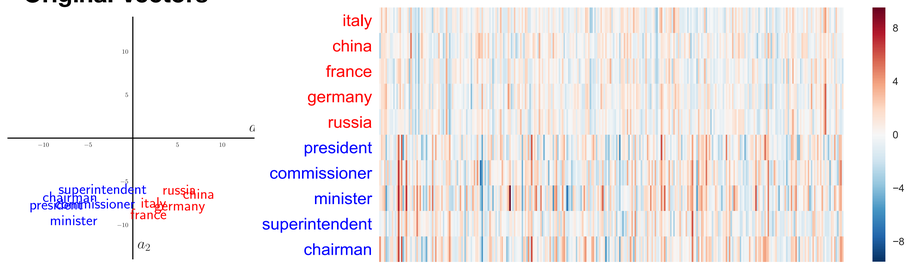
English Wikipedia
- 5.3M articles
- 83M sentences
- 1,676M tokens

2 Word Vectors
- Word2Vec, Glove
- 306,491 words
- 300 dimensions

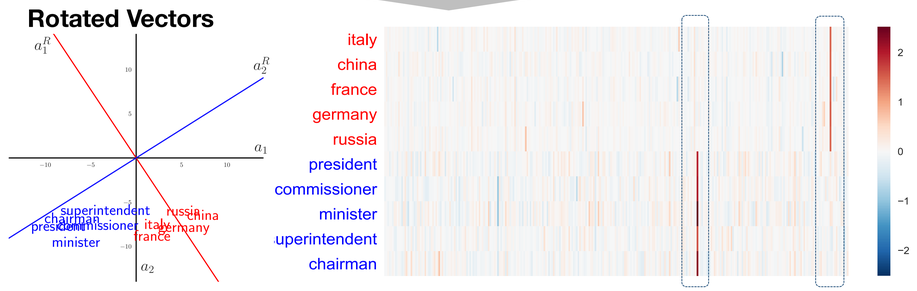
16 Rotated Representations
- For each kappa (4)
- For each Embedding (2)
- For each constraint (2)

Rotated Word Vectors

Original Vectors

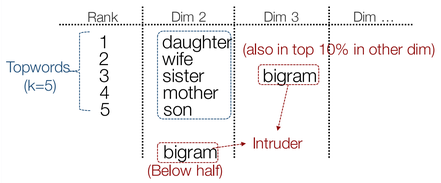


Rotated Vectors



Interpretability

Word Intrusion



Measure: Overall Distance Ratio

- $\mathrm{DR}_{\mathrm{overall}} = \frac{1}{d}\frac{\sum_{a=1}^{d} D_{\mathrm{inter}}^{a}}{\sum_{a=1}^{d} D_{\mathrm{intra}}^{a}}$
- $\mathrm{DR}_{\mathrm{inter}}^{a} = \frac{\sum_{w_i} \mathrm{dist}(w_i, w_{\mathrm{intruder}})}{k}$ (avg. dist btw topwords & intruder)
- $\mathrm{DR}_{\mathrm{intra}}^{a} = \frac{\sum_{w_i}\sum_{w_j} \mathrm{dist}(w_i, w_j)}{k(k-1)}$ (avg. dist btw topwords)

Results

Overall Distance Ratio	SG	Glove
Original	1.258	1.095
SOV (sparse overcomplete vector)	1.089	1.050
SOV (non-neg)	1.081	1.074
Quartimax (orthogonal)	1.479	1.248
Varimax (orthogonal)	1.477	**1.289**
Parsimax (orthogonal)	**1.596**	1.261
FacParsim (orthogonal)	1.300	1.102
Quartimax (oblique)	1.385	**1.225**
Varimax (oblique)	**1.398**	1.222
Parsimax (oblique)	1.386	1.174
FacParsim (oblique)	1.145	1.081

Qualitative Examples

Skip-Gram
- householder, asked, indicted, there, ethnic
- score, two, best, three, four
- mining, footballer, population, laps, settled
- density, census, fourier, editor, photos
- money, toured, season, announced, banned

Rotated Skip-Gram
- twitter, facebook, youtube, myspace, internet
- receptors, receptor, neurons, apoptosis, neuronal
- pennsylvania,ohio,maryland,philadelphia,illinois
- paintings, portraits, painting, drawings, painter
- that, which, when, where, but

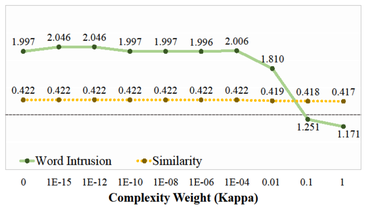
Expressive Performance

NLP tasks Rotated word vectors show comparable performance to that of the SOV and the original.

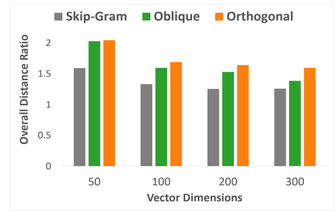
Performance	Word Simil.	Analg. (sem)	Analg. (syn)	Sent.	Ques.	Topics	NP Brackt.
Original	0.374	0.668	**0.652**	0.741	0.920	**0.960**	0.812
SOV	0.390	0.640	0.594	0.751	0.910	0.955	**0.836**
SOV (non-neg)	0.384	0.566	0.480	**0.761**	0.918	**0.960**	0.829
Quartimax (orthogonal)	0.374	0.668	**0.652**	0.744	0.922	0.956	0.822
Varimax (orthogonal)	0.374	0.668	**0.652**	0.744	0.922	0.956	0.822
Parsimax (orthogonal)	0.374	0.668	**0.652**	0.744	0.922	0.956	0.819
FacParsim (orthogonal)	0.374	0.668	**0.652**	0.744	0.922	0.956	0.822
Quartimax (oblique)	**0.422**	**0.673**	0.624	0.755	**0.932**	0.955	0.820
Varimax (oblique)	**0.422**	**0.673**	0.624	0.755	**0.932**	0.955	0.820
Parsimax (oblique)	0.421	0.671	0.623	0.752	**0.932**	0.956	0.826
FacParsim (oblique)	0.417	0.660	0.620	0.751	0.928	0.952	0.820

Understanding Rotated Vectors

Effect of kappa



Effect of # of dimensions



Conclusion

- Observed increased interpretability in both directions and the positive relation between absolute value of the dimension and interpretability.
- Rotation algorithm can be applied to any kind of word embeddings.
- The vectors can be used to
  - Understand what the word vectors are comprised of.
  - Remove irrelevant dimensions for a specific task of interest.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIP) (No. 2016R1A2B4016048)