Encoding Prior Knowledge with Eigenword Embeddings

Dominique Osborne[1], Shashi Narayan[2] & Shay Cohen[2]

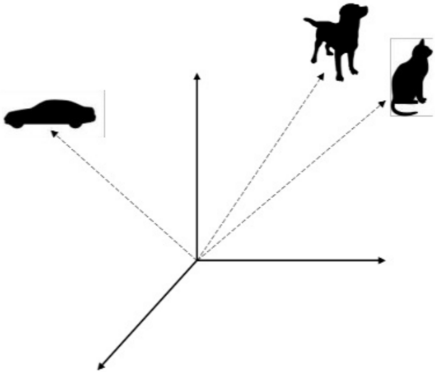
[1]Department of Mathematics and Statistics, University of Strathclyde

[2]School of Informatics, University of Edinburgh

EACL 2017

Word embeddings ...

cat	(0.1, 0.2, 0, 0.2, 0.03, ...)
dog	(0.2, 0.02, 0.1, 0.1, 0.02, ...)
car	(0.001, 0, 0, 0.1, 0.3, ...)



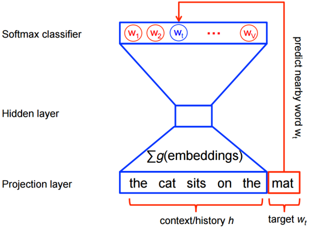
Learning dense representations

Neural networks

Matrix factorization

	context$_1$	context$_2$	...	context$_n$
word$_1$				
word$_2$				
...				
word$_n$				

- LSA (word-document) (Deerwester et al., 1990)
- GloVe (word-neighbourWords) (Pennington et al., 2014)
- CCA based Eigenword (word-neighbourWords) (Dhillon et al., 2015)



- NLM (word-neighbourWords) (Bengio et al., 2003)
- Word2Vec (Mikolov et al., 2013)

Distributional hypothesis (Harris, 1954)

Adding knowledge to word embeddings

- Refining vector space representations using semantic lexicons such as WordNet, FrameNet, and the Paraphrase Database, to
- encourage linked words to have similar vector representations.
- Often operates as a post processing step, e.g., Retrofitting (Faruqui et at, 2015) and AutoExtend (Rothe and Schutze, 2015).

In this talk ...

Encode semantic knowledge to CCA-based eigenword embeddings

- Spectral learning algorithms are interesting for their speed, scalability, globally optimal solution, and performance in various NLP applications.

In this talk ...

Encode semantic knowledge to CCA-based eigenword embeddings

- Spectral learning algorithms are interesting for their speed, scalability, globally optimal solution, and performance in various NLP applications.
- We introduce prior knowledge in the CCA derivation itself.
  - Preserves the properties of spectral learning algorithms for learning word embeddings.
  - Applicable for incorporating prior knowledge into any CCA.

CCA-based Eigenword embeddings (Dhillon et al., 2015)

Training set: $\{(w_1^{(i)}, \ldots, w_k^{(i)}, w^{(i)}, w_{k+1}^{(i)}, \ldots, w_{2k}^{(i)}) \mid i \in [n]\}$

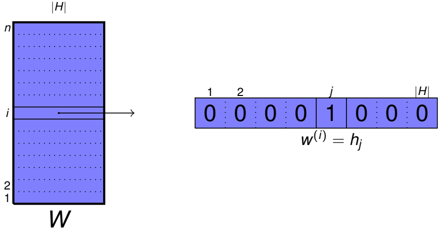
- Pivot word: $w^{(i)}$
- Left context: $\{w_1^{(i)}, \ldots, w_k^{(i)}\}$
- Right context: $\{w_{k+1}^{(i)}, \ldots, w_{2k}^{(i)}\}$

CCA finds projections for the contexts and for the pivot words which are most correlated (follows distributional hypothesis of Harris, 1954)

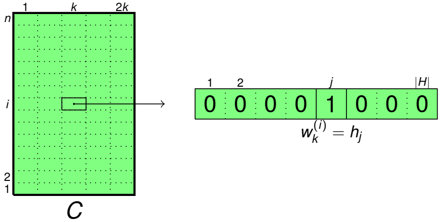
Defining two views for CCA

Training set: $\{(w_1^{(i)}, \ldots, w_k^{(i)}, w^{(i)}, w_{k+1}^{(i)}, \ldots, w_{2k}^{(i)}) \mid i \in [n]\}$

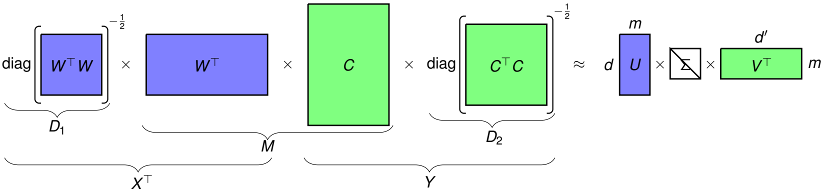
Word matrix $W \in \mathbb{R}^{n \times |H|}$



Context matrix $C \in \mathbb{R}^{n \times 2k|H|}$



Dimensionality reduction with SVD



Eigenword embedding

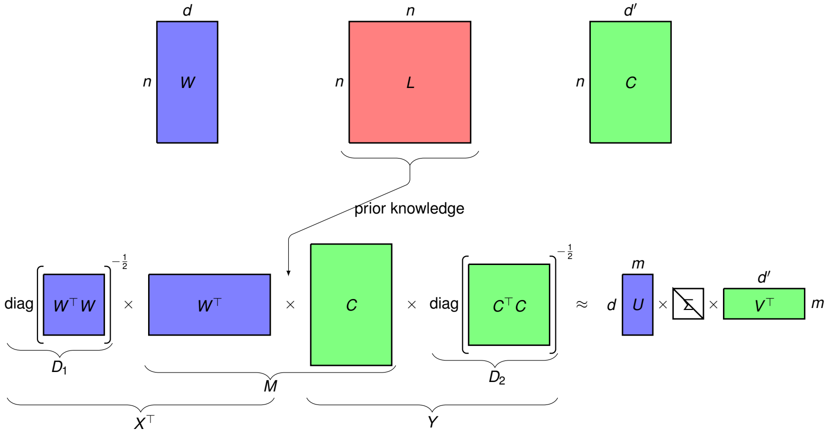
$$E = D_1^{-1/2} U \in \mathbb{R}^{|H| \times m}$$

Adding prior knowledge to Eigenword embeddings

Introduce prior knowledge in the CCA derivation itself to preserves the properties of spectral learning algorithms

Prior knowledge $\Leftarrow$ WordNet, FrameNet and the Paraphrase Database

Adding prior knowledge to Eigenword embeddings

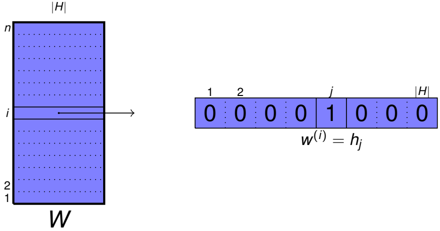


Improve the optimization of correlation between the two views by weighing them using the external source of prior knowledge

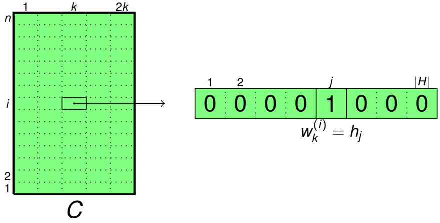
Two views for CCA

Training set: $\{(w_1^{(i)}, \ldots, w_k^{(i)}, w^{(i)}, w_{k+1}^{(i)}, \ldots, w_{2k}^{(i)}) \mid i \in [n]\}$

Word matrix $W \in \mathbb{R}^{n \times |H|}$



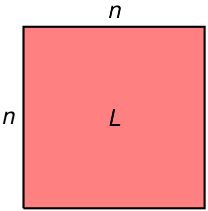
Context matrix $C \in \mathbb{R}^{n \times 2k|H|}$



Prior knowledge as the weight matrix

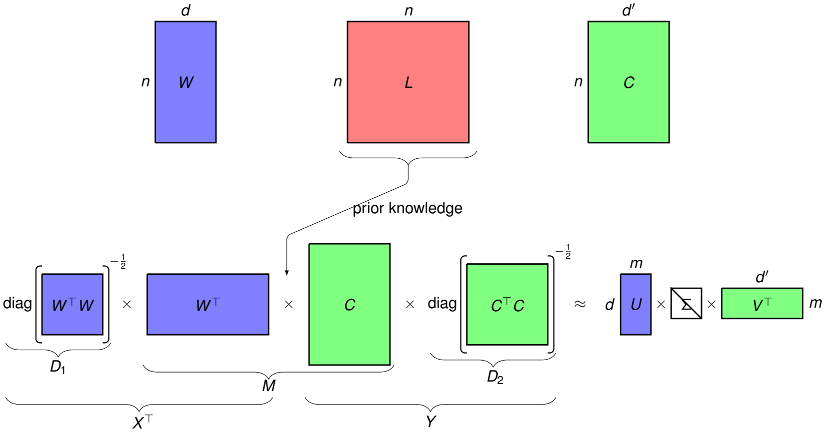
Training set: $\{(w_1^{(i)}, \ldots, w_k^{(i)}, w^{(i)}, w_{k+1}^{(i)}, \ldots, w_{2k}^{(i)}) \mid i \in [n]\}$

Weight matrix over examples: $L \in R^{nxn}$



Captures adjacency information in the semantic lexicons, such as WordNet, FrameNet, and the Paraphrase Database

Adding prior knowledge to Eigenword embeddings



Do we still find projections for the contexts and for the pivot words which are most correlated?

Generalisation of CCA

Yes, if L is a Laplacian matrix!

Laplacian matrix $L \in R^{nxn}$

A symmetric positive semi-definite square matrix such that the sum over rows (or columns) is 0.

$$L_{ij} = \begin{cases} n-1 & \text{if } i = j \\ -1 & \text{if } i \neq j. \end{cases}$$

Lemma

$X^\top LY$ equals $X^\top Y$ up to a multiplication by a positive constant.

Optimizes same objective function!

Generalisation of CCA

$$
\begin{aligned}
max(\sum_{k=1}^{m}(Xu_k)^\top L\,(Yv_k)) &= max(\sum_{i,j} -L_{ij}\left(d_{ij}^{m}\right)^2) \\
&= max(\sum_{i,j}\left(d_{ij}^{m}\right)^2 - n\sum_{i=1}^{n}(d_{ii}^{m})^2)
\end{aligned}
$$

where d_{ij}^{m} is the distance between projections of ith word and jth context views.

CCA follows distributional hypothesis, with additional constraints from prior knowledge.

Experiments

- Evaluation Benchmarks
  - Word Similarity: 11 different widely used benchmarks, e.g., the WS-353-ALL dataset (Finkelstein et al., 2002) and the SimLex-999 dataset (Hill et al., 2015)
  - Geographic Analogies: *"Greece (a) is to Athens (b) as Iraq (c) is to (d)"* (Mikolov et al. 2013)
    - $d = c - (a - b)$
  - NP Bracketing: *"annual (price growth)"* vs *"(annual price) growth"* (Lazaridou et al., 2013)

Experiments

- Prior Knowledge Resources: WordNet, the Paraphrase Database (PPDB), and FrameNet.

- Baselines
  - Off-the-shelf Word Embeddings: Glove (Pennington et al., 2014), Skip-Gram (Mikolov et al., 2013), Global Context (Huang et al., 2012), Multilingual (Faruqui and Dyer, 2014) and Eigen word embeddings (Dhillon et al. (2015)
  - Retrofitting (Faruqui et al., 2015)

All embeddings were trained on the first 5 billion words from Wikipedia.

Results

NPK: No prior knowledge, WN: WordNet, PD: the paraphrase database and FN: FrameNet.

		Word similarity average				Geographic analogies				NP bracketing			
		NPK	WN	PD	FN	NPK	WN	PD	FN	NPK	WN	PD	FN
Retrofitting	Glove	59.7	63.1	64.6	57.5	**94.8**	75.3	80.4	**94.8**	78.1	79.5	79.4	78.7
	Skip-Gram	64.1	65.5	**68.6**	62.3	87.3	72.3	70.5	87.7	79.9	80.4	81.5	80.5
	Global Context	44.4	50.0	50.4	47.3	7.3	4.5	18.2	7.3	79.4	79.1	80.5	80.2
	Multilingual	62.3	66.9	68.2	62.8	70.7	46.2	53.7	72.7	81.9	81.8	**82.7**	82.0
	Eigen (CCA)	59.5	62.2	63.6	61.4	89.9	79.2	73.5	89.9	81.3	81.7	81.2	80.7
CCAPrior		-	**60.7**	60.6	60.0	-	89.1	**93.2**	92.9	-	81.8	**82.4**	81.0
CCAPrior+RF		-	63.4	**64.9**	61.6	-	78.0	71.9	**92.5**	-	**81.9**	81.7	81.2

Results

NPK: No prior knowledge, WN: WordNet, PD: the paraphrase database and FN: FrameNet.

		Word similarity average				Geographic analogies				NP bracketing			
		NPK	WN	PD	FN	NPK	WN	PD	FN	NPK	WN	PD	FN
Retrofitting	Glove	59.7	63.1	64.6	57.5	94.8	75.3	80.4	94.8	78.1	79.5	79.4	78.7
	Skip-Gram	64.1	65.5	68.6	62.3	87.3	72.3	70.5	87.7	79.9	80.4	81.5	80.5
	Global Context	44.4	50.0	50.4	47.3	7.3	4.5	18.2	7.3	79.4	79.1	80.5	80.2
	Multilingual	62.3	66.9	68.2	62.8	70.7	46.2	53.7	72.7	81.9	81.8	82.7	82.0
	Eigen (CCA)	59.5	62.2	63.6	61.4	89.9	79.2	73.5	89.9	81.3	81.7	81.2	80.7
CCAPrior		-	**60.7**	**60.6**	**60.0**	-	**89.1**	**93.2**	**92.9**	-	**81.8**	**82.4**	**81.0**
CCAPrior+RF		-	63.4	64.9	61.6	-	78.0	71.9	92.5	-	81.9	81.7	81.2

Adding prior knowledge to eigenword embeddings does improve the quality of word vectors

Results

NPK: No prior knowledge, WN: WordNet, PD: the paraphrase database and FN: FrameNet.

		Word similarity average				Geographic analogies				NP bracketing			
		NPK	WN	PD	FN	NPK	WN	PD	FN	NPK	WN	PD	FN
Retrofitting	Glove	59.7	63.1	64.6	57.5	94.8	75.3	80.4	94.8	78.1	79.5	79.4	78.7
	Skip-Gram	64.1	65.5	68.6	62.3	87.3	72.3	70.5	87.7	79.9	80.4	81.5	80.5
	Global Context	44.4	50.0	50.4	47.3	7.3	4.5	18.2	7.3	79.4	79.1	80.5	80.2
	Multilingual	62.3	66.9	68.2	62.8	70.7	46.2	53.7	72.7	81.9	81.8	82.7	82.0
	Eigen (CCA)	59.5	62.2	63.6	61.4	89.9	79.2	73.5	89.9	81.3	81.7	81.2	80.7
CCAPrior		-	**60.7**	**60.6**	**60.0**	-	**89.1**	**93.2**	**92.9**	-	**81.8**	**82.4**	**81.0**
CCAPrior+RF		-	63.4	64.9	61.6	-	78.0	71.9	92.5	-	81.9	81.7	81.2

Retrofitting further improves eigenword embeddings

Results

NPK: No prior knowledge, WN: WordNet, PD: the paraphrase database and FN: FrameNet.

		Word similarity average				Geographic analogies				NP bracketing			
		NPK	WN	PD	FN	NPK	WN	PD	FN	NPK	WN	PD	FN
Retrofitting	Glove	59.7	63.1	64.6	57.5	94.8	75.3	80.4	94.8	78.1	79.5	79.4	78.7
	Skip-Gram	64.1	65.5	68.6	62.3	87.3	72.3	70.5	87.7	79.9	80.4	81.5	80.5
	Global Context	44.4	50.0	50.4	47.3	7.3	4.5	18.2	7.3	79.4	79.1	80.5	80.2
	Multilingual	62.3	66.9	68.2	62.8	70.7	46.2	53.7	72.7	81.9	81.8	82.7	82.0
	Eigen (CCA)	59.5	62.2	63.6	61.4	89.9	79.2	73.5	89.9	81.3	81.7	81.2	80.7
CCAPrior		-	**60.7**	**60.6**	**60.0**	-	**89.1**	**93.2**	**92.9**	-	**81.8**	**82.4**	**81.0**
CCAPrior+RF		-	63.4	64.9	61.6	-	78.0	71.9	92.5	-	81.9	81.7	81.2

CCA results are more stable than retrofitting

Conclusion

- We described a method for incorporating prior knowledge into CCA-based eigenword embeddings.
- Adding prior knowledge to eigenword embeddings improves the quality of word vectors.
- We proposed a general framework for incorporating prior knowledge into any CCA.