Evaluating Neural Word Representations in Tensor-Based Compositional Settings

Dmitrijs Milajevs[QM], Dimitri Kartsaklis[OX], Mehrnoosh Sadrzadeh[QM], Matthew Purver[QM]

[QM]Queen Mary University of London
School of Electronic Engineering and Computer Science
Mile End Road, London, UK

[OX]University of Oxford
Department of Computer Science
Parks Road, Oxford, UK

EPSRC
Engineering and Physical Sciences Research Council

Modelling word and sentence meaning

Formal semantics

John: j

Mary: m

saw: λx.λy.saw(y,x)

John saw Mary: saw(j, m)

Distributional hypothesis

- Word similarity
  - **John** is more similar to **Mary** that to **idea**.
- Sentence similarity
  - **Dogs chase cats** vs. **Hounds pursue kittens**
    vs. **Cats chase dogs**
    vs. **Students chase deadline**

Distributional approach

For each **target** word

A lorry might **carry** sweet apples

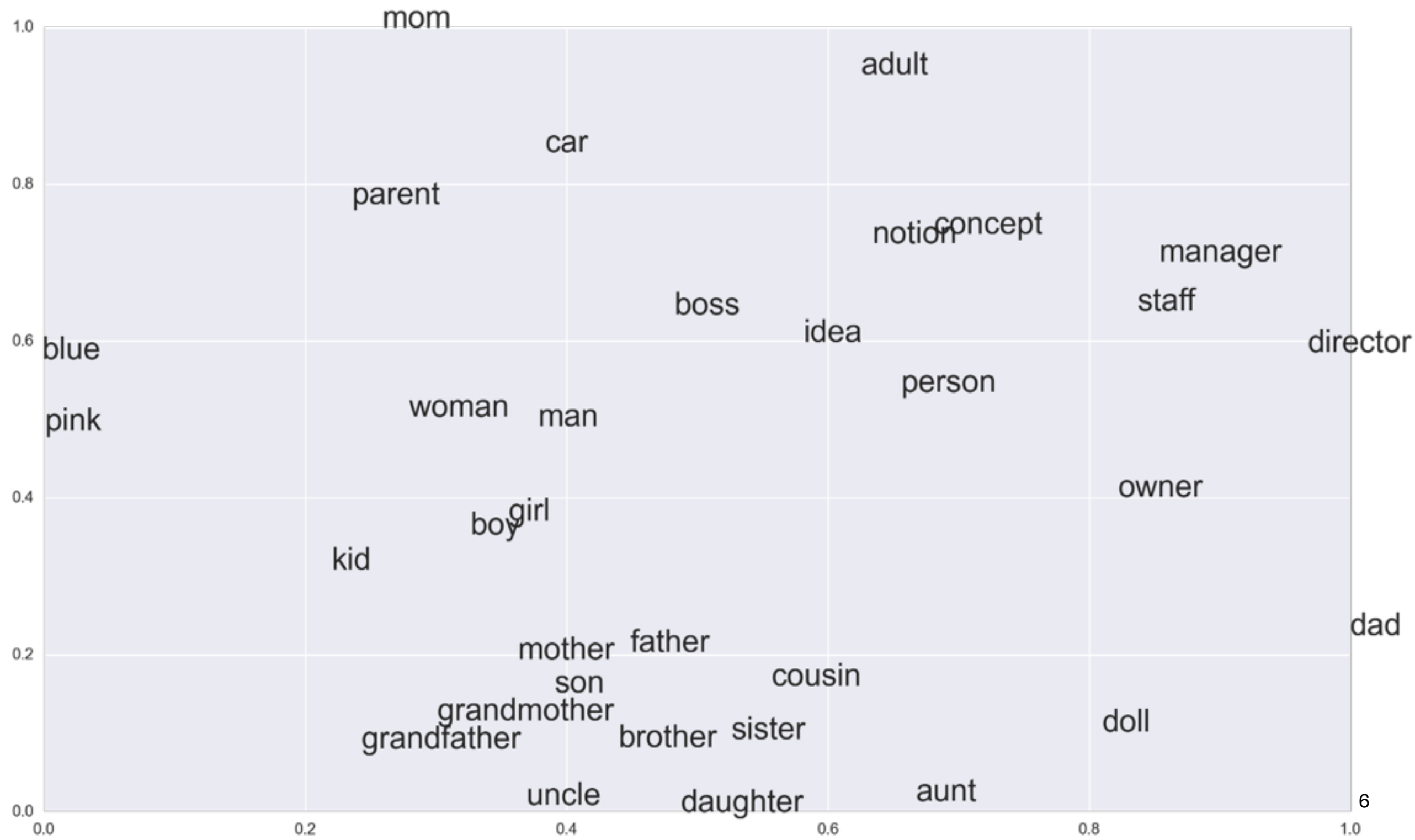
and a neighbouring ***context*** *words*

A lorry *might* **carry** *sweet* apples

update a co-occurrence matrix

	might	sweet	red	…
carry	+1	+1	+0	…

Similarity of two words ~ distance between vectors



Neural word embeddings (language modelling)

Corpus: **The cat is walking in the bedroom**

Unseen **A dog was running in a room** should be almost as likely, because of similar semantic and grammatical roles. Bengio et al., 2006

Mikolov et al. scaled up the estimation procedure to a large corpus and provided a dataset to test extracted relations.

Tensor based models

Representing verb as a matrix

General duality theorem: tensors are in one–one correspondence with multilinear maps. Bourbaki, ‘89

$$\overline{z} \in V \otimes W \otimes \cdots \otimes Z \;\cong\; f_{\overline{z}} \colon V \to W \to \cdots \to Z$$

In a tensor based model, transitive verbs are matrices.

Relational

$$\overline{\text{Verb}} = \sum_i \overrightarrow{\text{Sbj}_i} \otimes \overrightarrow{\text{Obj}_i}$$

Kronecker

$$\widetilde{\text{Verb}} = \overrightarrow{\text{Verb}} \otimes \overrightarrow{\text{Verb}}$$

Compositional models for (Obj, Verb, Sbj)

Mitchell and Lapata ‘08

Addition

Multiplication

Kartsaklis et al. ‘12

Copy object: $\overrightarrow{\text{Sbj}} \odot (\overline{\text{Verb}} \times \overrightarrow{\text{Obj}})$

Copy subject: $\overrightarrow{\text{Obj}} \odot (\overline{\text{Verb}}^{\text{T}} \times \overrightarrow{\text{Sbj}})$

Grefenstette and Sadrzadeh ‘11

Relational: $\overline{\text{Verb}} \odot (\overrightarrow{\text{Sbj}} \otimes \overrightarrow{\text{Obj}})$

Kronecker: $\widetilde{\text{Verb}} \odot (\overrightarrow{\text{Sbj}} \otimes \overrightarrow{\text{Obj}})$

Kartsaklis and Sadrzadeh ‘14

Frobenius addition

Frobenius multiplication

Frobenius outer

Experiments

Vector spaces

GS11: BNC, lemmatised, 2000 dimensions, PPMI

KS14: ukWaC, lemmatised, 300 dimensions, LMI, SVD

NWE: Google news, 300 dimensions, word2vec

Disambiguation

Grefenstette and Sadrzadeh ’11 and ‘14

satisfies

System **meets** specification

visits

Similarity of sentences

Grefenstette and Sadrzadeh ’11 and ‘14

System **satisfies** specification

System **meets** specification

System **visits** specification

Verb only baseline

satisfy

System **meets** specification

visit

Disambiguation results

Method	GS11	KS14	NWE
Verb only	0.212	0.325	0.107
Addition	0.103	0.275	0.149
Multiplication	0.348	0.041	0.095
Kronecker	0.304	0.176	0.117
Relational	0.285	0.341	0.362
Copy subject	0.089	0.317	0.131
Copy object	0.334	0.331	**0.456**
Frobenius add.	0.261	0.344	0.359
Frobenius mult.	0.233	0.341	0.239
Frobenius out.	0.284	0.350	0.375

Spearman rho

Sentence similarity

Kartsaklis, Sadrzadeh, Pulman (CoNLL ’12) Kartsaklis, Sadrzadeh (EMNLP ‘13)

panel discuss issue

project present problem

man shut door

gentleman close eye

paper address question

study pose problem

Sentence similarity

Method	GS11	KS14	NWE
Verb only	0.491	0.602	0.561
Addition	0.682	**0.732**	0.689
Multiplication	0.597	0.321	0.341
Kronecker	0.581	0.408	0.561
Relational	0.558	0.437	0.618
Copy subject	0.370	0.448	0.405
Copy object	0.571	0.306	0.655
Frobenius add.	0.566	0.460	0.585
Frobenius mult.	0.525	0.226	0.387
Frobenius out.	0.560	0.439	0.662

Spearman rho

Paraphrasing

- MS Paraphrasing corpus
- Compute similarity of a pair of sentences
- Choose a threshold similarity value on training data
- Evaluate on the test set

Paraphrase results

Method	GS11	KS14	NWE
Addition	0,62 (0,79)	0,70 (0,80)	**0,73 (0,82)**
Multiplication	0,52 (0,58)	0,66 (0,80)	0,42 (0,34)

Accuracy (F-Score)

Dialogue act tagging

Milajevs and Purver ’14, Serafin et al. ’03

Switchboard: telephone conversation corpus.

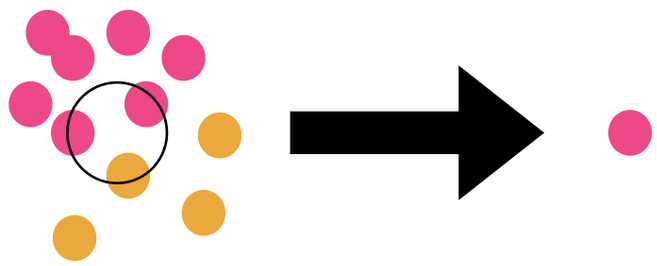
1. Utterance-feature matrix

$$\overline{\texttt{I}} \oplus \overline{\texttt{wonder}} \oplus \overline{\texttt{if}} \oplus \overline{\texttt{that}} \oplus \overline{\texttt{worked}} \oplus \overline{\texttt{.}}$$

2. Utterance vectors are reduced using SVD to 50 dimensions

$$M \approx U\tilde{\Sigma}V^T = \tilde{M}$$

3. k-nearest neighbours classification



Dialogue act tagging results

Method	GS11	KS14	NWE lemmatised	NWE
Addition	0,35 (0,35)	0,40 (0,35)	0,44 (0,40)	**0,63 (0,60)**
Multiplication	0,32 (0,16)	0,39 (0,33)	0,43 (0,38)	0,58 (0,53)

Accuracy (F-Score)

Discussion

"*context-predicting models obtain a thorough and resounding victory against their count-based counterparts*" Baroni et al. (2014)

"*analogy recovery is not restricted to neural word embeddings [...] a similar amount of relational similarities can be recovered from traditional distributional word representations*" Levy et al. (2014)

"*shallow approaches are as good as more computationally intensive alternatives on phrase similarity and paraphrase detection tasks"* Blacoe and Lapata (2012)

Improvement over baselines

Task	GS11	KS14	NWE
Disambiguation	+	+	**+**
Sentence similarity	+	**-**	+
Paraphrase	-	+	**+**
Dialog act tagging	-	-	**+**

Conclusion

- The choice of compositional operator seems to be more important than the word vector nature and more task specific.
- Tensor-based composition does not yet always outperform simple compositional operators.
- Neural word embeddings are more successful than the co-occurrence based alternatives.
    - Corpus size might contribute a lot.