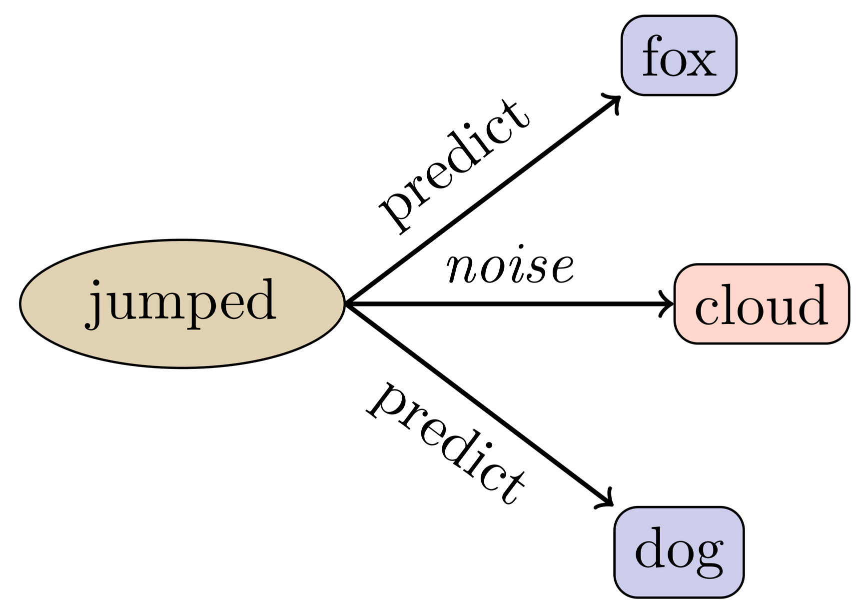
Neural-based Noise Filtering from Word Embeddings

Kim Anh Nguyen, Sabine Schulte im Walde, Ngoc Thang Vu
Institute for Natural Language Processing - University of Stuttgart, Germany
`{nguyenkh,schulte,thangvu}@ims.uni-stuttgart.de`

Motivation

- Hypothesis: Word embeddings contain unnecessary information, i.e. *noise*.
- Goal: Improve word embeddings by reducing their noise.
- Procedure: Strengthen salient contexts and weaken unnecessary contexts.
- Example sentence:

`The quick brown fox gazing at the cloud jumped over the lazy dog.`



Contributions

We propose two neural models to filter noise from word embeddings:

1. The complete word denoising embeddings model *(CompEmb)*
2. The overcomplete word denoising embeddings model *(OverCompEmb)*

The denoising embeddings outperform the originally state-of-the-art embeddings on several benchmark tasks.

Models

1. Complete Word Denoising Embeddings (CompEmb):

- Learn a denoising matrix $\mathbf{Q_c}$ by optimizing the following objective function:

$$\underset{\mathbf{X},\mathbf{Q_c},\mathbf{S}}{\operatorname{argmin}} \sum_{i=1}^{V} \|\mathbf{x_i} - f(\mathbf{x_i}, \mathbf{Q_c}, \mathbf{S})\| + \alpha\|\mathbf{S}\|_1 \quad (1)$$

- Project the original embeddings $\mathbf{X}$ into $\mathbf{Q_c}$ to generate the *CompEmb* $\mathbf{X}^*$:

$$\mathbf{X}^* = \mathcal{G}(\mathbf{X}\mathbf{Q_c}) \quad (2)$$

2. Overcomplete Word Denoising Embeddings (OverCompEmb):

- Transform the original embeddings $\mathbf{X}$ into the overcomplete embeddings $\mathbf{Z}$.
- Learn a denoising matrix $\mathbf{Q_o}$ by optimizing the following objective function:

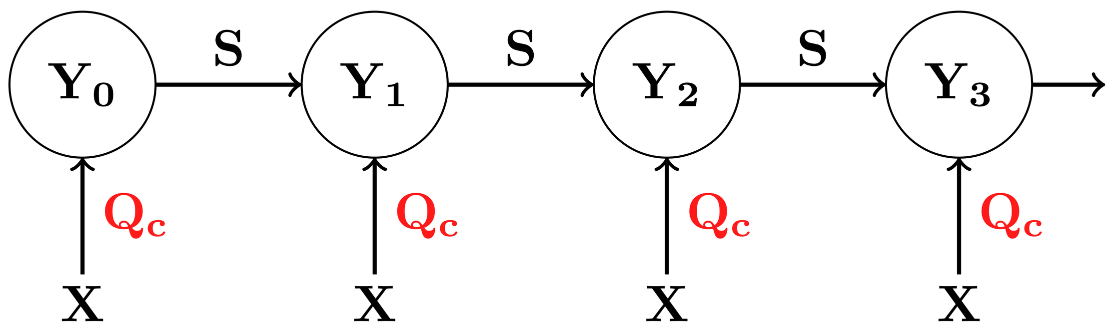
$$\underset{\mathbf{X},\mathbf{Q_o},\mathbf{S}}{\operatorname{argmin}} \sum_{i=1}^{V} \|\mathbf{z_i} - f(\mathbf{x_i}, \mathbf{Q_o}, \mathbf{S})\| + \alpha\|\mathbf{S}\|_1 \quad (3)$$

- Project the original embeddings $\mathbf{X}$ into $\mathbf{Q_o}$ to generate the *OverCompEmb* $\mathbf{Z}^*$:

$$\mathbf{Z}^* = \mathcal{G}(\mathbf{X}\mathbf{Q_o}) \quad (4)$$

3. The architecture of filters f:

- The architecture of filter f for learning $\mathbf{Q_c}$:



- The architecture of filter f for learning $\mathbf{Q_o}$:

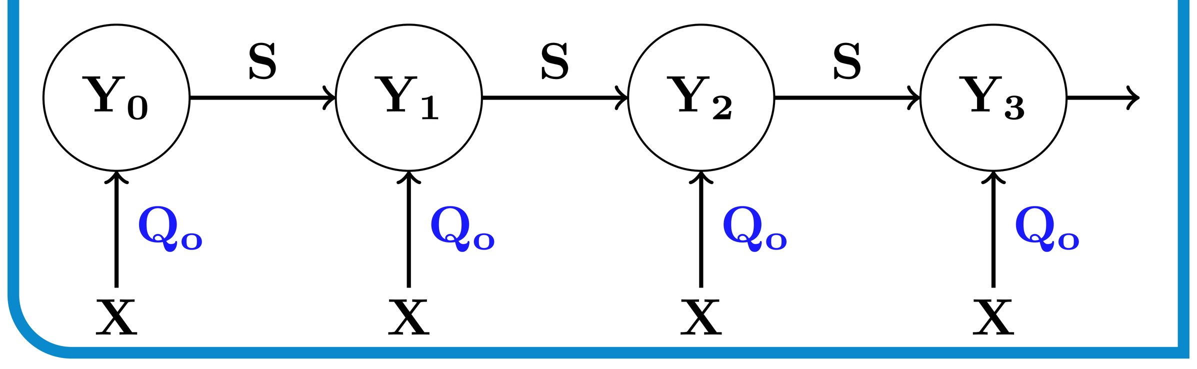
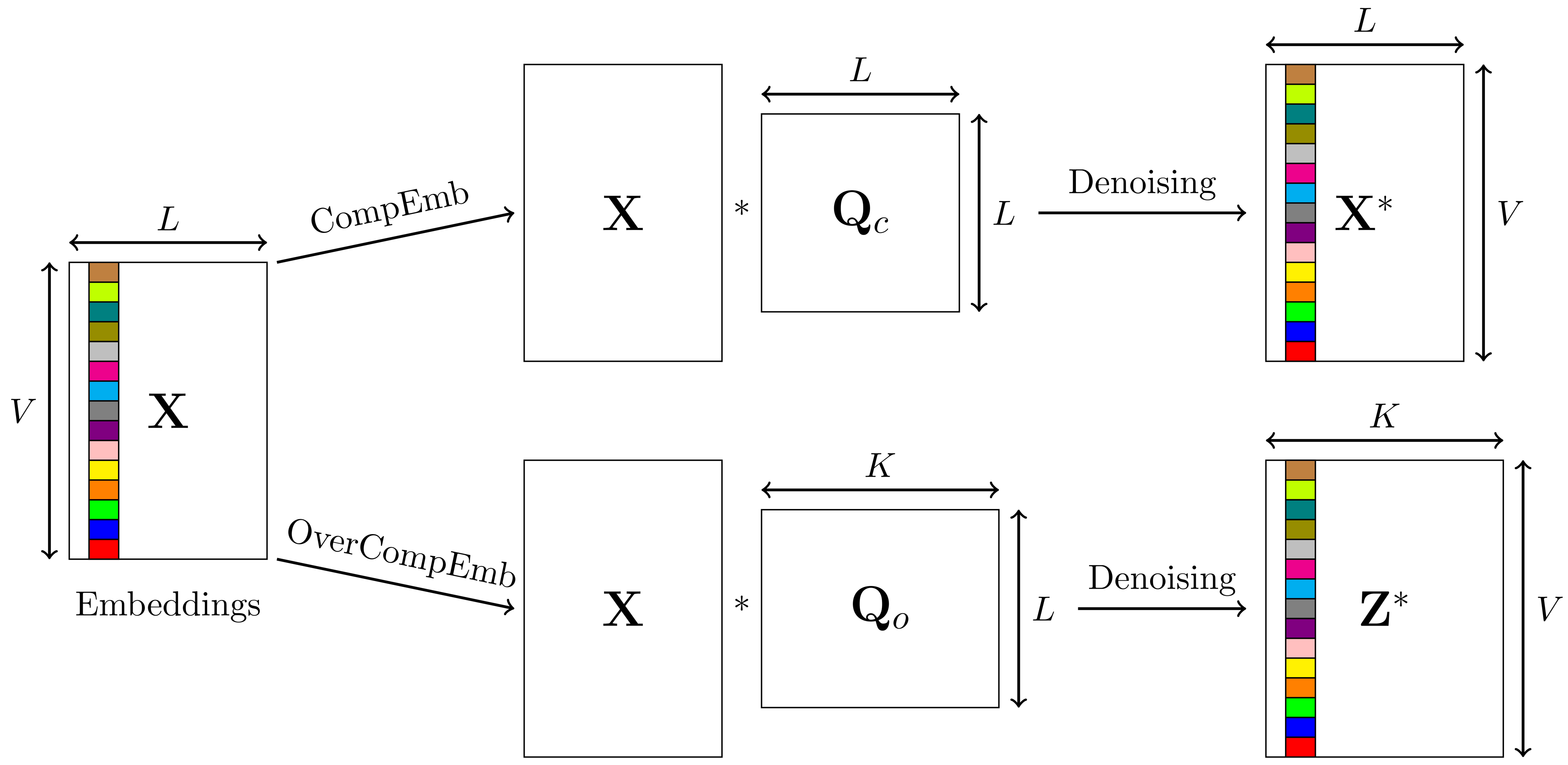


Illustration of Models



Experimental Settings

Task	Dataset	#Instance
Similarity	SimLex-999	999
	WS353-SIM	203
Relatedness	MEN	3000
	WS353-REL	252
Synonymy	TOEFL	80
	ESL	50
NP Classification	Lazaridou et al. (2013)	2227

Parameters	Values
Embeddings	dim = 300; dim=100
Overcomplete factor	$\gamma = 10$
ℓ_1 regularization	$\alpha = 0.5$; $\lambda = 10^{-6}$
Filter depth	$T = 3$
Corpus size	14.5B tokens

Results

Vectors		Simlex-999 Corr.	MEN Corr.	WS353 Corr.	WS353-SIM Corr.	WS353-REL Corr.	ESL Acc.	TOEFL Acc.	NP Acc.
SG-100	X	33.7	72.9	69.7	74.5	65.5	48.9	62.0	72.8
	X*	33.2	72.8	70.6	74.8	66.0	53.0	**64.5**	78.5
	Z*	**35.9**	**74.4**	**71.2**	**75.2**	**68.1**	53.0	62.0	**79.1**
	A	32.5	69.8	65.5	69.5	60.2	**55.1**	51.8	78.8
	B	31.9	70.4	65.8	72.6	62.2	53.0	58.2	74.1
SG-300	X	36.1	74.7	71.0	75.9	66.1	**59.1**	72.1	77.9
	X*	**37.1**	**75.8**	**71.8**	**76.4**	**66.9**	**59.1**	74.6	**79.3**
	Z*	36.5	75.0	70.6	**76.4**	64.4	57.1	**77.2**	78.6
	A	32.9	72.4	67.5	71.9	63.4	53.0	65.8	78.3
	B	32.7	71.2	63.3	68.7	56.2	51.0	70.8	78.6
GloVe-100	X	29.7	69.3	52.9	60.3	49.5	46.9	82.2	76.4
	X*	**31.7**	**70.9**	**58.0**	**63.8**	**57.3**	53.0	**88.6**	**77.4**
	Z*	30.0	**70.9**	56.0	62.8	53.8	**57.0**	81.0	77.3
	A	30.7	70.7	54.9	62.2	51.2	55.1	78.4	77.1
	B	31.0	69.2	57.3	62.3	53.7	46.9	73.4	76.4
GloVe-300	X	37.0	74.8	60.5	66.3	57.2	**61.2**	89.8	74.3
	X*	**40.2**	**76.8**	**64.9**	**69.8**	**62.0**	**61.2**	**92.4**	76.3
	Z*	39.0	75.2	63.0	67.9	59.7	57.1	86.0	75.7
	A	36.7	74.1	61.5	67.7	57.8	55.1	87.3	79.9
	B	33.1	70.2	57.0	62.2	53.0	51.0	91.4	**80.0**

Vectors $\mathbf{X}$ represent the baselines; vectors $\mathbf{A}$ and $\mathbf{B}$ were suggested by Faruqui et al. (2015); vectors $\mathbf{X}^*$ and $\mathbf{Z}^*$ are the denoising embeddings in which the vector length $\mathbf{Z}^*$ is equal to 10 times of vector length $\mathbf{X}$.