Uncovering divergent linguistic information in word embeddings

VL Embeddings

Uni Heidelberg

SS 2019

Uncovering linguistic information in word embeddings

Artetxe et al (2018): Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation

- Word embeddings capture more information than we can directly observe
    - We can apply linear transformations to pretrained embeddings to adjust performance to different tasks along the axes of similarity/relatedness and semantics/syntax

Uncovering linguistic information in word embeddings

Artetxe et al (2018): Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation

- Word embeddings capture more information than we can directly observe
  - We can apply linear transformations to pretrained embeddings to adjust performance to different tasks along the axes of similarity/relatedness and semantics/syntax

	FastText	Dep-based embeddings	GloVe
good for	syntactic analogies	functional similarities	semantic analogies

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let X be the word embeddings matrix
- Let X_i be the embedding of the ith word in the vocabulary

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let X be the word embeddings matrix
- Let X_i be the embedding of the ith word in the vocabulary
- The dot product $sim(i, j) = X_i \cdot X_j$ is a measure of the similarity between the ith and the jth word

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let X be the word embeddings matrix
- Let X_i be the embedding of the ith word in the vocabulary
- The dot product $sim(i, j) = X_i \cdot X_j$ is a measure of the similarity between the ith and the jth word

- Define the similarity matrix $M(X) := XX^\top$ so that $sim(i, j) = M(X)_{ij}$ $\Rightarrow$ first order similarity

Linear transformation of embedding matrix

Artetxe et al (2018)

$$X = \begin{array}{lrrr} \text{cat} & -0.19 & 0.45 & -0.40 \\ \text{dog} & -0.28 & 0.43 & -0.39 \\ \text{blue} & 0.02 & -0.40 & -0.39 \\ \text{red} & 0.03 & -0.22 & -0.31 \\ \text{happy} & -0.03 & -0.07 & -0.19 \end{array}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$X = \begin{array}{lrrr} \text{cat} & -0.19 & 0.45 & -0.40 \\ \text{dog} & -0.28 & 0.43 & -0.39 \\ \text{blue} & 0.02 & -0.40 & -0.39 \\ \text{red} & 0.03 & -0.22 & -0.31 \\ \text{happy} & -0.03 & -0.07 & -0.19 \end{array}$$

$$X^\top = \begin{array}{rrrrr} \text{cat} & \text{dog} & \text{blue} & \text{red} & \text{happy} \\ -0.19 & -0.28 & 0.02 & 0.03 & -0.03 \\ 0.45 & 0.43 & -0.40 & -0.22 & -0.07 \\ -0.40 & -0.39 & -0.39 & -0.31 & -0.19 \end{array}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$X = \begin{array}{l} \text{cat} \\ \text{dog} \\ \text{blue} \\ \text{red} \\ \text{happy} \end{array} \begin{array}{ccc} \text{-0.19} & \text{0.45} & \text{-0.40} \\ \text{-0.28} & \text{0.43} & \text{-0.39} \\ \text{0.02} & \text{-0.40} & \text{-0.39} \\ \text{0.03} & \text{-0.22} & \text{-0.31} \\ \text{-0.03} & \text{-0.07} & \text{-0.19} \end{array}$$

$$sim(i, j) = X_i \cdot X_j$$

$$X^\top = \begin{array}{ccccc} \text{cat} & \text{dog} & \text{blue} & \text{red} & \text{happy} \\ \text{-0.19} & \text{-0.28} & \text{0.02} & \text{0.03} & \text{-0.03} \\ \text{0.45} & \text{0.43} & \text{-0.40} & \text{-0.22} & \text{-0.07} \\ \text{-0.40} & \text{-0.39} & \text{-0.39} & \text{-0.31} & \text{-0.19} \end{array}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$X = \begin{array}{lrrr} \text{cat} & -0.19 & 0.45 & -0.40 \\ \text{dog} & -0.28 & 0.43 & -0.39 \\ \text{blue} & 0.02 & -0.40 & -0.39 \\ \text{red} & 0.03 & -0.22 & -0.31 \\ \text{happy} & -0.03 & -0.07 & -0.19 \end{array}$$

$$sim(i,j) = X_i \cdot X_j$$

$$X^\top = \begin{array}{rrrrr} \text{cat} & \text{dog} & \text{blue} & \text{red} & \text{happy} \\ -0.19 & -0.28 & 0.02 & 0.03 & -0.03 \\ 0.45 & 0.43 & -0.40 & -0.22 & -0.07 \\ -0.40 & -0.39 & -0.39 & -0.31 & -0.19 \end{array}$$

$$M(X) := XX^\top$$

$$sim(i,j) = M(X)_{ij}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let X be the word embeddings matrix
- Let X_i be the embedding of the ith word in the vocabulary
- The dot product $sim(i, j) = X_i \cdot X_j$ is a measure of the similarity between the ith and the jth word

- Define the similarity matrix $M(X) := XX^\top$ so that $sim(i, j) = M(X)_{ij}$ $\Rightarrow$ first order similarity

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let X be the word embeddings matrix
- Let X_i be the embedding of the ith word in the vocabulary
- The dot product $sim(i, j) = X_i \cdot X_j$ is a measure of the similarity between the ith and the jth word

- Define the similarity matrix $M(X) := XX^\top$ so that $sim(i, j) = M(X)_{ij}$ $\Rightarrow$ first order similarity

- We can also define a second order similarity measure:
  - first order: How similar are w_i and w_j?
  - second order: How similar are the contexts of w_i and w_j?

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let X be the word embeddings matrix
- Let X_i be the embedding of the ith word in the vocabulary
- The dot product $sim(i, j) = X_i \cdot X_j$ is a measure of the similarity between the ith and the jth word

- Define the similarity matrix $M(X) := XX^\top$ so that $sim(i, j) = M(X)_{ij}$ $\Rightarrow$ first order similarity

- We can also define a second order similarity measure:
  - first order: How similar are w_i and w_j?
  - second order: How similar are the contexts of w_i and w_j?
- We can even define a third, fourth or nth order similarity

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let X be the word embeddings matrix
- Let X_i be the embedding of the ith word in the vocabulary
- The dot product $sim(i, j) = X_i \cdot X_j$ is a measure of the similarity between the ith and the jth word

- Define the similarity matrix $M(X) := XX^\top$ so that $sim(i, j) = M(X)_{ij}$ $\Rightarrow$ first order similarity

- We can also define a second order similarity measure:
  - first order: How similar are w_i and w_j?
  - second order: How similar are the contexts of w_i and w_j?
- We can even define a third, fourth or nth order similarity

Idea: Some higher order similarities might be better at capturing specific aspects of language.

Linear transformation of embedding matrix

Artetxe et al (2018)

- Define the first order similarity matrix as
  $M(X) := XX^\top$ so that $sim(i,j) = M(X)_{ij}$
- Define the second order similarity matrix as
  $M_2(X) := M(M(X))$ so that $sim_2(i,j) = M_2(X)_{ij}$
  where $M_2(X) = XX^\top XX^\top$

Linear transformation of embedding matrix

Artetxe et al (2018)

Definition: $M_2(X) := M(M(X))$

Linear transformation of embedding matrix

Artetxe et al (2018)

Definition: $M_2(X) := M(M(X))$

$$M_2(X) \quad = \quad M(M(X))$$

Linear transformation of embedding matrix

Artetxe et al (2018)

Definition: $M_2(X) := M(M(X))$

$$\begin{aligned} M_2(X) &= M(M(X)) \\ &= M(X)M(X)^\top \end{aligned}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

Definition: $M_2(X) := M(M(X))$

$$
\begin{aligned}
M_2(X) &= M(M(X)) \\
&= M(X)M(X)^\top \\
&= XX^\top(XX^\top)^\top
\end{aligned}
$$

$$\boxed{(AB)^\top = B^\top A^\top}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

Definition: $M_2(X) := M(M(X))$

$$
\begin{aligned}
M_2(X) &= M(M(X)) \\
&= M(X)M(X)^\top \\
&= XX^\top(XX^\top)^\top \\
&= XX^\top X^{\top\top} X^\top
\end{aligned}
$$

$$\boxed{(AB)^\top = B^\top A^\top}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

Definition: $M_2(X) := M(M(X))$

$$
\begin{aligned}
M_2(X) &= M(M(X)) \\
&= M(X)M(X)^\top \\
&= XX^\top (XX^\top)^\top \\
&= XX^\top X^{\top\top} X^\top \\
&= XX^\top XX^\top
\end{aligned}
$$

$$\boxed{(AB)^\top = B^\top A^\top}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

- Define the second order similarity matrix as
  $M_2(X) := XX^\top XX^\top$ so that $sim_2(i,j) = M_2(X)_{ij}$
  where $M_2(X) = M(M(X))$

Linear transformation of embedding matrix

Artetxe et al (2018)

- Define the second order similarity matrix as
  $M_2(X) := XX^\top XX^\top$ so that $sim_2(i,j) = M_2(X)_{ij}$
  where $M_2(X) = M(M(X))$
- Define the n-th order similarity matrix as
  $M_n(X) = (XX^\top)^n$ so that $sim_n(i,j) = M_n(X)_{ij}$

Linear transformation of embedding matrix

Artetxe et al (2018)

- Define the second order similarity matrix as
  $M_2(X) := XX^\top XX^\top$ so that $sim_2(i,j) = M_2(X)_{ij}$
  where $M_2(X) = M(M(X))$
- Define the n-th order similarity matrix as
  $M_n(X) = (XX^\top)^n$ so that $sim_n(i,j) = M_n(X)_{ij}$

Instead of changing the similarity measure, we can also change the word embeddings themselves through a linear transformation so they directly capture this second or $n-$th order similarity

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let $X^\top X = Q\Lambda Q^\top$ be the eigendecomposition of $X^\top X$
  - Λ is a positive diagonal matrix whose entries are the eigenvalues of $X^\top X$ and Q is an orthogonal matrix with their respective eigenvectors as columns

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let $X^\top X = Q \Lambda Q^\top$ be the eigendecomposition of $X^\top X$
  - Λ is a positive diagonal matrix whose entries are the eigenvalues of $X^\top X$ and Q is an orthogonal matrix with their respective eigenvectors as columns
- Define the linear transformation matrix $W = Q \sqrt{\Lambda}$
- Apply W to the original embeddings X : $XW = X = XQ\sqrt{\Lambda}$
- $M(X) = M_2(X)$: transformed embeddings X exploit the second-order similarity as defined for the original embeddings

Linear transformation of embedding matrix

Artetxe et al (2018)

$$\underset{m \times n}{X} = \underset{m \times m}{U} \cdot \underset{m \times n}{\Sigma} \cdot \underset{n \times n}{Q^\top}$$

SVD

Linear transformation of embedding matrix

Artetxe et al (2018)

$$\underset{m \times n}{X} = \underset{m \times m}{U} \cdot \underset{m \times n}{\Sigma} \cdot \underset{n \times n}{Q^\top}$$

SVD

$$X^\top X = (U \cdot \Sigma \cdot Q^\top)^\top \cdot (U \cdot \Sigma \cdot Q^\top)$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$\underset{m \times n}{X} = \underset{m \times m}{U} \cdot \underset{m \times n}{\Sigma} \cdot \underset{n \times n}{Q^\top}$$

SVD

$$X^\top X = (U \cdot \Sigma \cdot Q^\top)^\top \cdot (U \cdot \Sigma \cdot Q^\top)$$

$$= (Q^\top)^\top \cdot \Sigma^\top \cdot U^\top \cdot U \cdot \Sigma \cdot Q^\top$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$\underset{m \times n}{X} = \underset{m \times m}{U} \cdot \underset{m \times n}{\Sigma} \cdot \underset{n \times n}{Q^\top} \qquad \text{SVD}$$

$$X^\top X = (U \cdot \Sigma \cdot Q^\top)^\top \cdot (U \cdot \Sigma \cdot Q^\top)$$

$$= (Q^\top)^\top \cdot \Sigma^\top \cdot U^\top \cdot U \cdot \Sigma \cdot Q^\top \qquad U \text{ orthonormal}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$\underset{m \times n}{X} = \underset{m \times m}{U} \cdot \underset{m \times n}{\Sigma} \cdot \underset{n \times n}{Q^\top} \qquad \text{SVD}$$

$$X^\top X = (U \cdot \Sigma \cdot Q^\top)^\top \cdot (U \cdot \Sigma \cdot Q^\top)$$

$$= (Q^\top)^\top \cdot \Sigma^\top \cdot U^\top \cdot U \cdot \Sigma \cdot Q^\top \qquad U \text{ orthonormal}$$

$$= Q \cdot \Sigma^\top \cdot \Sigma \cdot Q^\top$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$\underset{m \times n}{X} = \underset{m \times m}{U} \cdot \underset{m \times n}{\Sigma} \cdot \underset{n \times n}{Q^\top} \qquad \text{SVD}$$

$$X^\top X = (U \cdot \Sigma \cdot Q^\top)^\top \cdot (U \cdot \Sigma \cdot Q^\top)$$

$$= (Q^\top)^\top \cdot \Sigma^\top \cdot U^\top \cdot U \cdot \Sigma \cdot Q^\top \qquad U \text{ orthonormal}$$

$$= Q \cdot \Sigma^\top \cdot \Sigma \cdot Q^\top \qquad \Sigma \text{ Diagonalmatrix mit } \sqrt{EW}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$\underset{m \times n}{X} = \underset{m \times m}{U} \cdot \underset{m \times n}{\Sigma} \cdot \underset{n \times n}{Q^\top} \qquad \text{SVD}$$

$$X^\top X = (U \cdot \Sigma \cdot Q^\top)^\top \cdot (U \cdot \Sigma \cdot Q^\top)$$

$$= (Q^\top)^\top \cdot \Sigma^\top \cdot U^\top \cdot U \cdot \Sigma \cdot Q^\top \qquad U \text{ orthonormal}$$

$$= Q \cdot \Sigma^\top \cdot \Sigma \cdot Q^\top \qquad \Sigma \text{ Diagonalmatrix mit } \sqrt{EW}$$

$$= Q \cdot \Lambda \cdot Q^\top$$

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let $X^\top X = Q\Lambda Q^\top$ be the eigendecomposition of $X^\top X$
  - Λ is a positive diagonal matrix whose entries are the eigenvalues of $X^\top X$ and Q is an orthogonal matrix with their respective eigenvectors as columns

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let $X^\top X = Q \Lambda Q^\top$ be the eigendecomposition of $X^\top X$
  - Λ is a positive diagonal matrix whose entries are the eigenvalues of $X^\top X$ and Q is an orthogonal matrix with their respective eigenvectors as columns
- Define the linear transformation matrix $W := Q\sqrt{\Lambda}$

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let $X^\top X = Q\Lambda Q^\top$ be the eigendecomposition of $X^\top X$
  - Λ is a positive diagonal matrix whose entries are the eigenvalues of $X^\top X$ and Q is an orthogonal matrix with their respective eigenvectors as columns
- Define the linear transformation matrix $W := Q\sqrt{\Lambda}$
- Apply W to the original embeddings $X \quad \Rightarrow X' = XW$

Linear transformation of embedding matrix

Artetxe et al (2018)

- Let $X^\top X = Q\Lambda Q^\top$ be the eigendecomposition of $X^\top X$
  - Λ is a positive diagonal matrix whose entries are the eigenvalues of $X^\top X$ and Q is an orthogonal matrix with their respective eigenvectors as columns
- Define the linear transformation matrix $W := Q\sqrt{\Lambda}$
- Apply W to the original embeddings $X \quad \Rightarrow X' = XW$

- $M(X') = M_2(X) \quad \Rightarrow$ transformed embeddings X' capture the second order similarity as defined for the original embeddings

Linear transformation of embedding matrix

Artetxe et al (2018)

$$M(X') = X' \cdot X'^{\top}$$ first order similarity of X'

Linear transformation of embedding matrix

Artetxe et al (2018)

$$M(X') = X' \cdot X'^{\top} \qquad \text{first order similarity of } X'$$

$$= X \cdot W \cdot (XW)^{\top}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$M(X') = X' \cdot X'^{\top} \qquad \text{first order similarity of } X'$$

$$= X \cdot W \cdot (XW)^{\top} \qquad (AB)^{\top} = B^{\top} A^{\top}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(X') &= X' \cdot X'^{\top} && \text{first order similarity of } X' \\
&= X \cdot W \cdot (XW)^{\top} && (AB)^{\top} = B^{\top} A^{\top} \\
&= X \cdot W \cdot W^{\top} \cdot X^{\top}
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(X') &= X' \cdot X'^{\top} && \text{first order similarity of } X' \\
&= X \cdot W \cdot (XW)^{\top} && (AB)^{\top} = B^{\top} A^{\top} \\
&= X \cdot W \cdot W^{\top} \cdot X^{\top} && W := Q\sqrt{\Lambda}
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(X') &= X' \cdot X'^{\top} && \text{first order similarity of } X' \\
&= X \cdot W \cdot (XW)^{\top} && (AB)^{\top} = B^{\top} A^{\top} \\
&= X \cdot W \cdot W^{\top} \cdot X^{\top} && W := Q\sqrt{\Lambda} \\
&= X \cdot Q\sqrt{\Lambda} \cdot (Q\sqrt{\Lambda})^{\top} \cdot X^{\top}
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(X') &= X' \cdot X'^{\top} && \text{first order similarity of } X' \\
&= X \cdot W \cdot (XW)^{\top} && (AB)^{\top} = B^{\top} A^{\top} \\
&= X \cdot W \cdot W^{\top} \cdot X^{\top} && W := Q\sqrt{\Lambda} \\
&= X \cdot Q\sqrt{\Lambda} \cdot (Q\sqrt{\Lambda})^{\top} \cdot X^{\top} && (AB)^{\top} = B^{\top} A^{\top}
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(X') &= X' \cdot X'^{\top} && \text{first order similarity of } X' \\
&= X \cdot W \cdot (XW)^{\top} && (AB)^{\top} = B^{\top}A^{\top} \\
&= X \cdot W \cdot W^{\top} \cdot X^{\top} && W := Q\sqrt{\Lambda} \\
&= X \cdot Q\sqrt{\Lambda} \cdot (Q\sqrt{\Lambda})^{\top} \cdot X^{\top} && (AB)^{\top} = B^{\top}A^{\top} \\
&= X \cdot Q\sqrt{\Lambda} \cdot \sqrt{\Lambda}^{\top} \cdot Q^{\top}X^{\top}
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(X') &= X' \cdot X'^{\top} && \text{first order similarity of } X' \\
&= X \cdot W \cdot (XW)^{\top} && (AB)^{\top} = B^{\top}A^{\top} \\
&= X \cdot W \cdot W^{\top} \cdot X^{\top} && W := Q\sqrt{\Lambda} \\
&= X \cdot Q\sqrt{\Lambda} \cdot (Q\sqrt{\Lambda})^{\top} \cdot X^{\top} && (AB)^{\top} = B^{\top}A^{\top} \\
&= X \cdot Q\sqrt{\Lambda} \cdot \sqrt{\Lambda}^{\top} \cdot Q^{\top}X^{\top} \\
&= X \cdot Q \cdot \Lambda \cdot Q^{\top} \cdot X^{\top}
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(X') &= X' \cdot X'^{\top} && \text{first order similarity of } X' \\
&= X \cdot W \cdot (XW)^{\top} && (AB)^{\top} = B^{\top}A^{\top} \\
&= X \cdot W \cdot W^{\top} \cdot X^{\top} && W := Q\sqrt{\Lambda} \\
&= X \cdot Q\sqrt{\Lambda} \cdot (Q\sqrt{\Lambda})^{\top} \cdot X^{\top} && (AB)^{\top} = B^{\top}A^{\top} \\
&= X \cdot Q\sqrt{\Lambda} \cdot \sqrt{\Lambda}^{\top} \cdot Q^{\top}X^{\top} \\
&= X \cdot Q \cdot \Lambda \cdot Q^{\top} \cdot X^{\top} \\
&= X \cdot X^{\top} \cdot X \cdot X^{\top}
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(X') &= X' \cdot X'^{\top} && \text{first order similarity of } X' \\
&= X \cdot W \cdot (XW)^{\top} && (AB)^{\top} = B^{\top}A^{\top} \\
&= X \cdot W \cdot W^{\top} \cdot X^{\top} && W := Q\sqrt{\Lambda} \\
&= X \cdot Q\sqrt{\Lambda} \cdot (Q\sqrt{\Lambda})^{\top} \cdot X^{\top} && (AB)^{\top} = B^{\top}A^{\top} \\
&= X \cdot Q\sqrt{\Lambda} \cdot \sqrt{\Lambda}^{\top} \cdot Q^{\top}X^{\top} \\
&= X \cdot Q \cdot \Lambda \cdot Q^{\top} \cdot X^{\top} \\
&= X \cdot X^{\top} \cdot X \cdot X^{\top} = M_2(X) && \text{second order similarity of } X
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

More generally

- Define $W_\alpha := Q\Lambda^\alpha$

  where α is a parameter of the transformation that adjusts to the desired similarity order:

first order similarity	$\alpha = 0$	$\Rightarrow$	$M(XW_0) = M(X)$
second order similarity	$\alpha = 0.5$	$\Rightarrow$	$M(XW_{0.5}) = M_2(X)$
n-th order similarity	$\alpha = (n-1)/2$	$\Rightarrow$	$M(XW_\alpha) = M_n(X)$

Linear transformation of embedding matrix

Artetxe et al (2018)

More generally

- Define $W_\alpha := Q\Lambda^\alpha$

  where α is a parameter of the transformation that adjusts to the desired similarity order:

first order similarity	$\alpha = 0$	$\Rightarrow$	$M(XW_0) = M(X)$
second order similarity	$\alpha = 0.5$	$\Rightarrow$	$M(XW_{0.5}) = M_2(X)$
n-th order similarity	$\alpha = (n-1)/2$	$\Rightarrow$	$M(XW_\alpha) = M_n(X)$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$M(XW_0) \quad = \quad M(X \cdot Q \cdot \Lambda^0) \qquad \Lambda^0 \quad \text{Einheitsmatrix}$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$M(XW_0) = M(X \cdot Q \cdot \Lambda^0) \qquad \Lambda^0 \quad \text{Einheitsmatrix}$$

$$= M(X \cdot Q)$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(XW_0) &= M(X \cdot Q \cdot \Lambda^0) \qquad \Lambda^0 \quad \text{Einheitsmatrix} \\
&= M(X \cdot Q) \\
&= X \cdot Q(XQ)^\top
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(XW_0) &= M(X \cdot Q \cdot \Lambda^0) \qquad \Lambda^0 \quad \text{Einheitsmatrix} \\
&= M(X \cdot Q) \\
&= X \cdot Q(XQ)^\top \\
&= X \cdot Q \cdot Q^\top \cdot X^\top
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(XW_0) &= M(X \cdot Q \cdot \Lambda^0) \qquad \Lambda^0 \quad \text{Einheitsmatrix} \\
&= M(X \cdot Q) \\
&= X \cdot Q(XQ)^\top \\
&= X \cdot Q \cdot Q^\top \cdot X^\top \\
&= X \cdot X^\top
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(XW_0) &= M(X \cdot Q \cdot \Lambda^0) \qquad \Lambda^0 \quad \text{Einheitsmatrix} \\
&= M(X \cdot Q) \\
&= X \cdot Q(XQ)^\top \\
&= X \cdot Q \cdot Q^\top \cdot X^\top \\
&= X \cdot X^\top \\
&= M(X)
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

More generally

- Define $W_\alpha := Q\Lambda^\alpha$

  where α is a parameter of the transformation that adjusts to the desired similarity order:

first order similarity	$\alpha = 0$	$\Rightarrow$	$M(XW_0) = M(X)$
second order similarity	$\alpha = 0.5$	$\Rightarrow$	$M(XW_{0.5}) = M_2(X)$
n-th order similarity	$\alpha = (n-1)/2$	$\Rightarrow$	$M(XW_\alpha) = M_n(X)$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$M(XW_{0.5}) \quad = \quad M(X \cdot Q\sqrt{\Lambda})$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(XW_{0.5}) &= M(X \cdot Q\sqrt{\Lambda}) \\
&= X \cdot Q\sqrt{\Lambda}\,(X \cdot Q\sqrt{\Lambda})^{\top}
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(XW_{0.5}) &= M(X \cdot Q\sqrt{\Lambda}) \\
&= X \cdot Q\sqrt{\Lambda}\,(X \cdot Q\sqrt{\Lambda})^{\top} \\
&= X \cdot Q\sqrt{\Lambda}\,\sqrt{\Lambda}^{\top} Q^{\top} \cdot X^{\top}
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(XW_{0.5}) &= M(X \cdot Q\sqrt{\Lambda}) \\
&= X \cdot Q\sqrt{\Lambda}\,(X \cdot Q\sqrt{\Lambda})^{\top} \\
&= X \cdot Q\sqrt{\Lambda}\,\sqrt{\Lambda}^{\top} Q^{\top} \cdot X^{\top} \\
&= X \cdot Q\sqrt{\Lambda}\,\sqrt{\Lambda} \cdot Q^{\top} \cdot X^{\top}
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(XW_{0.5}) &= M(X \cdot Q\sqrt{\Lambda}) \\
&= X \cdot Q\sqrt{\Lambda}\,(X \cdot Q\sqrt{\Lambda})^{\top} \\
&= X \cdot Q\sqrt{\Lambda}\,\sqrt{\Lambda}^{\top} Q^{\top} \cdot X^{\top} \\
&= X \cdot Q\sqrt{\Lambda}\,\sqrt{\Lambda} \cdot Q^{\top} \cdot X^{\top} \\
&= X \cdot Q \cdot \Lambda \cdot Q^{\top} \cdot X^{\top}
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(XW_{0.5}) &= M(X \cdot Q\sqrt{\Lambda}) \\
&= X \cdot Q\sqrt{\Lambda}\,(X \cdot Q\sqrt{\Lambda})^{\top} \\
&= X \cdot Q\sqrt{\Lambda}\,\sqrt{\Lambda}^{\top} Q^{\top} \cdot X^{\top} \\
&= X \cdot Q\sqrt{\Lambda}\,\sqrt{\Lambda} \cdot Q^{\top} \cdot X^{\top} \\
&= X \cdot Q \cdot \Lambda \cdot Q^{\top} \cdot X^{\top} \\
&= X \cdot X^{\top} \cdot X \cdot X^{\top}
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

$$
\begin{aligned}
M(XW_{0.5}) &= M(X \cdot Q\sqrt{\Lambda}) \\
&= X \cdot Q\sqrt{\Lambda}\,(X \cdot Q\sqrt{\Lambda})^{\top} \\
&= X \cdot Q\sqrt{\Lambda}\,\sqrt{\Lambda}^{\top} Q^{\top} \cdot X^{\top} \\
&= X \cdot Q\sqrt{\Lambda}\,\sqrt{\Lambda} \cdot Q^{\top} \cdot X^{\top} \\
&= X \cdot Q \cdot \Lambda \cdot Q^{\top} \cdot X^{\top} \\
&= X \cdot X^{\top} \cdot X \cdot X^{\top} \\
&= M_2(X)
\end{aligned}
$$

Linear transformation of embedding matrix

Artetxe et al (2018)

More generally

- Define $W_\alpha := Q\Lambda^\alpha$

  where α is a parameter of the transformation that adjusts to the desired similarity order:

first order similarity	$\alpha = 0$	$\Rightarrow$	$M(XW_0) = M(X)$
second order similarity	$\alpha = 0.5$	$\Rightarrow$	$M(XW_{0.5}) = M_2(X)$
n-th order similarity	$\alpha = (n-1)/2$	$\Rightarrow$	$M(XW_\alpha) = M_n(X)$

Linear transformation of embedding matrix

Artetxe et al (2018)

- Assuming that the embeddings X capture some second order similarity, it is possible to transform them so that they capture the corresponding first order similarity
- One can easily generalise this to higher order similarities by using smaller values of α

$\Rightarrow$ Parameter α can be used to either increase or decrease the similarity order that we want our embeddings to capture

$\Rightarrow$ α can be continuous

Linear transformation of different embeddings

Artetxe et al (2018)

		Word analogy		Word similarity	
		Semantic	Syntactic	Similarity (SimLex-999)	Relatedness (MEN)
word2vec	Original	76.49	74.87	44.21	76.96
	Best				
glove	Original	83.17	76.19	40.70	80.06
	Best				
fasttext	Original	89.76	82.44	50.48	83.55
	Best				

Table 1: Results in intrinsic evaluation for the original embeddings and the best post-processed model with the corresponding value of α. The evaluation measure is accuracy for word analogy and Spearman correlation for word similarity.

Linear transformation of different embeddings

Artetxe et al (2018)

		Word analogy		Word similarity	
		Semantic	Syntactic	Similarity (SimLex-999)	Relatedness (MEN)
word2vec	Original	76.49	74.87	44.21	76.96
	Best	81.00 α = -0.65	74.96 α = 0.10	47.81 α = -0.70	78.09 α = -0.30
glove	Original	83.17	76.19	40.70	80.06
	Best	86.73 α = -0.85	76.51 α = -0.10	51.54 α = -0.85	84.00 α = -0.45
fasttext	Original	89.76	82.44	50.48	83.55
	Best	90.85 α = -0.45	84.45 α = 0.25	51.55 α = -0.25	84.06 α = -0.15

Table 1: Results in intrinsic evaluation for the original embeddings and the best post-processed model with the corresponding value of α. The evaluation measure is accuracy for word analogy and Spearman correlation for word similarity.

Lessons learned for intrinsic and extrinsic evaluations

Artetxe et al (2018)

- Standard intrinsic evaluation is static and incomplete

  $\Rightarrow$ Intrinsic evaluation not a good predictor for performance in downstream applications

  $\Rightarrow$ Systems that use embeddings as features can learn task-specific optimal balance between the two axes

References

- Mikolov, Yih and Zweig: (2013): Linguistic regularities in continuous space word representations. NAACL 2013.
- Faruqui, Tsvetkov, Rastogi and Dyer (2016): Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. The 1st Workshop on Evaluating Vector Space Representations for NLP, Berlin, Germany.
- Artetxe, Labaka, Lopez-Gazpio and Agirre (2018): Uncovering Divergent Linguistic Information in Word Embeddings with Lessons for Intrinsic and Extrinsic Evaluation. CoNLL 2018. Brussels, Belgium.
- Rubenstein and Goodenough (1965): Contextual correlates of synonymy. Communications of the ACM 8(10):627–633.
- Harris, Z. (1954). Distributional structure. Word, 10(23): 146-162.
- Multimodal Distributional Semantics E. Bruni, N. K. Tran and M. Baroni. Journal of Artificial Intelligence Research 49: 1-47.
- Collobert, Weston Bottou, Karlen, Kavukcuoglu and Kuksa (2011): Natural Language Processing (almost) from Scratch. Journal of Machine Learning Research 12 (2011) 2461-2505.
- Lu, Wang, Bansal, Gimpel and Livescu (2015): Deep multilingual correlation for improved word embeddings. NAACL 2015.
- Rastogi, Van Durme and Arora (2015): Multiview LSA: Representation learning via generalized CCA. NAACL 2015.
- Chiu, Korhonen and Pyysalo (2016): Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance. ACL 2016.

- Data and Code
  - Code for Artetxe etal. (2018): `https://github.com/artetxem/uncovec`
  - The MEN dataset `https://staff.fnwi.uva.nl/e.bruni/MEN`
  - Datasets for word vector evaluation `https://github.com/vecto-ai/word-benchmarks`